Operations Research

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

Optimal Rate Scheduling via Utility-Maximization for J-User MIMO Markov Fading Wireless Channels with Cooperation

Wanyang Dai

Operations Research
Vol. 61, No. 6, November–December 2013, pp. 1450–1462
ISSN 0030-364X (print) | ISSN 1526-5463 (online)

http://dx.doi.org/10.1287/opre.2013.1224

Optimal Rate Scheduling via Utility-Maximization for J-User MIMO Markov Fading Wireless Channels with Cooperation

Wanyang Dai
Department of Mathematics and State Key Laboratory of Novel Software Technology, Nanjing University, Nanjing 210093, China, nan5lu8@netra.nju.edu.cn

We design a dynamic rate scheduling policy of Markov type by using the solution (a social optimal Nash equilibrium point) to a utility-maximization problem over a randomly evolving capacity set for a stochastic system of generalized processor-sharing queues in a random environment whose job arrivals to each queue follow a *doubly stochastic renewal process* (DSRP). Both the random environment and the random arrival rate of each DSRP are driven by a finite state continuous time Markov chain. The scheduling policy optimizes in a *greedy* fashion with respect to each queue and environmental state. Since the closed-form solution for the performance of such a queuing system under the policy is difficult to obtain, we establish a reflecting diffusion with regime-switching model for its measures of performance. Furthermore, we justify its asymptotic optimality by deriving the stochastic fluid and diffusion limits for the corresponding system under heavy traffic. In addition, we identify a cost function related to the utility function, which is minimized by minimizing the workload process in the diffusion limit. More importantly, our queuing model includes typical systems in the future wireless networks, such as the J-user multi-input multioutput multiple access channel and the broadcast channel under Markov fading with cooperation and admission control as special cases.

Subject classifications: processor-sharing queues; random environment; MIMO wireless channel; Markov fading; doubly stochastic renewal process; utility-maximization scheduling; concave game; heavy traffic; asymptotic optimality; reflecting diffusion with regime-switching.
Area of review: Stochastic Models.
History: Received June 2010; revisions received December 2010, December 2011, August 2012, April 2013, June 2013; accepted September 2013.

1. Introduction

In current cellular systems, each base station is considered as a separate entity with no cooperation among base stations. Infrastructure cooperation among base stations, which has been proposed in several studies (e.g., Acampora et al. 2006, Kumaran and Viswanathan 2005, Viswanathan and Kumaran 2005), considers the base stations as one end of a multi-input multioutput (MIMO) system, and it has attracted considerable attention as a method for achieving high data rates over wireless links. In this paper, we study a J-user MIMO multiple access channel (MAC) uplink system and a J-user MIMO broadcast channel (BC) downlink system. Each of them can be considered as a cellular system with multiple users and multiple cooperating base station antennas: either multiple cooperating base stations each with a single antenna, a single-cell cellular system with a multiantenna base station, or a combination thereof. In either the MAC or the BC, data are buffered at the transmit end and the channel is time varying owing to multipath fading, a typical feature of wireless channels that increases the complexity of system design and performance analysis. We suppose that the fading process is a finite state continuous time Markov chain (FS-CTMC) whose discrete time version is widely used for modeling wireless channels (see, e.g., Wang and Moayeri 1995, Stolyar 2004, Viswanathan and Kumaran 2005, Dai and Wang 2009, and the references therein). Therefore, the J-user capacity regions of the MAC and BC are both time-varying set-valued stochastic processes driven by the FS-CTMC. Moreover, in each state of the Markov chain, it is well known that one can improve the capacity by cooperation; e.g., the sum of the rates at which data can be served for the J users is greater than the single-user capacity for any user (see, e.g., Bhardwaj et al. 2007). In addition, owing to the impact of the random environmental fading factor and the cooperative design, the service rates of the corresponding queuing system for the J users in the MAC or BC are random processes driven by the FS-CTMC.

Motivated by the observations described above, we consider a stochastic system of generalized processor-sharing queues in a random environment, where job arrivals to each queue follow a *doubly stochastic renewal process* (DSRP). Both the random environment and the random arrival rate of each DSRP are driven by an FS-CTMC. For such a queuing system, a reasonable online rate scheduling policy to minimize the average delay for a given load has not been established thus far, and exact solutions for the average delay are not available even for many simple policies. Therefore, simulation is often conducted for a meaningful evaluation of the system. In order to bridge the gap between dynamic rate scheduling and performance optimization of the system to some extent, we design a dynamic rate scheduling policy of Markov type by using the solution (a social optimal Nash equilibrium point) to an optimization problem that maximizes a general utility function over each of the randomly evolving capacity regions under the Karush-Kuhn-Tucker (KKT) optimality conditions (see, e.g., Luenberger 1984). Moreover, to overcome the intractability of performance evaluation for the system under the designed policy, we develop stochastic fluid and diffusion models by suitably scaling time and space and justifying related limit theorems for a heavily loaded queuing system operating under this policy. The limit models for queue lengths and workloads are a random process driven by, respectively the FS-CTMC and a reflecting diffusion with regime-switching model (i.e., a reflecting stochastic differential equation (SDE) with regime switching). In addition, we identify a cost function related to the utility function, which is minimized by minimizing the workload process in the diffusion limit. Hence, our policy is shown to be asymptotically optimal in deriving the stochastic fluid and diffusion limits.

Finally, in order to incorporate the J-user MIMO MAC and MIMO BC into our general queuing framework, we justify that the J-user capacity region for the MAC or the BC in any particular channel state is a convex set formed by a number of linear or smooth curved facets through applying the method of convex optimization, the implicit function theorem, and the duality of capacity regions between the MAC and the BC. Moreover, to realize the DSRP in the MAC or in the BC, we adopt a cross-layer design methodology to switch the arrival rates with the FS-CTMC channel fading process according to the current channel state information (CSI) through admission control.

Literature Review. The randomly evolving capacity region used to design our utility-maximization rate scheduling policy is a generalization of the so-called MIMO channel capacity region in the Shannon theoretic sense. For a single-user time-invariant channel, the Shannon capacity is defined as the maximum mutual information between input and output, which is shown by Shannon's capacity theorem to be the maximum data rate that can be transmitted over the channel with arbitrarily small error probability. For a J-user time-invariant MIMO channel, the corresponding capacity region is a J-dimensional set of all rate vectors $(c_1, \ldots, c_J)'$ simultaneously achievable by all J users. In particular, the region for the Gaussian MAC is a convex set that is the union of rate regions corresponding to every product input distribution satisfying the user-by-user power constraints (see, e.g., Goldsmith et al. 2003, Yu et al. 2004). The Gaussian BC differs from the Gaussian MAC in two fundamental aspects (see, e.g., Jindal et al. 2004). In the MAC, each transmitter has an individual power constraint, whereas in the BC, there is only a single power constraint on the transmitter. Moreover, signal and interference come from different transmitters in the MAC and are multiplied by different channel gains (known as the near-far effect) before being received, whereas in the BC, the entire received signal comes from the same source, and therefore it has the same channel gain. Nevertheless, the capacity region for the Gaussian BC can be obtained through the duality between the Gaussian MAC and the Gaussian BC (see, e.g., Jindal et al. 2004 and Goldsmith et al. 2003); i.e., it is the convex hull of the union over the set of capacity regions of the dual Gaussian MACs such that the total MAC power is the same as the power in the BC. Moreover, Liu and Hou (2008) have provided an analytical and numerical characterization in terms of the shape of the capacity boundaries for both the MAC and the BC.

However, in both the Gaussian MAC and the Gaussian BC, the exact characterization concerning piecewise smoothness of the capacity boundaries has not been determined thus far. Therefore, in order to apply our utility maximization rate scheduling algorithm to such wireless systems, we will present a more accurate analysis of the capacity regions. In addition, when the J-user MIMO channels are stochastic and time-varying fading ones, the capacity regions have multiple definitions (see, e.g., Goldsmith et al. 2003). Nevertheless, to capture the exact capacity region at each time instant for the MAC or the BC, we consider the capacity regions as a set-valued stochastic process evolving with the FS-CTMC rather than as a fixed one in an average sense, such as an ergodic capacity region (see, e.g., Goldsmith et al. 2003).

With regard to the scheduling algorithms, Acampora et al. (2006), Bhardwaj et al. (2007), and Bhardwaj and Williams (2009) considered a quasi-static downlink channel that is assumed to be fixed for all transmissions over the period of interest. In this case, the FS-CTMC and the random packet arrival rates assumed in the current paper reduce to constants. Moreover, without considering utility and cost optimization, Acampora et al. (2006), Bhardwaj et al. (2007), and Bhardwaj and Williams (2009) designed a simple rate scheduling policy of Markov type, which was shown to be throughput-optimal for a fixed convex capacity region by Acampora et al. (2006); additionally, a limit theorem was proved to justify the diffusion approximation of the queue length process for a heavily loaded system operating under their policy with two users by Bhardwaj et al. (2007) and with multiple users by Bhardwaj and Williams (2009). Their approximating model is a reflecting Brownian motion in the two-dimensional positive quadrant or in the general-dimensional positive orthant.

Stolyar (2004), Shakkotai et al. (2004), and Dai and Wang (2009) considered some scheduling policies for certain heavily loaded wireless systems with a finite-state discrete-time Markov fading process. In particular, a MaxWeight scheduling policy was considered by Stolyar

(2004) for a generalized switch. It was shown that the diffusively scaled workload processes converge to a one-dimensional RBM, and the MaxWeight policy asymptotically minimizes the workload under certain conditions. Moreover, an exponential scheduling rule was designed for wireless channels by Shakkotai et al. (2004) and for a generalized switch by Dai and Wang (2009), which was proved to be throughput-optimal and under which similar results concerning the workload process were obtained and justified as done by Stolyar (2004). In addition, Ye and Yao (2008) designed a utility-maximizing resource allocation policy for a class of stochastic networks with concurrent occupancy of resources and established its asymptotic optimality for the associated heavily loaded queuing system. Their policy covers the generalized $c\mu$-rule of Mandelbaum and Stolyar (2004) and the MaxWeight policy of Stolyar (2004) as special cases.

The current study differs from those of Stolyar (2004), Shakkotai et al. (2004), Dai and Wang (2009), and Ye and Yao (2008) in three aspects.

First, the scheduling policies of Stolyar (2004), Shakkotai et al. (2004), Dai and Wang (2009), and Ye and Yao (2008) depend only on a fixed capacity region that is a convex polyhedral, whereas ours depends on a time-varying and stochastic evolving capacity region process (a random environment) that at each time instant is a more general convex region rather than a convex polyhedral.

Second, the rates of packet arrivals to the J-user buffers are random processes rather than constants as used by Stolyar (2004), Shakkotai et al. (2004), Dai and Wang (2009), and Ye and Yao (2008). Thus, our input traffic to each user buffer is assumed to be a DSRP. Moreover, the tails of its interarrival time random variables are assumed to satisfy some integrability condition (see Dai 2011a and Dai 2012 for an interpretation). The particular case of our DSRP is the well-known doubly stochastic Poisson process (see, e.g., Brémaud 1981) that is very popular in modeling voice, video, and data source traffic in telecommunication systems, and it is called a Markovian modulated Poisson process (MMPP) or ON/OFF source.

Third, our discussion is based on a continuous time horizon rather than a discrete one, as in the case of Stolyar (2004), Shakkotai et al. (2004), and Dai and Wang (2009). Therefore, our vector-valued random service rate process depends on the FS-CTMC whose holding time at each environmental state has an important impact on the limiting processes; e.g., the limiting fluid model is a random process driven by the FS-CTMC rather than a deterministic function of time, and the limiting diffusion model is a more general RDRS rather than an RBM as derived by Stolyar (2004), Shakkotai et al. (2004), and Dai and Wang (2009). If one wants to directly generalize the studies of Stolyar (2004), Shakkotai et al. (2004), and Dai and Wang (2009) to the corresponding ones in a discrete-time random environment, a geometric distribution may be imposed on the holding time at each environmental state.

Finally, without considering optimal dynamic scheduling with utility/cost and performance optimizations, CTMCs have been used to model random environments in the studies of some queuing systems under certain static service disciplines (see, e.g., Choudhury et al. 1997 and the references therein for further details).

The remainder of the paper is organized as follows. In §2, we introduce our generalized processor-sharing queues under a random environment and we design our optimal rate scheduling policy. In §3, we introduce our heavy traffic condition and present our main asymptotic optimality theorem. In §4, we illustrate the use of our optimal policy and we present our main results in the J-user MIMO uplink and downlink wireless channels; we also present associated results concerning the piecewise smoothness of capacity boundaries of the J-user MIMO MAC and MIMO BC. In §5 and the e-companion to the paper (available as supplemental material at http://dx.doi.org/10.1287/opre.2013.1224), we prove our main theorem and associated lemmas.

2. Optimizing Processor-Sharing Queues Under Random Environment

2.1. Primitive Data

The queuing system under consideration is a stochastic system of generalized processor-sharing queues in a random environment evolving according to a stationary FS-CTMC $\alpha = \{\alpha(t), t \in [0, \infty)\}$. The process $\alpha(\cdot)$ takes values in a finite state space $\mathcal{K} \equiv \{1, \ldots, K\}$ with generator matrix $G = (g_{il})$ $(i, l \in \mathcal{K})$ and

$$g_{il} = \begin{cases} -\gamma(i) & \text{if } i = l, \\ \gamma(i)q_{il} & \text{if } i \neq l, \end{cases} \tag{1}$$

where $\gamma(i)$ is the holding rate for the chain in an environmental state $i \in \mathcal{K}$ and $Q = (q_{il})$ is the transition matrix of its embedded discrete-time Markov chain (see, e.g., Resnick 1992). Moreover, the queuing system has J queues in parallel, which correspond to J users for a given positive integer J. Each queue of infinite buffer capacity buffers packets (jobs) arrived for a given user. The queues can be served simultaneously by a single server (alternatively, J parallel servers) with rate allocation vector $c(t) = (c_1(t), \ldots, c_J(t))'$ that takes values in a time-varying and randomly evolving capacity set $\mathcal{R}(\alpha(t))$.

More precisely, for each state $i \in \mathcal{K}$, $\mathcal{R}(i)$ is a convex set that contains the origin and has L $(> J)$ boundary pieces. J of them are $(J-1)$-dimensional linear facets along the coordinate axes. The remaining ones are in the interior of R_+^J and form the so-called *capacity surface* denoted by $\mathcal{O}(i)$, which consists of $B = L - J$ (> 0) linear or smooth curved facets $h_k(c, i)$ on R_+^J for $k \in \mathcal{U} \equiv \{1, 2, \ldots, B\}$; i.e.,

$$\mathcal{R}(i) \equiv \{c \in R_+^J\colon\ h_k(c, i) \leqslant 0, k \in \mathcal{U}\}. \tag{2}$$

Figure 1. Two-user capacity set in the two-dimensional space in a particular environmental state.

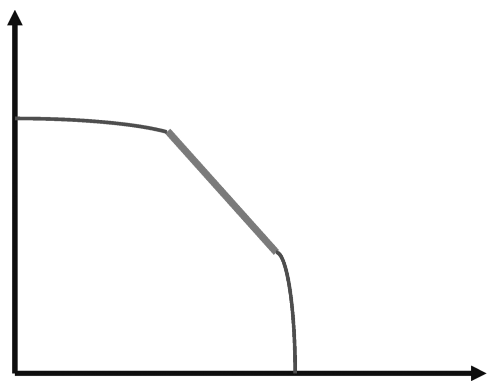
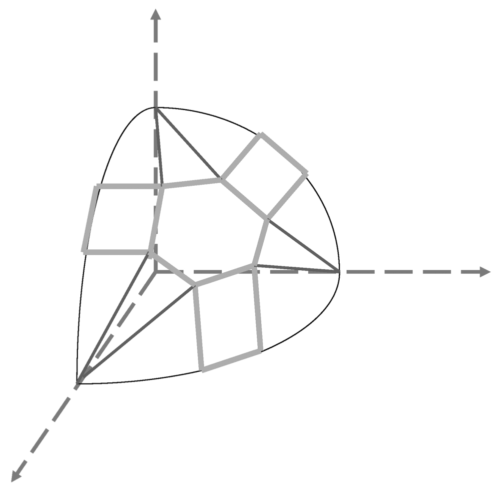


Figure 2. Three-user capacity set in the three-dimensional space in a particular environmental state.



Furthermore, let $C_U(i)$ denote the sum capacity upper bound for the capacity region. Then the facet in the center of the capacity surface is linear and is supposed to be a nondegenerate $(J-1)$-dimensional region, which can be expressed by

$$h_{k_U}(c, i) = \sum_{j=1}^{J} c_j - C_U(i), \tag{3}$$

where $k_U \in \mathcal{U}$ is the index corresponding to $C_U(i)$. In addition, we suppose that any one of the J linear facets along the coordinate axes forms a $(J-1)$-user capacity region corresponding to a particular group of $J-1$ users, who are the only users in the systems. Similarly, we can define the $(J-j)$-user capacity region for each $j \in \{2, \ldots, J-1\}$. Examples of such capacity sets in two- and three-dimensional spaces for a particular state $i \in \mathcal{K}$ are shown in Figures 1 and 2.

Now, let τ_n for all $n \in \{0, 1, \ldots\}$ denote the jump times of $\alpha(\cdot)$; i.e.,

$$\tau_0 \equiv 0, \quad \tau_n \equiv \inf\{t > \tau_{n-1}\colon \alpha(t) \neq \alpha(t^-)\}. \tag{4}$$

Then following from the description of a delayed renewal process (see, e.g., pages 174–175 of Resnick 1992), we can introduce the definition of DSRP as follows.

Definition 1. A process $A_j(\cdot)$ with $j \in \mathcal{J}$ is called a DSRP if $A_j(\tau_n + \cdot)$ is the counting process corresponding to a (conditional) delayed renewal process with arrival rate $\lambda_j(\alpha(\tau_n))$ and squared coefficient of variation $\alpha_j^2(\alpha(\tau_n)) \in (0, \infty)$ during time interval $[\tau_n, \tau_{n+1})$ for each $n \in \{0, 1, \ldots\}$.

Therefore, we suppose that there is a J-dimensional packet arrival process $A = \{A(t) = (A_1(t), \ldots, A_J(t))', t \geqslant 0\}$, where $A_j(t)$ with $j \in \mathcal{J} \equiv \{1, \ldots, J\}$ and $t \geqslant 0$ is the number of packets that arrive at the jth queue during $(0, t]$ and $A_j(\cdot)$ is assumed to be a DSRP. Note that here and elsewhere in the paper, the prime denotes the transpose of a vector or a matrix.

The packet interarrival times are assumed to be i.i.d. during the time interval corresponding to a specific environmental state $i \in \mathcal{K}$. Let $\{u_j(k), k = 1, 2, \ldots\}$ denote the sequence of times between the arrivals of the $(k-1)$th and the kth packets at the jth queue. Moreover, let $\{v_j(k), k = 1, 2, \ldots\}$ denote the sequence of packet lengths (in bits) for the successive arrivals at queue j, which is assumed to be a sequence of strictly positive i.i.d. random variables with average packet length $1/\mu_j \in (0, \infty)$ and squared coefficient of variation $\beta_j^2 \in (0, \infty)$. In addition, we suppose that all interarrival and service time processes are mutually (conditionally) independent when the environmental state is fixed. For each $j \in \mathcal{J}$ and each nonnegative constant h (in bits), we use $S_j(\cdot)$ to denote the renewal counting process associated with $\{v_j(k), k = 1, 2, \ldots\}$; i.e.,

$$S_j(h) = \sup\left\{n \geqslant 0\colon \sum_{k=1}^{n} v_j(k) \leqslant h\right\}. \tag{5}$$

Here, we note that according to the statistical analysis by Cao et al. (2001) and recent findings in Dai (2011a) and (2012), packet arrival processes and packet sizes can be approximated by DSRPs and i.i.d. variables, respectively.

2.2. Utility-Maximization Scheduling Algorithm and Queuing Dynamics

First, we remark that the service discipline used in this paper is the so-called head of line discipline; i.e., the service goes to the packet at the head of the line for a serving queue where packets are stored in the order of their arrivals. The service rates are determined by a function of the environmental state and the number of packets in each of the queues. At each state $i \in \mathcal{K}$ and for a given queue length vector $q = (q_1, \ldots, q_J)'$, let $\Lambda(q, i)$ denote the corresponding rate vector (in bps) of serving the J queues, which is a solution of the utility maximization problem

$$\max_{c \in \mathcal{R}(i)} \sum_{j \in \mathcal{J}} U_j(q_j, c_j), \tag{6}$$

where $c=(c_1,\dots,c_J)'$ is a J-dimensional vector and $U_j(q_j,c_j)$ for each $j\in\mathcal{J}$ is a utility function defined on R_+^J, which is second-order differentiable and satisfies the following conditions

$$U_j(0,c_j)=0, \tag{7}$$

$$U_j(q_j,c_j)=\Phi_j(q_j)\Psi(c_j) \text{ is strictly increasing and concave in } c_j \text{ for each } q_j>0, \tag{8}$$

$$\Psi(\nu_j c_j)=\Psi(\nu_j)\Psi(c_j) \quad\text{or}\quad \Psi(\nu_j c_j)=\Psi(\nu_j)+\Psi(c_j) \quad \text{for each nonnegative constant } \nu_j \tag{9}$$

$$\frac{\partial U_j(q_j,c_j)}{\partial c_j} \text{ is strictly increasing in } q_j\geqslant 0, \tag{10}$$

$$\frac{\partial U_j(0,c_j)}{\partial c_j}=0 \quad\text{and}\quad \lim_{q_j\to\infty}\frac{\partial U_j(q_j,c_j)}{\partial c_j}=+\infty \quad \text{for each } c_j>0. \tag{11}$$

Now, for a given $m\in\mathcal{J}$, let $k_j\in\mathcal{J}$ for each $j\in\{1,\dots,m\}$ and $k_j\neq k_l$ if $j\neq l$. Let $\mathcal{Q}(k_1,\dots,k_m)$ be the set of all $q\in R_+^J$ that have exactly m components q_{k_j} ($j\in\{1,\dots,m\}$) to be zero. Then from condition (8), we know that there must exist an optimal solution in the following form for a given $q\in\mathcal{Q}(k_1,\dots,k_m)$,

$$\Lambda_j(q,i)=\begin{cases}0 & \text{if } q_j=0,\\ \text{solution to the } (J-m)\text{-dimensional problem with all } q_j>0 \text{ in (6)–(11)} & \text{if } q_j>0.\end{cases} \tag{12}$$

Remark 1. For each $q>0$ (and similarly for a lower dimensional case), it follows from (8) that every point on the capacity surface defined in (2) is a Nash equilibrium point to a concave game in the sense of Rosen (1965). Therefore, the solution to (6) is a social optimal Nash equilibrium point to the concave game.

In addition, we assume that $\{U_j(q_j,c_j), j\in\mathcal{J}\}$ satisfies the so-called radial homogeneity condition; i.e., for any scalar $a>0$, each $q>0$ and each $i\in\mathcal{K}$, its maximizer satisfies

$$\Lambda_j(aq,i)=\Lambda_j(q,i). \tag{13}$$

Interested readers are referred to Ye and Yao (2008) for examples of the utility function that satisfies conditions (7)–(11) and (13), such as the so-called proportional fair allocation, minimal delay allocation, and (β,α)-proportionally fair allocation, which are widely used in communication protocols.

2.3. Dual Cost Minimization Problem

In this subsection, we consider the following cost minimization problem for each $i\in\mathcal{K}$, a given $c\in\mathcal{R}(i)$, and a given parameter $w\geqslant 0$:

$$\min_q \; V(q,c)$$
$$\text{s.t.} \; \sum_{j=1}^J \frac{q_j}{\mu_j}\geqslant w, \tag{14}$$
$$q_j\geqslant 0 \quad \text{for each } j\in\mathcal{J},$$

where the function V is defined by

$$V(q,c)=\sum_{j=1}^J C_j(q_j,c_j) \tag{15}$$

and C_j is the cost function associated with the utility function U_j in (6); i.e.,

$$C_j(q_j,c_j)=\frac{1}{\mu_j}\int_0^{q_j}\frac{\partial U_j(u,c_j)}{\partial c_j}\,du. \tag{16}$$

In other words, when the environment is in state $i\in\mathcal{K}$, we try to identify a queue state q corresponding to a given $c\in\mathcal{R}(i)$ and a given parameter $w\geqslant 0$ such that the total cost over the system is minimized and the (average) workload meets or exceeds w.

2.4. Performance Measure Processes

Let $Q_j(t)$ denote the queue length for the jth queue with $j\in\mathcal{J}$ at each time $t\in[0,\infty)$, and let $D_j(t)$ be the number of packet departures from the jth queue in $(0,t]$. Then

$$Q_j(t)=Q_j(0)+A_j(t)-D_j(t). \tag{17}$$

Furthermore, let $T_j(t)$ denote the cumulative amount of service (measured in bits) given to the jth queue up to time t; i.e.,

$$T_j(t)=\int_0^t \Lambda_j(Q(s),\alpha(s))\,ds. \tag{18}$$

Then we have that $D_j(t)=S_j(T_j(t))$. In addition, let $W(t)$ denote the (expected) workload at time t and $Y(t)$ denote the unused capacity up to time t; i.e.,

$$W(t)=\sum_{j=1}^J\frac{Q_j(t)}{\mu_j},$$
$$Y(t)=\sum_{j=1}^J\left(\int_0^t\rho_j(\alpha(s))\,ds-T_j(t)\right), \tag{19}$$

where for each $i\in\mathcal{K}$, $\rho(i)=(\rho_1(i),\dots,\rho_J(i))'$ is a given point on the capacity surface $\mathcal{O}(i)$.

3. Main Theorem: Asymptotic Optimality

In this section, we present the optimality result for our scheduling policy by considering the operation of the queueing system in the asymptotic regime where it is heavily loaded. More precisely, we define three sequences of diffusion-scaled processes $\hat{Q}^r(\cdot)$, $\hat{W}^r(\cdot)$, and $\hat{Y}^r(\cdot)$ by

$$\hat{Q}_j^r(t) \equiv \frac{Q_j^r(r^2 t)}{r}, \quad \hat{W}^r(t) \equiv \frac{W^r(r^2 t)}{r}, \quad \hat{Y}^r(t) \equiv \frac{Y^r(r^2 t)}{r} \tag{20}$$

for each $t \geqslant 0$, $j \in \mathcal{J}$, which is indexed by $r \in \mathcal{R}$ (a strictly increasing sequence of positive real numbers tending to infinity), and associate them with a sequence of independent Markov processes $\{\alpha^r(\cdot), r \in \mathcal{R}\}$. These systems indexed by $r \in \mathcal{R}$ all have the same basic structure as described in the last section, except the arrival rates $\lambda_j^r(i)$ and the holding time rates $\gamma^r(i)$ for all $i \in \mathcal{K}$, which may vary with $r \in \mathcal{R}$ and satisfy the following heavy traffic condition

$$r(\lambda_j^r(i) - \lambda_j(i)) \to \theta_j(i) \quad \text{as } r \to \infty, \quad \gamma^r(i) = \frac{\gamma(i)}{r^2} \tag{21}$$

for each $j \in \mathcal{J}$, where $\theta_j(i) \in R$ are some constants, and

$$\lambda_j(i) \equiv \mu_j \rho_j(i) \quad \text{with } \rho_j(i) = \nu_j \bar{\rho}_j(i), \tag{22}$$

$$\sum_{j=1}^J \nu_j = J, \quad \nu_j \geqslant 0 \text{ are constants for all } j \in \mathcal{J}, \tag{23}$$

$$\sum_{j=1}^J \bar{\rho}_j(i) = \max_{c \in \mathcal{R}(i)} \left(\sum_{j=1}^J c_j \right) = \mathcal{C}_U(i) \quad \text{and} \quad \bar{\rho}_1(i) = \cdots = \bar{\rho}_J(i). \tag{24}$$

From the equation in (3), $\bar{\rho}_j(i)$ with $j \in \mathcal{J}$, which satisfy the second condition in (24), can indeed be chosen. Moreover, if $\nu_j = 1$ for all $j \in \mathcal{J}$, the conditions in (22)–(24) imply that the system wishes to maximally and fairly allocate capacity to different users and the system design parameters $\lambda_j(i)$ for all $j \in \mathcal{J}$ are then determined through (22) in each channel state $i \in \mathcal{K}$. Furthermore, if the ratios of the nominal average packet arrival rates (e.g., estimates of the true average arrival rates) are preassumed (as in Bhardwaj et al. 2007, Bhardwaj and Williams 2009), the nominal service rate vector in each channel state $i \in \mathcal{K}$ for the J users can be determined via (22)–(24). In addition, the second condition in (24) and the separable conditions in (8)–(9) are technically useful in proving Lemmas 7–8. However, when only a constant environment (e.g., a pseudo channel in a wireless system) is concerned, the separable conditions in (8)–(9) can be removed.

Therefore, if the environment is in a state $\alpha^r(t) = i \in \mathcal{K}$ at time t, we can assume that the interarrival time corresponding to the kth arriving customer to the system indexed by $r \in \mathcal{R}$ is given as

$$u_j^r(k, i) = \frac{\hat{u}_j(k)}{\lambda_j^r(i)} \quad \text{for each } j \in \mathcal{J},\ k \in \{1, 2, \ldots\}, \tag{25}$$

where the $\hat{u}_j(k)$ do not depend on r and i; moreover, they have mean one and finite square coefficient of variation α_j^2. The packet length $\{v_j(k), k \in \{1, 2, \ldots\}\}$ is assumed not to change with r.

Note that from the heavy traffic condition in (21) for the rth environmental state process $\alpha^r(\cdot)$ with $r \in \mathcal{R}$, we know that $\alpha^r(r^2 \cdot)$ and $\alpha(\cdot)$ are equal to each other in distribution since they own the same generator matrix (see, e.g., the definition in pages 384–388 of Resnick 1992). Hence, in the sense of distribution, all of the systems indexed by $r \in \mathcal{R}$ in (20) share the same random environment over any time interval $[0, t]$.

Furthermore, let $B^E(\cdot)$ and $B^S(\cdot)$ denote the two independent J-dimensional standard Brownian motions, and for each $i \in \mathcal{K}$, let

$$\lambda(i) = (\lambda_1(i), \ldots, \lambda_J(i))', \tag{26}$$

$$\rho(i) = (\rho_1(i), \ldots, \rho_J(i))', \tag{27}$$

$$\theta(i) = (\theta_1(i), \ldots, \theta_J(i))', \tag{28}$$

$$\Gamma^E(i) = (\Gamma_{kl}^E(i))_{J \times J} \equiv \operatorname{diag}(\lambda_1(i)\alpha_1^2, \ldots, \lambda_J(i)\alpha_J^2), \tag{29}$$

$$\Gamma^S(i) = (\Gamma_{kl}^S(i))_{J \times J} \equiv \operatorname{diag}(\lambda_1(i)\beta_1^2, \ldots, \lambda_J(i)\beta_J^2), \tag{30}$$

$$H^e(t) = (H_1^e(t)', \ldots, H_J^e(t))' \quad \text{with } e \text{ denotes } E \text{ or } S, \tag{31}$$

$$H_j^e(t) = \int_0^t (\Gamma_{jj}^e(\alpha(s)))^{\frac{1}{2}} \, dB_j^e(s). \tag{32}$$

Then we have the following definition about RDRS.

Definition 2. A stochastic process $\hat{W}(\cdot)$ is called an RDRS if it can be represented uniquely as

$$\hat{W}(t) = \hat{X}(t) + \hat{Y}(s) \geqslant 0, \tag{33}$$

where

$$d\hat{X}(t) = \sum_{j=1}^J \frac{1}{\mu_j} (\theta_j(\alpha(t))dt + dH_j^E(t) + dH_j^S(t)). \tag{34}$$

Furthermore, $(\hat{W}(\cdot), \hat{Y}(\cdot))$ is the unique solution of (33) with the following complementary properties almost surely:

1. $\hat{Y}(0) = 0$,
2. $\hat{Y}(\cdot)$ is nondecreasing,
3. $\hat{Y}(\cdot)$ can increase only at a time $t \in [0, \infty)$ such that $\hat{W}(t) = 0$.

In addition, let $\hat{Q}^{r,G}(\cdot)$ and $\hat{W}^{r,G}(\cdot)$ denote the diffusion-scaled queue-length and workload processes under an arbitrarily feasible rate scheduling policy G, e.g., a simple Markovian policy as studied by Bhardwaj and Williams (2009) or a policy $\Lambda^G(Q^r(t), \alpha(t))$ that may not be the optimal solution to the utility maximization problem (6). Then we have the following theorem.

Theorem 1. *Suppose that $Q^r(0) = 0$ for all $r \in \mathcal{R}$ and the conditions* (21)–(25) *hold. Then under the scheduling*

policy (12), *we have the claims as stated in the following two parts*:
Part A: *Along* $r \in \mathcal{R}$, *the following convergence in distribution is true*:

$$(\hat{Q}^r(\cdot), \hat{W}^r(\cdot), \hat{Y}^r(\cdot)) \Rightarrow (\hat{Q}(\cdot), \hat{W}(\cdot), \hat{Y}(\cdot)) \tag{35}$$

and the limits $\hat{Q}(\cdot)$, $\hat{W}(\cdot)$ *and* $\hat{Y}(\cdot)$ *are continuous a.s., thereby satisfying the definition of an RDRS. In addition, we have*

$$\hat{Q}(t) = q^*(\hat{W}(t), \rho(\alpha(t))) \tag{36}$$

with $q^*(w, \rho(i))$ *being the solution to the cost minimization problem* (14) *in terms of each given* w *and* $i \in \mathcal{K}$.
Part B: *The workload* $\hat{W}(\cdot)$ *and the cost* $\sum_{j=1}^{J} C_j(\hat{Q}_j(\cdot), \rho(\alpha(\cdot)))$ *are minimal with probability one in the sense that for all* $t \geqslant 0$,

$$\liminf_{r\to\infty} \hat{W}^{r,G}(t) \geqslant \hat{W}(t), \tag{37}$$

$$\liminf_{r\to\infty} \sum_{j=1}^{J} C_j(\hat{Q}_j^{r,G}(t), \rho_j(\alpha(t))) \geqslant \sum_{j=1}^{J} C_j(\hat{Q}_j(t), \rho_j(\alpha(t))) \tag{38}$$

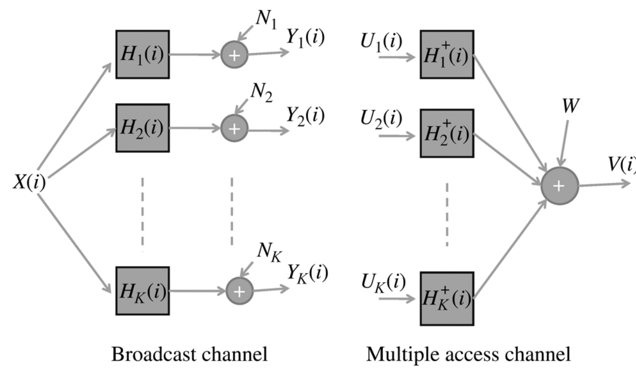
over some common supporting probability space.

Remark 2. As compared with the RBM widely studied in queuing literature, the RDRS model derived in (33) exhibits a new feature: it is a Markovian-modulated process. In the case of a constant environment (e.g., a quasi-static channel in a wireless system), the model in (33) reduces to an RBM. Moreover, from the discussions of Dai (1996), Dai and Dai (1999), and Harrison and Reiman (1981), we know that the unique solution $(\hat{W}(t), \hat{Y}(t))$ to (33) can be represented by $(\hat{W}, \hat{Y}) = (\Phi(\hat{X}), \Psi(\hat{X}))$, where $\Phi(\cdot)$ and $\Psi(\cdot)$ are Lipschitz continuous mappings. In addition, an RDRS is different from a conventional SDE since its drift and diffusion coefficients are not adapted to the filtration generated by the driving Brownian motions. Such SDEs without boundary reflections have attracted considerable attention in the field of financial engineering (see, e.g., Dai 2011b, Zhou and Yin 2003).

4. Applications to J-User MIMO Uplink and Downlink Wireless Channels

In this section, we apply the discussions from the previous sections to a cellular system where base stations cooperate among noise-free infinite-capacity links. We do not make any distinction between a single-cell cellular system having multiple base-station antennas and the traditional cellular system with cooperating single-antenna base stations. Here, cooperation means that the base stations can perform joint beamforming and/or power control, but there

Figure 3. BC and MAC channels in a particular environmental state.



is a constraint on the total power that the base stations can share. Therefore, our wireless system can be considered as a base station having M antennas and J users (mobiles), each of which has N antennas. Thus the uplink channel can be modeled as a J-user MIMO MAC and the downlink channel can be modeled as a J-user MIMO BC (see, e.g., Figure 3). The channel fading is supposed to obey the stationary FS-CTMC $\alpha = \{\alpha(t), t \in [0, \infty)\}$ that is described in the previous sections. Moreover, we suppose that the receive or transmit end (the cooperating base stations) has perfect CSI. For each channel state $i \in \mathcal{K}$, we let $H_j(i)$ ($j \in \mathcal{J} \equiv \{1, \ldots, J\}$) denote the downlink channel matrix from the base station to user j. Assuming that the same channel is used on the uplink and downlink, the uplink matrix of user j is $H_j^\dagger(i)$, which is the conjugate transpose of $H_j(i)$.

Furthermore, at the transmit end, arriving packets for each user are buffered before transmission and the rate of arrivals is a random process that switches with the FS-CTMC channel fading through admission control. Therefore, the processor-sharing queues presented in the previous section can be used to model the channel dynamics for both J-user MIMO MAC and J-user MIMO BC. The remaining issue is the characterization of the MAC and BC capacity region processes, which is also a central topic in information theory literature.

4.1. MIMO MAC Capacity Region

In the MAC and for each channel state $i \in \mathcal{K}$, let $U_j(i) \in \mathcal{C}^{N\times 1}$ be the transmitted signal of user j, where $\mathcal{C}^{N\times 1}$ denotes the $N \times 1$ complex matrix. Let $V(i) \in \mathcal{C}^{M\times 1}$ denote the received signal, and let $W \in \mathcal{C}^{M\times 1}$ denote the noise vector where $W \sim \bar{N}(0, I)$ is circularly symmetric complex Gaussian with identity covariance (note that the notation W here has a different meaning from the workload process $W(t)$ defined in (19)). Then the received signal at the base station is given by

$$V(i) = H^\dagger(i)U'(i) + W, \tag{39}$$

where $H^\dagger(i) = [H_1^\dagger(i), \ldots, H_J^\dagger(i)]$ and $U(i) = [U_1'(i), \ldots, U_J'(i)]$ (see, e.g., Figure 3). Moreover, each user j is subject to an individual power constraint P_j. The transmit covariance matrix of user j is defined as $\Gamma_j(i) \equiv E[U_j(i)U_j^\dagger(i)]$. The power constraint implies that $\mathrm{Tr}(\Gamma_j(i)) \leqslant P_j$ for $j \in \mathcal{J}$. During each channel state $i \in \mathcal{K}$, it follows from Goldsmith et al. (2003) and Yu et al. (2004) that the MAC capacity region is a J-dimensional closed convex set in $R_+^J \equiv \{c \in R^J\colon c_j \geqslant 0, j \in \mathcal{J}\}$; i.e.,

$$\begin{aligned}\mathcal{R}(i) &= \mathcal{C}_{\mathrm{MAC}}(P_1, \ldots, P_J, H^\dagger(i))\\ &= \bigcup_{\{\Gamma_j(i)\geqslant 0,\, \mathrm{Tr}(\Gamma_j(i))\leqslant P_j,\, j\in\mathcal{J}\}} \Bigg\{c \in R_+^J\colon \sum_{j\in S} c_j \\ &\qquad \leqslant \frac{1}{2}\log\left|I + \sum_{j\in S} H_j^\dagger(i)\Gamma_j(i)H_j(i)\right|, \forall S \subset \mathcal{J}\Bigg\}, \end{aligned} \tag{40}$$

where S is a subset of $\mathcal{J}$ and $|\cdot|$ denotes the determinant of a matrix. Moreover, $\Gamma_j(i) \geqslant 0$ denotes a Hermitian matrix that is positive semidefinite. In addition, every point in $\mathcal{R}(i)$ can be achieved by Shannon's source coding theorem and successive decoding (see, e.g., Gamal and Cover 1980 and Goldsmith et al. 2003). However, in designing a utility maximization-based rate scheduling policy, we need to know a more detailed boundary characterization of the MAC capacity region since it frequently relies on the KKT optimality conditions (see, e.g., Luenberger 1984 and Liu and Hou 2008). Thus, we have the following lemma.

Lemma 1. *For the J-user MIMO MAC and each channel state $i \in \mathcal{K}$, $\mathcal{R}(i)$ contains the origin and has L linear or smooth curved facets with L given by*

$$L = J! + \sum_{j=2}^{J} C_J^j (J - j + 1)! + J. \tag{41}$$

Moreover, J of these pieces are $(J-1)$-dimensional linear facets along the coordinate axes, and the remaining $B = L - J$ ones are in the interior of R_+^J and form $\mathcal{O}(i)$, which are linear or smooth curved facets $h_k(c, i)$ on R_+^J for $k \in \mathcal{U} \equiv \{1, 2, \ldots, B\}$; i.e.,

$$\mathcal{R}(i) \equiv \{c \in R_+^J\colon h_k(c, i) \leqslant 0, k \in \mathcal{U}\}. \tag{42}$$

In addition, if $C_{\mathrm{MAC}}(P, H(i))$ is used to denote the sum capacity upper bound for the MAC capacity region, then

$$h_{k_{\mathrm{MAC}}}(c, i) = \sum_{j=1}^{J} c_j - C_{\mathrm{MAC}}(P, H(i)), \tag{43}$$

where $k_{\mathrm{MAC}} \in \mathcal{U}$ is the index corresponding to $C_{\mathrm{MAC}}(P, H(i))$.

Proof of Lemma 1. For reader's convenience, we only outline the proof of the lemma; the reader can find the detailed proof in the electronic companion (e-companion) to the paper.

First, we use the optimization technique studied by Goldsmith et al. (2003) and Yu et al. (2004) to characterize the boundary of the MAC capacity region presented in (40); i.e., the region in (40) is convex, and thus its boundary can be fully characterized by maximizing the function $\sum_{j=1}^J \nu_j r_j$ over all rate vectors in the region and for all nonnegative priority vectors $\nu = (\nu_1, \ldots, \nu_J)$ such that $\sum_{j=1}^J \nu_j = 1$. Then based on priority vectors and permutation schemes, we can determine the number of boundary pieces of the region, which is consistent with that obtained by Liu and Hou (2008). Finally, by applying the KKT optimality conditions and the implicit function theorem, we can prove that the boundary of the MAC capacity region consists of the derived number of linear or smooth curved facets. □

Example 1. For the MAC channel and each $i \in \mathcal{K}$, when $J = 2$ and $N = 1$ (i.e., each of the user's mobiles has only a single transmit antenna), it follows from Goldsmith et al. (2003) that

$$\begin{aligned} g_1(c, i) &= c_1 - \log|I + H_1^\dagger(i)P_1H_1(i)|,\\ g_2(c, i) &= c_1 + c_2 - \log|I + H_1^\dagger(i)P_1H_1(i) + H_2^\dagger(i)P_2H_2(i)|,\\ g_3(c, i) &= c_2 - \log|I + (I + H_1^\dagger(i)P_1H_1(i))^{-1}H_2^\dagger(i)P_2H_2(i)|. \end{aligned}$$

4.2. MIMO BC Capacity Region

In the MIMO BC and for each channel state $i \in \mathcal{K}$, let $X(i) \in \mathcal{C}^{M\times 1}$ denote the transmitted vector signal from the base station and let $Y_j(i) \in \mathcal{C}^{N\times 1}$ be the received signal by the user j. The noise at user j is represented by $N_j \in \mathcal{C}^{N\times 1}$ and is assumed to be circularly symmetric complex Gaussian noise ($N_j \sim N(0, I)$). The received signal of user j (see, e.g., Figure 3) is given by

$$Y_j(i) = H_j(i)X(i) + N_j.$$

The transmit covariance matrix of the input signal is $\Gamma_X(i) \equiv E[X(i)X^\dagger(i)]$. The base station is subject to an average power constraint, which implies that $\mathrm{Tr}(\Gamma_X(i)) \leqslant P$. During each channel state $i \in \mathcal{K}$, the J-user MIMO BC capacity region denoted by $\mathcal{R}(i)$ can be calculated by the duality of the MAC and the BC, as shown by Jindal et al. (2004) and Goldsmith et al. (2003), where the BC capacity region is obtained by taking the convex hull of the union over the set of capacity regions of the dual MIMO MACs such that the total MAC power is the same as the power in the BC; i.e.,

$$\begin{aligned}\mathcal{R}(i) &= \mathcal{C}_{BC}(P, H(i))\\ &= \bigcup_{\{(P_1,\ldots,P_J)\colon \sum_{j=1}^J P_j = P\}} \mathcal{C}_{\mathrm{MAC}}(P_1, \ldots, P_J, H^\dagger(i)). \end{aligned}\tag{44}$$

Moreover, the dirty paper coding (DPC) proposed by Costa (1983) achieves the capacity for the MIMO BC (see, e.g., Weingarten et al. 2006). In particular, if each user has only a single receive antenna, we have the following lemma.

Lemma 2. *For the J-user MIMO BC with $N = 1$ and each $i \in \mathcal{K}$, and L given in* (41), *$\mathcal{R}(i)$ contains the origin and has L boundary pieces. J of them are $(J-1)$-dimensional linear facets along the coordinate axes, and the remaining $B = L - J$ ones are in the interior of R_+^J and form $\mathcal{O}(i)$, which are linear or smooth curved facets $h_k(c, i)$ on R_+^J for $k \in \mathcal{U} \equiv \{1, 2, \ldots, B\}$; i.e.,*

$$\mathcal{R}(i) \equiv \{c \in R_+^J \colon h_k(c, i) \leqslant 0, k \in \mathcal{U}\}. \tag{45}$$

Moreover, if $C_{\text{Sato}}(P, H(i))$ denotes the sum capacity upper bound (called the Sato upper bound) of the BC capacity region, then

$$h_{k_{\text{Sato}}}(c, i) = \sum_{j=1}^{J} c_j - C_{\text{Sato}}(P, H(i)), \tag{46}$$

where $k_{\text{Sato}} \in \mathcal{U}$ is the index corresponding to $C_{\text{Sato}}(P, H(i))$.

Proof of Lemma 2. For the proof of Lemma 2, we use the duality of the capacity regions between MAC and BC to transform the discussion for BC to one for MAC (see, e.g., Goldsmith et al. 2003). Readers can find the detailed proof in the e-companion to the paper. □

Remark 3. The concept of *Sato upper bound* of the J-user MIMO BC can be found in Goldsmith et al. (2003) and the references therein. Moreover, it is interesting to know whether the corresponding property stated in the lemma is true if each user has multiple receive antennas in the MIMO BC channel.

Example 2. Considering the BC channel with $J = 2$ and $N = 1$ for each $i \in \mathcal{K}$, we can derive $h_k(c, i)$ explicitly for $k = 1, 2, 3$ by employing the results of Goldsmith et al. (2003), Vishwanath et al. (2003), and Weingarten et al. (2006) as follows:

$$h_1(c, i) = e^{2(c_1+c_2)} - \Bigg| I + (H_1^\dagger(i)H_1(i) - H_2^\dagger(i)H_2(i)) \cdot \frac{e^{2c_1} - 1}{|H_{111}|^2(i) + |H_{122}|^2(i)} + H_2^\dagger(i)H_2(i)P \Bigg|^2, \tag{47}$$

$$h_2(c, i) = c_1 + c_2 - \mathcal{C}_{\text{Sato}}(P, H(i)), \tag{48}$$

$$h_3(c, i) = e^{2(c_1+c_2)} - \Bigg| I + (H_2^\dagger(i)H_2(i) - H_1^\dagger(i)H_1(i)) \cdot \frac{e^{2c_2} - 1}{|H_{211}|^2(i) + |H_{222}|^2(i)} + H_1^\dagger(i)H_1(i)P \Bigg|^2. \tag{49}$$

5. Proof of Theorem 1

We here note that for the reader's convenience, we provide an outline to summarize the lengthy proof of Theorem 1. Owing to length limitations, we put the full proof in the e-companion to the paper. In the remainder of this main body to the paper, we introduce key lemmas and main techniques that are used in the proof of Theorem 1. Most of their detailed technical proofs are also presented in the e-companion to the paper.

5.1. Preliminary Lemmas on the Utility-Maximization and Dual Cost Minimization Problems

Lemma 3. *Consider the utility-maximization problem in* (6) *and suppose that conditions* (7)–(11) *and* (13) *are imposed; then for a sequence of queue states, $\{q^l, l \in \mathcal{R}\}$, which satisfies $q^l \to q \in R_+^J$ as $l \to \infty$, we have*

$$\Lambda_j(q^l, i) \to \Lambda_j(q, i) \quad \textit{as } l \to \infty \tag{50}$$

for each $i \in \mathcal{K}$ and any $j \in \mathcal{J}$ such that $q_j > 0$.

Proof of Lemma 3. Consider each specific state $i \in \mathcal{K}$. The proof can be accomplished similarly as in the case of Lemma 6.2 of Ye et al. (2005), and hence, we omit it. □

Lemma 4. *For each state $i \in \mathcal{K}$, the following claims are true.*

1. *Under the policy in* (12) *and the associated convention, if $c^* = \rho(i)$ is an optimal solution to the maximization problem in* (6) *with a queue state q in the utility function, then $q^* = q$ must be the optimal solution to the minimization problem in* (14) *with $c = c^*$ in the cost function and with $w = \sum_{j=1}^J q_j^*/\mu_j$ in the constraints, i.e., $q^*(w, \rho(i)) = q^*$.*
2. *Conversely, if q^* is the optimal solution to the minimization problem in* (14) *with $w > 0$ and $\Lambda(q^*, i) = \rho(i)$ for each $i \in \mathcal{K}$ in the cost function, i.e., $q^*(w, \rho(i)) = q^*$, then $q^* > 0$ and $\Lambda^*(q^*, i) = \Lambda(q^*, i)$ must be an optimal solution to the maximization problem in* (6) *with $q = q^*$ in the utility function.*

The proof of Lemma 4 is provided in the e-companion to the paper. Now let $\|\cdot\|$ denote the norm of a vector $q \in R^J$ in the sense that $\|q\| = \sum_{j=1}^J |q_j|$. Then we have the following lemma.

Lemma 5. *For each state $i \in \mathcal{K}$, the following claims are true.*

1. *The cost minimization problem* (14) *has a unique optimal solution $q^*(w, \rho(i))$ when $c = \rho(i)$ is in the cost function for each $i \in \mathcal{K}$. Moreover, $q^*(w, \rho(i))$ is continuous in terms of w.*
2. *Assuming that for any given constant $\epsilon > 0$, there exists another constant $\sigma > 0$ that depends only on ϵ, such that for any $q \in \mathcal{V}(\epsilon, \sigma, i)$ with*

$$\mathcal{V}(\epsilon, \sigma, i) \equiv \Bigg\{ q \in R_+^J \colon \|q - q^*(w, \rho(i))\| \leqslant \sigma \textit{ and } w = \sum_{j=1}^J \frac{1}{\mu_j} q_j \geqslant \epsilon \Bigg\}, \tag{51}$$

we have

$$\sum_{j=1}^J \Lambda_j(q, i) = \sum_{j=1}^J \rho_j(i). \tag{52}$$

The proof of Lemma 5 will be provided in the e-companion to the paper. Here we note that the unique optimal solution $q^*(w, \rho(i))$ will be referred to as a *fixed point* in the following discussion.

5.2. Functional Central Limit Theorem with Regime Switching

From the second condition in (21), we know that the processes $\alpha^r(r^2\cdot)$ for each $r\in\mathcal{R}$ and $\alpha(\cdot)$ are equal in distribution. Thus, without loss of generality, we can suppose that

$$\alpha^r(r^2t)=\alpha(t)\quad\text{for each } r\in\mathcal{R} \text{ and } t\in[0,\infty). \tag{53}$$

Now define

$$E_j^r(\cdot)\equiv A_j^r(r^2\cdot), \tag{54}$$

$$\bar{T}_j^r(\cdot)\equiv\int_0^{\cdot}\Lambda_j(\bar{Q}^r(s),\alpha(s))\,ds=\frac{1}{r^2}T_j^r(r^2\cdot), \tag{55}$$

$$\bar{Q}_j^r(t)\equiv\frac{1}{r^2}Q_j^r(r^2t), \tag{56}$$

where we have used the radial homogeneity of $\Lambda(q,i)$ in (13) for (55). Thus, it follows from (17), (53), and the assumptions among the arrival and service processes that

$$\hat{Q}_j^r(\cdot)=\frac{1}{r}E_j^r(\cdot)-\frac{1}{r}S_j^r(\bar{T}_j^r(\cdot)). \tag{57}$$

Furthermore, let

$$\hat{E}^r(\cdot)=(\hat{E}_1^r(\cdot),\ldots,\hat{E}_J^r(\cdot))'\quad\text{with } \hat{E}_j^r(\cdot)=\frac{1}{r}(A_j^r(r^2\cdot)-r^2\bar{\lambda}_j^r(\cdot)), \tag{58}$$

$$\hat{S}^r(\cdot)=(\hat{S}_1^r(\cdot),\ldots,\hat{S}_J^r(\cdot))'\quad\text{with } \hat{S}_j^r(\cdot)=\frac{1}{r}(S_j(r^2\cdot)-\mu_jr^2\cdot) \tag{59}$$

for each $j\in\{1,\ldots,J\}$ with

$$\bar{\lambda}_j^r(\cdot)\equiv\int_0^{\cdot}\lambda_j^r(\alpha(s))\,ds=\int_0^{\cdot}\lambda_j^r(\alpha^r(r^2s))\,ds=\frac{1}{r^2}\int_0^{r^2\cdot}\lambda_j^r(\alpha^r(s))\,ds. \tag{60}$$

In addition, define

$$\bar{\lambda}^r(\cdot)=(\bar{\lambda}_1^r(\cdot),\ldots,\bar{\lambda}_J^r(\cdot))'. \tag{61}$$

Then we have the following lemma.

Lemma 6. *For the diffusion-scaled processes in (58)–(59), the following convergence in distribution is true as $r\to\infty$; i.e.,*

$$(\hat{E}^r(\cdot),\hat{S}^r(\cdot))\Rightarrow(H^E(\cdot),(\Gamma^B)^{1/2}B^S(\cdot)), \tag{62}$$

where $\Gamma^B=\mathrm{diag}(\mu_1\beta_1^2,\ldots,\mu_J\beta_J^2)$.

The main idea used in proving Lemma 6 stems from the related discussion in Dai (1996) and Dai and Dai (1999), and the concrete proving techniques include the conventional functional central limit theorem (see, e.g., Iglehart and Whitt 1971 and Prokhorov 1956), random change of time lemma (see, e.g., Billingsley 1999), establishment of oscillation inequality (see, e.g., Dai 1996, Dai and Dai 1999), equivalent conditions of relative compactness and Skorohod representation theorem (see, e.g., Ethier and Kurtz 1986), etc. However, owing to its length, the complete proof of Lemma 6 is provided in the e-companion to the paper.

5.3. Fluid Limiting Processes

For each $j\in\mathcal{J}$, $t\geqslant 0$ and $r>0$, we define the fluid-scaled processes,

$$\begin{aligned}\bar{W}^r(t)&\equiv\frac{1}{r^2}W^r(r^2t),\quad \bar{Y}^r(t)=\frac{1}{r^2}Y^r(r^2t),\\ \bar{E}_j^r(t)&\equiv\frac{1}{r^2}E_j^r(t),\quad \bar{S}_j^r(t)\equiv\frac{1}{r^2}S_j^r(r^2t)\end{aligned} \tag{63}$$

and use $\bar{Q}^r(\cdot)$, $\bar{E}^r(\cdot)$, $\bar{S}^r(\cdot)$, $\bar{T}^r(\cdot)$ to denote the corresponding vector processes. Further, let

$$\bar{Q}_j(t)=\bar{Q}_j(0)+\bar{\lambda}_j(t,\zeta_t(\cdot))-\mu_j\bar{T}_j(t)\quad\text{for each } j\in\mathcal{J}, \tag{64}$$

$$\bar{W}(t)=\sum_{j=1}^{J}\frac{\bar{Q}_j(t)}{\mu_j}=\bar{W}(0)+\bar{Y}(t), \tag{65}$$

$$\bar{Y}(t)=\sum_{j=1}^{J}\left(\int_0^t\rho_j(\alpha(s))\,ds-\bar{T}_j(t)\right), \tag{66}$$

$$\bar{\lambda}(t)=(\bar{\lambda}_1(t),\ldots,\bar{\lambda}_J(t))',\quad \bar{\lambda}_j(t)\equiv\int_0^t\lambda_j(\alpha(s))\,ds, \tag{67}$$

$$\bar{T}_j(t)=\int_0^t\bar{\Lambda}_j(\bar{Q}(s),\alpha(s))\,ds, \tag{68}$$

where it follows from (12) that for each $i\in\mathcal{K}$,

$$\bar{\Lambda}_j(q,i)=\begin{cases}\Lambda_j(q,i) & \text{if } q_j>0,\\ \rho_j(i) & \text{if } q_j=0.\end{cases} \tag{69}$$

Thus, we have the following lemma.

Lemma 7. *Suppose $\bar{Q}^r(0)\Rightarrow\bar{Q}(0)$ as $r\to\infty$. Then under the utility-maximization allocation policy $\Lambda(q,i)$ in (12), any subsequence of $\mathcal{R}$ has a further subsequence $\{r_l, l=1,2,\ldots\}$ such that the following convergence in distribution is true,*

$$(\bar{E}^{r_l}(\cdot),\bar{S}^{r_l}(\cdot),\bar{T}^{r_l}(\cdot),\bar{Q}^{r_l}(\cdot),\bar{W}^{r_l}(\cdot),\bar{Y}^{r_l}(\cdot))\Rightarrow(\bar{E}(\cdot),\bar{S}(\cdot),\bar{T}(\cdot),\bar{Q}(\cdot),\bar{W}(\cdot),\bar{Y}(\cdot)) \tag{70}$$

as $l\to\infty$, where the limit in (70) satisfies (64)–(69). Moreover, if $\bar{Q}(0)=0$, the convergence in (70) is true along the whole sequence $r\in\mathcal{R}$ with the limit satisfying

$$\bar{E}(t)=\bar{\lambda}(t),\quad \bar{S}(t)=\mu(t),\quad \bar{T}(t)=\bar{c}(t), \tag{71}$$

$$\bar{Q}_j(t)=\bar{W}(t)=\bar{Y}(t)=0, \tag{72}$$

for each $t\geqslant 0$ and $j\in\mathcal{J}$, where $\mu(t)\equiv(\mu_1,\ldots,\mu_J)'t$, and

$$\bar{c}(t)=(\bar{c}_1(t),\ldots,\bar{c}_J(t))',\quad \bar{c}_j(t)\equiv\int_0^t\rho_j(\alpha(s))\,ds. \tag{73}$$

The proof of Lemma 7 is provided in the e-companion to the paper owing to its length. Now since $y_j = C_j(q_j, \rho_j(i))$ is strictly increasing in q_j for each $i \in \mathcal{K}$ and $j \in \mathcal{J}$, its inverse $C_j^{-1}(y_j, \rho_j(i))$ is well defined and is strictly increasing in y_j. Hence, for each $\kappa \geqslant 0$, we can define

$$\tilde{g}(\kappa) \equiv \sum_{i=1}^{K} \max_{\|q\| \leqslant \kappa} \psi(q, i) \quad \text{and}$$
$$g(\kappa) \equiv \sum_{i=1}^{K} \sum_{j=1}^{J} C_j^{-1}(\tilde{g}(\kappa), \rho_j(i)). \tag{74}$$

Then we have the following lemma.

Lemma 8. *Under the same conditions as these used in Lemma 7, if $\|\bar{Q}(0)\| \leqslant \chi$ for some constant χ, $\bar{Q}(t)$ is bounded for each $t \geqslant 0$; i.e.,*

$$\|\bar{Q}(t)\| \leqslant g(\chi) \quad \text{for each } t \geqslant 0. \tag{75}$$

Furthermore, there exists a time $T_{\chi,\epsilon} > 0$ for any given $\epsilon > 0$ such that

$$\|\bar{Q}(t) - q^*(\bar{W}(t), \rho(\alpha(t)))\| < \epsilon \quad \text{for all } t \geqslant T_{\chi,\epsilon} \tag{76}$$

and in particular, if $\bar{Q}(0) = q^*(\bar{W}(0), \rho(\alpha(0)))$, then $\bar{Q}(t) = \bar{Q}(0)$ a.s. for all $t \in [\tau_0, \tau_1)$.

The proof of Lemma 8 is provided in the e-companion to the paper owing to its length. Instead, we remark here that our fluid limit derived in Lemma 7 and Lemma 8 is a random process driven by the FS-CTMC rather than a deterministic function of time as obtained in the existing studies. This new feature increases the complexity of proving the two lemmas; e.g., as compared with the study in Ye and Yao (2008), it requires more technical treatment in handling the FS-CTMC based jumps for the constructed Lyapunov function.

5.4. A Key Lemma on Finer Time-Scaling

It follows from (20), (19), (17), (54)–(60), and the similar argument as for (57) that

$$\hat{W}^r(\cdot) = \hat{X}^r(\cdot) + \hat{Y}^r(\cdot), \tag{77}$$

where for each $t \geqslant 0$,

$$\hat{Y}^r(t) = r \sum_{j=1}^{J} \left(\int_0^t \rho_j(\alpha(s))\, ds - \bar{T}_j^r(t) \right), \tag{78}$$

which is nondecreasing in $t \geqslant 0$ owing to (55), (18), (22)–(24). Moreover, from Lemma 6, we know that $(\hat{E}^r(\cdot), \hat{S}^r(\cdot))$ is C-tight. Thus, by a similar proof to that for (70), the convergence in (70) is still true along $r \in \mathcal{R}$ if $\hat{Q}^r(0) \equiv 0$ and $(\bar{E}^r(\cdot), \bar{S}^r(\cdot))$ on the left-hand side of (70) is replaced by $(\hat{E}^r(\cdot), \hat{S}^r(\cdot))$. Then it follows from the random change of time lemma (see, e.g., page 151 of Billingsley 1999) that

$$\begin{aligned}\hat{X}^r(t) &= \sum_{j=1}^{J} \frac{1}{\mu_j} (\hat{E}_j^r(t) - \hat{S}_j^r(\bar{T}_j^r(t))) \\ &\quad + \sum_{j=1}^{J} r \int_0^t (\rho_j^r(\alpha(s)) - \rho_j(\alpha(s)))\, ds \\ &\Rightarrow \hat{X}(t) \quad \text{as } r \to \infty,\end{aligned} \tag{79}$$

where $\rho_j^r(\alpha(\cdot)) = \lambda_j^r(\alpha(\cdot))/\mu_j$ and $\hat{X}(\cdot)$ is given by (34). Since $\hat{X}(\cdot)$ is a continuous process, it follows from the Skorohod representation theorem that the convergence in (79) can be assumed uniformly on all compact sets (u.o.c.). Moreover, the common supporting probability space can be chosen such that the convergence in (70) together with the convergence of $(\bar{E}^r(\cdot), \bar{S}^r(\cdot))$ corresponding to $\alpha(\cdot) \equiv i$ for each fixed $i \in \mathcal{K}$ are u.o.c. a.s. Therefore, in the remainder of this subsection, we will only consider an arbitrarily given sample path for which the above u.o.c. convergence holds.

Now for a time $\tau \geqslant 0$, a constant $\delta > 0$, a sufficiently large integer r, and a fixed time $T > 0$ of certain magnitude to be specified later, we divide the time interval $[\tau, \tau + \delta]$ into a total of $\lceil r\delta/T \rceil - 1$ segments with equal length T/r, except the last one, where $\lceil \cdot \rceil$ denotes the integer ceiling. The lth segment with $l \in \{0, 1, \ldots, \lceil r\delta/T \rceil - 1\}$ covers the time interval $[\tau + lT/r, (\tau + (l+1)T/r) \wedge T]$. Then any $t \in [\tau, \tau + \delta]$ has an equivalent transformation defined by

$$t = \tau + (lT + u)/r \equiv \eta^{r,l}(u) \tag{80}$$

for some $l \in \{0, 1, \ldots, \lceil r\delta/T \rceil - 1\}$ and $u \in [0, T]$. Hence, $\hat{W}^r(t)$ in (77) can be considered to have the form

$$\begin{aligned}\hat{W}^r(t) &= \frac{1}{r} W^r(r^2 t) = \frac{1}{r} W^r((r^2\tau + rlT) + ru) \\ &\equiv \bar{W}^{r,l}(u)\end{aligned} \tag{81}$$

for each $l \in \{0, 1, \ldots, \lceil r\delta/T \rceil - 1\}$ and $u \in [0, T]$. In other words, for each time point, we will study the behavior of $\hat{W}^r(t)$ through the fluid process $\bar{W}^{r,l}(u)$ over the time interval $[0, T]$ (see, e.g., Ye and Yao 2008 and references therein). Similarly, we can define $\bar{Q}^{r,l}(u)$ and $\bar{Y}^{r,l}(u)$ through $\hat{Q}^r(t)$ and $\hat{Y}^r(t)$. Furthermore, let

$$\bar{T}_j^{r,l}(u) = \frac{1}{r^2} \int_0^{r^2\tau + rlT + ru} \Lambda_j(\bar{Q}^r(s), \alpha(s))\, ds. \tag{82}$$

Then we have the following lemma.

Lemma 9. *Considering a given sample path, suppose that $\bar{Q}^{r,l_r}(0)$ is bounded for any given sequence $\{l_r, r \in \mathcal{R}\} \subseteq \{0, 1, \ldots\}$ and $\alpha(\cdot) \equiv i$ over $[0, T]$ for a fixed $i \in \mathcal{K}$. Then for any given subsequence $\mathcal{R}' \subseteq \mathcal{R}$, there is a further subsequence $\mathcal{R}'' \subseteq \mathcal{R}'$ such that the following u.o.c. convergence over $[0, T]$ is true as $r \to \infty$ along $r \in \mathcal{R}''$:*

$$\begin{aligned}&(\bar{E}^{r,l_r}(\cdot), \bar{S}^{r,l_r}(\cdot), \bar{T}^{r,l_r}(\cdot), \bar{Q}^{r,l_r}(\cdot), \bar{W}^{r,l_r}(\cdot), \bar{Y}^{r,l_r}(\cdot)) \\ &\quad \to (\bar{E}(\cdot), \bar{S}(\cdot), \bar{T}(\cdot), \bar{Q}(\cdot), \bar{W}(\cdot), \bar{Y}(\cdot)),\end{aligned} \tag{83}$$

where the limit in (83) satisfies (64)–(69) over $[0, T]$.

Proof of Lemma 9. For any given subsequence $\mathscr{R}' \subseteq \mathscr{R}$, it follows from an argument similar to that in the proof of Lemma 7 and the definition in (82) that the sequence $\bar{T}^{r,l_r}(\cdot)$ is uniformly Lipschitz continuous over $[0,T]$. Furthermore, since $(\bar{E}^{r,l_r}(\cdot), \bar{S}^{r,l_r}(\cdot))$ converges u.o.c over $[0,T]$ along $\{l_r, r \in \mathscr{R}'\}$ and the given path, it follows from the bounded property of $\bar{Q}^{r,l_r}(0)$ and the similar proof for Lemma 7 that the claim in (83) is true. □

Now, let c_1 and c_2 to be the following constants

$$c_1 = \max_{j \in \mathcal{J}}(1/\mu_j) \quad \text{and} \quad c_2 = \left(\min_{j \in \mathcal{J}}(1/\mu_j)\right)^{-1}. \tag{84}$$

Thus, for any given $w \geqslant 0$ and all $i \in \mathscr{K}$, we have

$$w \leqslant c_1 \|q^*(w, \rho(i))\| \quad \text{and} \quad \|q^*(w, \rho(i))\| \leqslant c_2 w \tag{85}$$

since

$$\min_{j \in \mathcal{J}}(1/\mu_j) \sum_{j=1}^{J} q_j^*(w, \rho(i)) \leqslant w$$
$$= \sum_{j=1}^{J} \frac{q_j^*(w, \rho(i))}{\mu_j} \leqslant \max_{j \in \mathcal{J}}(1/\mu_j) \sum_{j=1}^{J} q_j^*(w, \rho(i)).$$

In addition, for any $\epsilon > 0$, define

$$T_1 = \max\{T_{(c_2+1)\epsilon, \epsilon}, T_{\max\{b_{1,1}, b_{1,2}\}, \epsilon/2}, T_{\max\{b_{1,1}, b_{1,2}\}, \sigma/2}\} \tag{86}$$

where σ is determined in Lemma 5 and

$$b_{1,1} = g((c_2+1)\epsilon) + \epsilon, \quad b_{2,1} = c_1 b_{1,1} + \epsilon, \tag{87}$$

$$b_{2,2} = \max\{b_{2,1}, \nu + \epsilon\} + C + \epsilon, \quad b_{1,2} = c_2 b_{2,2} + \epsilon \tag{88}$$

where $g(\cdot)$ is defined in (74). Then we have the following lemma.

Lemma 10. *Consider the time interval $[\tau, \tau+\delta]$ with $\tau \geqslant 0$ and $\delta > 0$ and suppose that there is some constant $\nu \geqslant 0$ such that*

$$\lim_{r \to \infty} \hat{W}^r(\tau) = \nu \quad \text{and} \quad \lim_{r \to \infty} \hat{Q}^r(\tau) = q^*(\nu, \rho(\alpha(\tau))). \tag{89}$$

Furthermore, let C be an arbitrarily chosen positive constant such that

$$\sup_{t_1, t_2 \in [\tau, \tau+\delta]} \|\hat{X}(t_1) - \hat{X}(t_2)\| \leqslant C \quad \text{with } \hat{X}(\cdot) \text{ given by (34).} \tag{90}$$

Then for any given sufficiently small number $\epsilon > 0$ and a given $T \geqslant T_1$, the following claims are true for all sufficiently large $r \in \{1, 2, \ldots\}$ and all $l \in \{0, 1, \ldots, \lceil r\delta/T \rceil - 1\}$:

$$\|\bar{Q}^{r,l}(u) - q^*(\bar{W}^{r,l}(u), \rho(\alpha(\eta^{r,l}(u))))\| \leqslant \epsilon \quad \text{for all } u \in [0,T], \tag{91}$$

$$\bar{W}^{r,l}(u) \leqslant \nu + C + O(\epsilon) \quad \text{for all } u \in [0,T], \tag{92}$$

$$\bar{Y}^{r,l}(u) - \bar{Y}^{r,l}(0) = 0 \quad \text{for all } u \in [0,T] \quad \text{if } \bar{W}^{r,l}(u) > \epsilon \text{ for all } u \in [0,T], \tag{93}$$

where $\lim_{\epsilon \to 0} O(\epsilon) = 0$.

The proof of Lemma 10 will be provided in the e-companion to the paper.

5.5. The Remainder of Proof of Theorem 1

This part of proof is provided in the e-companion to the paper.

Supplemental Material

Supplemental material to this paper is available at http://dx.doi.org/10.1287/opre.2013.1224.

Acknowledgments

The author thanks the editors and two anonymous referees for helpful comments and suggestions. The author also gratefully acknowledges financial support from the National Natural Science Foundation of China [Grants 10971249 and 11371010].

References

Acampora AS, Bhardwaj S, Tamari RM (2006) On best-case throughput of cellular data networks with cooperating base stations. *Proc. Allerton Conf. Comm., Control, Comput., Monticello, IL.*

Bhardwaj S, Williams RJ (2009) Diffusion approximation for a heavily loaded multi-user wireless communication system with coorperation. *Queuing Systems* 62(4):345–382.

Bhardwaj S, Williams RJ, Acampora AS (2007) On the performance of a two-user MIMO downlink system in heavy traffic. *IEEE Trans. Inform. Theory* 53(5):1851–1859.

Billingsley P (1999) *Convergence of Probability Measures*, 2nd ed. (John Wiley & Sons, New York).

Brémaud P (1981) *Point Processes and Queues: Martingale Dynamics* (Springer-Verlag, New York).

Cao J, Cleverland WS, Lin D, Sun DX (2001) Internet traffic tends to Poisson and independent as the load increases. Bell Labs technical report, Bell Labs, Murray Hill, NJ.

Choudhury GL, Mandelbaum A, Reiman MI, Whitt W (1997) Fluid and diffusion limits for queues in slowly changing environments. *Stochastic Models* 13(1):121–146.

Costa MHM (1983) Writing on dirty paper. *IEEE Trans. Inform. Theory* 29(3):439–441.

Dai JG, Dai W (1999) A heavy traffic limit theorem for a class of open queuing networks with finite buffers. *Queuing Systems* 32(1):5–40.

Dai W (1996) Brownian approximations for queuing networks with finite buffers: Modeling, heavy traffic analysis and numerical implementations. Unpublished doctoral dissertation, Georgia Institute of Technology, Atlanta.

Dai W (2011a) On the conflict of truncated random variable vs. heavy-tail and long-range dependence in computer and network simulation. *J. Comput. Inform. Systems* 7(5):1488–1499.

Dai W (2011b) Mean-variance portfolio selection based on a generalized BNS stochastic volatility model. *Internat. J. Comput. Math.* 88(16):3521–3534.

Dai W (2012) Heavy traffic limit theorems for a queue with Poisson ON/OFF long-range dependent sources and general service time distribution. *Acta Mathematicae Applicatae Sinica, English Ser.* 28(4):807–822.

Dai W, Wang S (2009) Optimal control based on a general exponential scheduling rule for a generalized switch. *Proc. Comm. Mobile Comput.*, Vol. 2 (IEEE Computer Society Press, Washington, DC), 332–336.

Ethier SN, Kurtz TG (1986) *Markov Processes: Characterization and Convergence* (John Wiley & Sons Inc., New York).

Gamal AEL, Cover TM (1980) Multiple user information theory. *Proc. IEEE* 68(12):1466–1483.

Goldsmith A, Jafar SA, Jindal N, Vishwanath N (2003) Capacity limits of MIMO channels. *IEEE J. Selected Areas Comm.* 21(5):684–702.

Harrison JM, Reiman MI (1981) Reflected Brownian motion on an orthant. *Ann. Probab.* 9(2):302–308.

Iglehart DL, Whitt W (1971) The equivalence of functional central limit theorems for counting processes and associated partial sums. *Ann. Math. Statist.* 42(4):1372–1378.

Jindal N, Vishwanath S, Goldsmith A (2004) On the duality of Gaussian multiple-access and broadcast channels. *IEEE Trans. Inform. Theory* 50(5):768–783.

Kumaran K, Viswanathan S (2005) Joint power and bandwidth allocation in downlink transmission. *IEEE Trans. Wireless Comm.* 4(3): 1008–1016.

Liu J, Hou YT (2008) Weighted proportional fairness capacity of Gaussian MIMO broadcast channels. *Proc. IEEE INFOCOM 2008* (IEEE Computer Society Press, Washington, DC), 1058–1066.

Luenberger DG (1984) *Linear and Nonlinear Programming*, 2nd ed. (Addison-Wesley Publishing Company, Reading, MA).

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

Prokhorov YV (1956) Convergence of random processes and limit theorems in probability theory. *Theory Probab. Appl.* 1(2):157–214.

Resnick SI (1992) *Adventures in Stochastic Processes* (Birkhäuser, Boston).

Rosen JR (1965) Existence and uniqueness of equilibrium points for concave N-person games. *Econometrics* 33(3):520–534.

Shakkotai S, Srikant R, Stolyar AL (2004) Pathwise optimality of the exponential scheduling rule for wireless channels. *Adv. Appl. Probab.* 36(4):1021–1045.

Stolyar AL (2004) MaxWeight scheduling in a generalized switch: State space collapse and workload minimization in heavy traffic. *Ann. Appl. Probab.* 14(1):1–53.

Vishwanath S, Jindal N, Goldsmith A (2003) Duality, achievable rates and sum-rate capacity of Gaussian MIMO broadcast channels. *IEEE Trans. Inform. Theory* 49(10):2658–2668.

Viswanathan H, Kumaran K (2005) Rate scheduling in multiple antenna downlink wireless systems. *IEEE Trans. Comm.* 53(4):645–655.

Wang HS, Moayeri N (1995) Finite-state Markov channel—A useful model for radio communication channels. *IEEE Trans. Vehicular Tech.* 44(1):163–171.

Weingarten H, Steinberg Y, Shamai(Shitz) S (2006) The capacity region of the Gaussian MIMO broadcast channel. *IEEE Trans. Inform. Theory* (IEEE, Piscataway, NJ), 3936–3964.

Ye H, Yao DD (2008) Heavy traffic optimality of a stochastic network under utility-maximizing resource control. *Oper. Res.* 56(2):453–470.

Ye HQ, Ou J, Yuan XM (2005) Stability of data networks: Stationary and bursty models. *Oper. Res.* 53(1):107–125.

Yu W, Rhee W, Boyd S, Cioffi J (2004) Iterative water-filling for vector multiple access channels. *IEEE Trans. Inform. Theory* 50(1): 145–151.

Zhou XY, Yin G (2003) Markowitz's mean-variance portfolio selection with regime switching: A continuous-time model. *SIAM J. Control Optim.* 42(4):1466–1482.

Wanyang Dai is a professor of mathematics at Nanjing University of China. His research interests include wireless engineering and technology, optimization, stochastic processes and stochastic (ordinary/partial) differential equations, information engineering, and applied mathematics and physics.

OPERATIONS RESEARCH
Vol. e-companion, No. ec, September 10, 2013, pp. 1–26
ISSN 0030-364X | EISSN 1526-5463 | 13 | e-companion | 1-26

INFORMS
DOI : e-companion

Optimal Rate Scheduling via Utility-Maximization for J-User MIMO Markov Fading Wireless Channels with Cooperation

Wanyang Dai
Department of Mathematics and State Key Laboratory of Novel Software Technology
Nanjing University, Nanjing 210093, China
nan5lu8@netra.nju.edu.cn
http://maths.nju.edu.cn:8001/portals/blog/wydai/WanyangDai.pdf

In this e-companion, we first present the outline of proof of Theorem 1. Then, we provide the proofs of some lemmas and theorem appeared in the main body of the paper, which include Lemmas 1- 2, Lemmas 4-8, Lemma 10, and the Remainder of Proof of Theorem 1. Equation and lemma numbers, such as (1), (2), ..., Lemma 6, etc. refer to the main body of the paper.

1. Outline of Proof of Theorem 1

For the reader's convenience, we first outline the proof of Theorem 1, which consists of the following five parts.

First, in Subsection 5.1, we justify a dual relationship between the utility-maximization problem in (6) and the cost-minimization problem in (14), which is summarized in Lemma 4. Then, we prove a claim in Lemma 5, which states that when the system state is close to the unique optimal solution to the cost minimization problem (called *a fixed point*), the capacity of the system will be fully utilized. The claims stated in Lemmas 4-5 look similar to their counterparts in Ye and Yao (2008); nevertheless, their concrete proofs are different owing to the different problem formulations and the difference in the capacity constraints of the two studies. Moreover, our claims are channel-state dependent.

Second, in Subsection 5.2, we justify a functional central limit theorem (Lemma 6) for a DSRP whose arrival rate process is driven by the FS-CTMC. The main idea used in proving Lemma 6 stems from the related discussion in Dai (1996) and Dai and Dai (1999), and the concrete proof techniques include the conventional functional central limit theorem (see, e.g., Iglehart and Whitt (1971) and Prokhorov (1956)), random change of time lemma (see, e.g., Billingsley (1999)), establishment of oscillation inequality (see, e.g., Dai (1996), and Dai and Dai (1999)), equivalent conditions of relative compactness, and the Skorohod representation theorem (see, e.g., Ethier and Kurtz (1986)).

Third, in Subsection 5.3, we derive the fluid limit processes for the physical processes under fluid scaling in Lemma 7 and study the asymptotic behavior for the fluid limit processes as time evolves in Lemma 8. Fluid limits are widely used as an intermediate step in justifying diffusion approximations (see, e.g., Bramson and Dai (2001), Stolyar (2004), Ye and Yao (2008), Dai (2007), Bhardwaj et al. (2007), Bhardwaj and Williams (2009), and the references therein). Nevertheless, our fluid limit is a random process driven by the FS-CTMC rather than a deterministic function of time, as obtained in existing studies. This new feature increases the complexity of proving Lemma 7 and Lemma 8, e.g., as compared with the study of Ye and Yao (2008), it requires more technical treatment in handling the FS-CTMC based jumps for the constructed Lyapunov function.

Therefore, by noticing this new feature and the difference between our optimal scheduling policy and that of Ye and Yao (2008), we develop a theory by combining and generalizing the discussions in Ye and Yao (2008), Dai (1995), Bhardwaj et al. (2007), and Bhardwaj and Williams (2009) to justify Lemma 7 and Lemma 8.

Next, in Subsection 5.4, we study the convergence of the workload and queue length processes on a finer time-scale, which is an important step in justifying the main result of the paper. This method has appeared in several queueing studies (see, e.g., Ye and Yao (2008), Stolyar (2004), Mandelbaum and Stolyar (2004), Shakkotai et al. (2004), etc.) The main difference between our study and existing studies is as follows: all the processes concerned in our study involve jumps introduced by the random environment, whereas processes in existing studies do not involve such jumps. Therefore, we develop a scheme and incorporate it into the framework as used by Ye and Yao (2008) to complete the proof of the convergence properties for the processes on a finer time-scale.

Finally, in Subsection 5.5, we combine the results obtained in the previous subsections with the uniqueness of the solution to an associated Skorohod problem and the minimality of the Skorohod problem to provide a proof for Theorem 1. Such techniques have been used in studies on network scheduling (see, e.g., Ye and Yao (2008), Stolyar (2004), Mandelbaum and Stolyar (2004), Shakkotai et al. (2004), etc.) Nevertheless, our justification logic and technical treatment are somewhat different. More precisely, the method for justifying the u.o.c convergence of the diffusively scaled unused capacity processes is different and a Helly's theorem-based approach is involved in both parts of the proof. Moreover, the use of the key lemma on finer time-scaling is different owing to the random environment.

2. Proof of Lemma 1

Note that for a fixed priority vector ν, the optimization characterization described in the outline is equivalent to finding the point on the capacity boundary that is tangent to a line whose slope is defined by the priority vector. Owing to the structure of the capacity region, we can see that all boundary points of the region are corner points of polyhedrons corresponding to different sets of covariance matrices. In addition, the corner point should correspond to successive decoding in order of increasing priority, i.e., the user with the highest priority should be decoded last, and therefore, sees no interference. Hence, by Goldsmith et al. (2003) and Yu et al. (2004), the problem of finding the boundary point on the capacity region associated with a descending ordered priority vector ν can be written as

$$\max_{\{\Gamma_j(i)\geq 0,\mathrm{Tr}(\Gamma_j(i))\leq P_j, j\in\mathcal{J}\}} f(\Gamma_1(i),...,\Gamma_J(i),\nu), \tag{EC-1}$$

where

$$\begin{aligned}&f(\Gamma_1(i),...,\Gamma_J(i),\nu) &\text{(EC-2)}\\ =&\nu_J\log\left|I+\sum_{j=1}^{J}H_j^{\dagger}(i)\Gamma_j(i)H_j(i)\right|+\sum_{j=1}^{J-1}\left((\nu_j-\nu_{j+1})\log\left|I+\sum_{l=1}^{j}H_l^{\dagger}(i)\Gamma_l(i)H_l(i)\right|\right),\end{aligned}$$

which is concave in the covariance matrices.

Now, let $\tilde{\nu}_j=\nu_j-\nu_{j+1}$ for $j\in\{1,...,J-1\}$ and $\tilde{\nu}_J=\nu_J$. Then, for any integer $m\in\{1,...,J-1\}$, let $S(k_1,...,k_m)$ denote the following set corresponding to exactly m indices $k_1,...,k_m\in\{1,...,J-1\}$ such that $\tilde{\nu}_{k_1}=...=\tilde{\nu}_{k_m}=0$, i.e.,

$$\begin{aligned}S(k_1,...,k_m)\equiv\{&f(\Gamma_1(i),...,\Gamma_J(i),\nu):\Gamma_j(i)\geq 0,\tilde{\nu}_j\geq 0 \text{ for } j\in\mathcal{J},\tilde{\nu}_{k_1}=...=\tilde{\nu}_{k_m}=0, &\text{(EC-3)}\\ &k_j\neq k_l \text{ for } j\neq l \text{ and } j,l\in\{1,...,m\},\ \nu_J>0,\ \sum_{j=1}^{J}\nu_j=1\}.\end{aligned}$$

Moreover, if $m=0$, we use $S(k_0)$ to denote the set corresponding to $\tilde{\nu}_j>0$ for all $j\in\mathcal{J}$, i.e.,

$$S(k_0)\equiv\left\{f(\Gamma_1(i),...,\Gamma_J(i),\nu):\Gamma_j(i)\geq 0,\tilde{\nu}_j>0\text{ for }j\in\mathcal{J},\nu_J>0,\sum_{j=1}^{J}\nu_j=1\right\}. \tag{EC-4}$$

In addition, if $m=J$, $\nu_J=0$, we use $S(k_J)$ to denote the following set corresponding to $\nu_J=0$:

$$S(k_J)\equiv\left\{f(\Gamma_1(i),...,\Gamma_J(i),\nu):\Gamma_j(i)\geq 0\text{ for }j\in\mathcal{J},\ \nu_J=0,\ \sum_{j=1}^{J}\nu_j=1\right\}. \tag{EC-5}$$

Eventually, we can define

$$S(k_0,k_1,...,k_m)=\begin{cases}S(k_1,...,k_m) & \text{if } m\in\mathcal{J},\\ S(k_0) & \text{if } m=0,\\ S(k_J) & \text{if } m=J.\end{cases} \tag{EC-6}$$

Thus, we have

$$\begin{aligned}&\left\{f(\Gamma_1(i),...,\Gamma_J(i),\nu):\Gamma_j(i)\geq 0,\ \nu_j\geq 0,\ \tilde{\nu}_j\geq 0\text{ for }j\in\mathcal{J},\sum_{j=0}^{J}\nu_j=1\right\} \qquad \text{(EC-7)}\\ &=\bigcup_{m=1}^{J}\ \bigcup_{k_1,...,k_m\in\mathcal{J}}S(k_0,k_1,...,k_m).\end{aligned}$$

Note that the J users can be arbitrarily ordered, and hence, we have $J!$ such priority orders, e.g., $\nu_{j_1}\geq\nu_{j_2}\geq...\geq\nu_{j_J}$, where $(j_1,...,j_J)$ is a permutation of $(1,...,J)$. Thus, we can see that our capacity region is bounded by L boundary pieces with L given by (41). In fact, the first term $J!$ on the right-hand side of the first equality in (41) is the number of boundary pieces corresponding to all $\nu_{j_1}>\nu_{j_2}>...>\nu_{j_J}$, $C_J^j(J-j+1)!$ ($j\in\{2,...,J\}$) is the number of boundary pieces corresponding to all $\nu_{k_1}=...=\nu_{k_j}$ with $k_1,...,k_j\in\{1,...,J\}$, and $k_l\neq k_h$ for $l\neq h$ when $\nu_{j_J}>0$, and the last term J on the right-hand side of (41) is the number of boundary pieces corresponding to $\nu_{j_J}=0$. Here, we remark that the number of boundary pieces obtained through the above method is consistent with the one derived in Liu and Hou (2008).

Next, we show the smoothness of these boundary pieces. Without loss of generality, our discussion will focus on a specific set $S(k_0)$ in a particular user priority order since the discussions for all other cases are similar. Therefore, we have $\nu_1>\nu_2>...>\nu_J>0$. Now, let $(y_1,...,y_{2NNJ})'$ denote the $(2NNJ)$-dimensional vector formed by the real part and the imaginary part of entries of $\Gamma_1(i)$,...,$\Gamma_J(i)$ in a suitable order. Since $\Gamma_j(i)$ for each $i\in\mathcal{K}$ and $j\in\mathcal{J}$ is a Hermitian matrix, y is actually determined by its M components with $M=N(N+1)J$. Hence, without causing confusion, we use $y=(y_1,...,y_M)'$ to denote such an M-dimensional vector in a suitable order. Thus, we know that $f(\Gamma_1(i),...,\Gamma_J(i),\nu)=f(y,\nu)$ is concave in y for each given $\nu\geq 0$. Moreover, define $\mathcal{J}^+\equiv\{J+1,...,L\}$ with $L\equiv 2^NJ$ and let $-f_j(y)$ for each $j\in\mathcal{J}^+$ denote a principal minor obtained from one of $\Gamma_1(i)$,...,$\Gamma_J(i)$. Then, the optimization problem in (EC-1) can be restated as

$$\max_{y\in R^M}f(y,\nu) \tag{EC-8}$$

subject to

$$f_j(y)\equiv\text{Tr}(\Gamma_j(i))-P_j\leq 0\quad\text{for all }\ j\in\mathcal{J}, \tag{EC-9}$$

$$f_j(y)\leq 0\qquad\text{for all }\ j\in\mathcal{J}^+. \tag{EC-10}$$

Therefore, it follows from the KKT optimality conditions (see, e.g., Luenberger (1984)) that the solution to the optimization problem in (EC-8)-(EC-9) for a function $f(y,\nu) \in S(k_0)$ with the associated $\nu \geq 0$ can be obtained through the equations

$$y_l \left(\frac{\partial f(y,\nu)}{\partial y_l} + \sum_{j=1}^{\mathbf{L}} \eta_j \frac{\partial f_j(y)}{\partial y_l} \right) = 0 \text{ for each } l \in \{1,2,...,M\}, \tag{EC-11}$$

$$\eta_j f_j(y) = 0 \text{ for each } j \in \mathcal{J} \cup \mathcal{J}^+, \tag{EC-12}$$

where $\eta_j \geq 0$ for $j \in \mathcal{J} \cup \mathcal{J}^+$ are the Lagrangian multipliers. Then, our remaining discussion can be divided into the following two steps.

Step One: Define

$$\mathcal{N} \equiv \left\{ \nu = (\nu_1,...,\nu_J)' : \nu_1 > ... > \nu_J > 0, \sum_{j=1}^{J} \nu_j = 1 \right\}. \tag{EC-13}$$

If there exists some $\nu \in \mathcal{N}$ such that the problem in (EC-8)-(EC-9) for the function $f(y,\nu) \in S(k_0)$ has at least one optimal solution located in the interior of the associated feasible region, we have the following discussions.

First, we suppose that the optimal solution is unique, given by $y^* = (y_1^*,...,y_M^*)'$. Then, we know that $f(y,\nu)$ is strictly concave in y since it is sufficiently smooth in y for the given ν from the definition of f. Hence, it follows from (EC-11)-(EC-12) that

$$F_l(y^*,\nu) \equiv \frac{\partial f(y^*,\nu)}{\partial y_l} = 0 \text{ for all } l \in \{1,...,M\}. \tag{EC-14}$$

Moreover, it follows from Theorem 4.3.1 on page 115 of Hiriart-Urruty and Lemaréchal (2001) that the Hessian matrix

$$\bigtriangledown^2 f(y,\nu) \equiv \left(\frac{\partial^2 f(y,\nu)}{\partial y_l \partial y_k} \right)_{M\times M} \text{ for all } l,k \in \{1,...,M\} \tag{EC-15}$$

is positive definite at all y within the M-dimensional feasible region. Now, define

$$F(y,\nu) \equiv \{F_l(y,\nu),\, l \in \{1,...,M\}\}.$$

Thus, we know that $F(y^*,\nu) = 0$ and the Jacobian determinant of $F(y,\nu)$ with respect to y at (y^*,ν) is nonzero owing to (EC-15), i.e.,

$$\frac{D(F_1,...,F_M)}{D(y_1,...,y_M)} \neq 0. \tag{EC-16}$$

Therefore, $F(y,\nu)$ satisfies all the conditions stated in the implicit function theorem. Hence, $F(y,\nu) = 0$ uniquely determines an M-dimensional function $y^*(\nu)$ that is continuous and differentiable with respect to ν in a neighborhood $O(\nu,\epsilon)$ of ν. Moreover, (EC-16) and (EC-14) hold in $O(\nu,\epsilon)$, which implies that $y^*(\nu)$ is an optimal solution to the problem in (EC-8)-(EC-9) for each $\nu \in O(\nu,\epsilon)$.

Next, we suppose that the problem in (EC-8)-(EC-9) for the function $f(y,\nu) \in S(k_0)$ has multiple optimal solutions located in the interior of the associated feasible region. Without loss of generality, we suppose that these optimal points are all in a m-dimensional hyperplane that is parallel to each coordinate-axis corresponding to those y with part of its components, $y_{s_l} \in \mathcal{Y}$, where

$$\mathcal{Y} \equiv \{y_{s_l} \in R, l \in \{1,...,m\}, s_l \in \{1,...,M\}\} \text{ for some } m \in \{1,...,M\}.$$

(Here we remark that if this is not the case, we can employ the method of rotation transformation to make this case true.) Therefore, from (EC-2) and the concavity of $f(y,\nu)$ in y, we know that $f(y,\nu)$ is independent of $y_{s_l}\in\mathcal{Y}$. Thus, there exists a $(M-m)$-dimensional set P_ν corresponding to each ν such that $f(y,\nu)$ only depends on $y_{s_l}\in\mathcal{Y}^c$ (the complementary set of $\mathcal{Y}$) and is strictly concave in those y_{s_l}. Therefore, for any optimal point $y^*(\nu)$ in the set P_ν and by considering the similar $(M-m)$-dimensional problem as in (EC-14)-(EC-15), we can conclude that $y^*(\nu)$ is continuous and differentiable in a neighborhood $O(\nu,\epsilon)$ of ν.

Hence, if the optimal points of $f(y,\nu)$ are all strictly located within the feasible region when ν moves in $\mathcal{N}$, it follows from the above discussion that $f(y,\nu)$ keeps either strictly concave or flat with respect to $y_{s_l}\in\mathcal{Y}^c$ or $y_{s_l}\in\mathcal{Y}$ for all $\nu\in\mathcal{N}$. Thus, we can conclude that all the optimal paths $y^*(\nu)$ are continuous and differentiable with respect to $\nu\in\mathcal{N}$. In other words, any set $\{\Gamma_1^*(\nu,i),...,\Gamma_J^*(\nu,i)\}$ of the optimal covariance matrices is continuous and differentiable with respect to $\nu\in\mathcal{N}$. Hence, it follows from (40) that the corner points of the capacity region, which are determined by the following equations, form a smooth curved facet $f(\Gamma_1^*(\nu,i),...,\Gamma_J^*(\nu,i))$ when ν moves in the region $\mathcal{N}$. Moreover, the facet does not depend on the choice of the set $\{\Gamma_1^*(\nu,i),...,\Gamma_J^*(\nu,i)\}$ along $\nu\in\mathcal{N}$. In addition, from (40), for all $j\in\mathcal{J}$, we have

$$c_j(\nu)=\log\left|I+\sum_{l=1}^{j}H_l^\dagger(i)\Gamma_l^*(\nu,i)H_l(i)\right|-\log\left|I+\sum_{l=1}^{j-1}H_l^\dagger(i)\Gamma_l^*(\nu,i)H_l(i)\right|. \qquad \text{(EC-17)}$$

However, if some optimal point of $f(y,\nu)$ reaches one of the boundaries of the feasible region when ν moves in $\mathcal{N}$, then the associated justification for this case is part of the proof in the following step.

Step Two: Without loss of generality, we suppose that $f(y,\nu)$ is strictly concave in y for all $\nu\in\mathcal{N}$; otherwise we can employ the similar argument as above. Therefore, assume that $y^*=(y_1^*,...,y_M^*)'$ is the solution to the optimization problem in (EC-8)-(EC-10) and it is located on one of the boundary pieces, say $f_j(y)=0$ for some $j\in\mathcal{J}\cup\mathcal{J}^+$. Note that $f_j(y)$ depends only on part of components of y. Hence, we can use $\mathcal{Y}^*\equiv\{y_{s_l}^*,l\in\{1,...,m\},s_l\in\{1,...,M\}\}$ for some $m\in\{1,...,M\}$ to denote the set of those components of y^*, which determine the surface $f_j(y)=0$ for the given $j\in\mathcal{J}\cup\mathcal{J}^+$. Then, the remaining components of y^* are located in the interior of the corresponding $(M-m)$-dimensional feasible region. Now, let $\mathcal{L}_1\equiv(\mathcal{J}\cup\mathcal{J}^+)\cap\{j:f_j(y^*)=0\}$ and $\mathcal{L}\equiv\{l:y_{s_l}^*\in y^*\setminus\mathcal{Y}^*\}$. Hence, we have that

$$F_l(y^*,\eta^*,\nu)\equiv\frac{\partial f(y^*,\nu)}{\partial y_l}+\sum_{j\in\mathcal{L}_1}\eta_j^*\frac{\partial f_j(y^*)}{\partial y_l}=0 \text{ for each } l\in\mathcal{L}, \qquad \text{(EC-18)}$$

where η_j^* for all $j\in\mathcal{L}_1$ are the Lagrangian multipliers corresponding to y^*.

Next, let $y\in R_+^M$ denote the vector whose components y_{s_l} for all $l\in\{1,...,m\}$ are confined in $\mathcal{Y}^*$. Note that $y_{s_l}^*$ with $l\in\mathcal{L}$ is in the interior of the corresponding $(M-m)$-dimensional feasible region. Then, we know that $f(y,\nu)$ is strictly concave in the components of y, except those $y_{s_l}\in\mathcal{Y}^*$ since it is sufficiently smooth in y. Thus, it follows from Theorem 4.3.1 in page 115 of Hiriart-Urruty and Lemaréchal (2001) that the Hessian matrix

$$\bigtriangledown^2 f(y,\nu)\equiv\left(\frac{\partial^2 f(y,\nu)}{\partial y_l\partial y_k}\right)_{(M-m)\times(M-m)} \text{ for all } l,k\in\mathcal{L}.$$

is positive definite at all y whose components y_{s_l} for all $l\in\{1,...,m\}$ are confined in $\mathcal{Y}^*$. Moreover, if we define

$$F(y,\eta,\nu)\equiv\begin{cases}F_l(y,\eta,\nu) & \text{if } l\in\mathcal{L},\\ F_l(y)=f_l(y) & \text{if } l\in\mathcal{L}_1,\end{cases} \qquad \text{(EC-19)}$$

we can conclude that the Jacobian determinant of $F(y,\eta,\nu)$ with respect to y_l $(l\in\mathcal{L})$ and η_l $(l\in\mathcal{L}_1)$ is nonzero at y^*. Moreover, from the definition of $f(y,\nu)$ and $f_j(y)$ for $j\in\mathcal{J}$, we know that $F(y,\eta,\nu)$ satisfies all the conditions as stated in the implicit function theorem. Hence, $F(y,\eta,\nu)=0$ uniquely determines an $(M+\bar{J})$-dimensional function $(y^*(\nu),\eta^*(\nu))$ that is continuous and differentiable in $\nu\in\mathcal{N}$ (where $\bar{J}$ is the number of f_l $(l\in\mathcal{L}_1)$ such that $f_l(y^*)=0$). In addition, all the components $y^*_{s_l}(\nu)$ with $l\in\{1,...,m\}$ are confined in $\mathcal{Y}^*$ when $\nu\in\mathcal{N}$ moves. Therefore, the remaining proof of this boundary situation can be divided into the following three cases.

Case One: When $u\in\mathcal{N}$ continuously moves to a vector $\nu\in\mathcal{N}$, the optimal point $y^*(u)$ moves from the interior of the feasible region to the optimal point $y^*(\nu)(=y^*)$ on the boundary of the feasible region. Then, we need to prove that $y^*(u)$ and its associated derivatives converge to $y^*(\nu)$ and its corresponding derivatives as u converges to ν continuously within a neighborhood of $\nu\in\mathcal{N}$ in the whole M-dimensional feasible region, which implies that the components $y^*_{s_l}(u)$ for all $l\in\{1,...,m\}$ are not necessarily confined in $\mathcal{Y}^*$ when $u\in\mathcal{N}$ moves.

In fact, define the following constraints of parallel surfaces

$$\tilde{f}_j(y,b)\equiv f_j(y)-b_j=0 \text{ for each } j\in\mathcal{L}_1, \tag{EC-20}$$

where b is an arbitrary constant vector whose components are given by b_j $(j\in\mathcal{L}_1)$. Therefore, by applying the KKT optimality conditions, the optimal solution $\tilde{y}^*$ to the the problem (EC-8) with the constraints in (EC-20) should be given by the following equations,

$$F_l(\tilde{y}^*,\tilde{\eta}^*,\nu,b)\equiv\frac{\partial f(\tilde{y}^*,\nu)}{\partial y_l}+\sum_{j\in\mathcal{L}_1}\tilde{\eta}^*_j\frac{\partial\tilde{f}_j(\tilde{y}^*,b)}{\partial y_l}=0 \tag{EC-21}$$

for each $l\in\mathcal{L}\equiv\{l:y^*_{s_l}\in y^*\setminus\mathcal{Y}^*\}$, where $\tilde{\eta}^*_j$ $(j\in\mathcal{L}_1)$ are the related Lagrangian multipliers corresponding to $\tilde{y}^*$. Now, for each $\tilde{y}\in R^M$, define

$$F(\tilde{y},\tilde{\eta},\nu,b)\equiv\begin{cases}F_l(\tilde{y},\tilde{\eta},\nu,b) & \text{if } l\in\mathcal{L},\\ F_l(\tilde{y},b)=\tilde{f}_l(\tilde{y},b) & \text{if } l\in\mathcal{L}_1.\end{cases} \tag{EC-22}$$

Then, by the similar argument as used for (EC-19), we know that there is a unique $(M+\bar{J})$-dimensional optimal path $(\tilde{y}^*(u,b),\tilde{\eta}^*(u,b))$ which is continuous and differentiable with respect to $(u,b)\in\mathcal{N}\times R^{\bar{J}}$ (where $\bar{J}$ is the dimension of b). Moreover, all the components of $\tilde{y}^*(u,b)$ corresponding to $y^*_{s_l}\in\mathcal{Y}^*$ satisfy the constraints in (EC-20). Thus, we know that $\tilde{y}^*(u,b)$ and its associated derivatives converge to $y^*(\nu)$ and its corresponding derivatives as (u,b) converges to $(\nu,0)$ continuously. Moreover, note that $y^*(u)=\tilde{y}^*(u,b)$ when $b_l<0$ $(l\in\mathcal{L}_1)$ are all close to zero, which implies that all the components of $\tilde{y}^*(u,b)$ corresponding to $y^*_{s_l}\in\mathcal{Y}^*$ are also continuous and differentiable with respect to $u\in\mathcal{N}$ when $b_l<0$ $(l\in\mathcal{L}_1)$ are all close to zero. Hence, we can conclude that $y^*(u)$ and its associated derivatives converge to $y^*(\nu)$ and its associated derivatives as $u\to\nu$, which implies that $y^*(\nu)$ is continuous and differentiable at a neighborhood of ν in the whole M-dimensional feasible region.

Case Two: When $u\in\mathcal{N}$ moves to a vector $\nu\in\mathcal{N}$, the optimal point $y^*(u)$ moves to the optimal point $y^*(\nu)$ $(=y^*)$ from a boundary piece of the feasible region next to the boundary piece on which $y^*(\nu)$ is located. The proof for this case is similar to the one as used in *Case One*. Hence, we omit it.

Case Three: When $u\in\mathcal{N}$ moves to a vector $\nu\in\mathcal{N}$, the optimal point $y^*(u)$ moves to the optimal point $y^*(\nu)$ $(=y^*)$ from a boundary piece of the feasible region that is not next to the boundary piece on which $y^*(\nu)$ is located. Owing to the concavity of $f(y,u)$, the optimal point $y^*(u)$ must go first into the interior of the feasible region and then to the other boundary piece. Therefore, the proof for this case is the same as the one as used in *Case One*.

Finally, we note that the boundary piece corresponding to $S(k_1,...,k_J)$ is a J-dimensional linear facet, which is determined by the sum-rate capacity bound (see, e.g., Yu et al. (2004) for more details). $\square$

3. Proof of Lemma 2

It follows from Goldsmith et al. (2003) that the capacity region for the J-user MIMO BC with $N=1$ and each $i\in\mathcal{K}$ is given by

$$\begin{aligned}
&\mathcal{R}(i)=\mathcal{C}_{BC}(P,H(i)) \qquad \text{(EC-23)}\\
=&\bigcup_{\{(P_1,\dots,P_J):\sum_{j=1}^{J}P_j=P\}}\mathcal{C}_{MAC}(P_1,\dots,P_J,H^{\dagger}(i))\\
=&\bigcup_{\{(P_1,\dots,P_J):\sum_{j=1}^{J}P_j=P\}}\left\{c\in R_+^J:\sum_{j\in S}c_j\leq\frac{1}{2}\log\left|I+\sum_{j\in S}H_j^{\dagger}(i)P_jH_j(i)\right|,\forall\ S\subset\mathcal{J}\right\}.
\end{aligned}$$

Hence, owing to the similarity of structures between $\mathcal{M}(i)$ in (40) and $\mathcal{R}(i)$ in (EC-23), we can apply the similar discussion as for the MIMO MAC and the discussion in Vishwanath et al. (2003) to conclude that the claims in the lemma are true. □

4. Proof of Lemma 4

First, without loss of generality, we suppose that $q>0$. Then, it follows from the KKT optimality conditions (see, e.g., Luenberger (1984)) that the solution to the utility maximization problem in (6) can be obtained through the equations

$$c_j\left(\frac{\partial U_j(q_j,c_j)}{\partial c_j}+\sum_{k=1}^{B}\eta_k\frac{\partial h_k(c,i)}{\partial c_j}\right)=0 \text{ for } j\in\mathcal{J}, \qquad \text{(EC-24)}$$

$$\eta_k h_k(c,i)=0 \text{ for each } k\in\mathcal{U}, \qquad \text{(EC-25)}$$

where B and $\mathcal{U}$ are defined in (2), $\eta_k\geq 0$ for all $k\in\mathcal{U}$ are the Lagrangian multipliers, and $h_k(c,i)$ for each $k\in\mathcal{U}$ and $i\in\mathcal{K}$ is defined in (2). Similarly, the solution to the cost minimization problem (14) can be obtained through the equations

$$q_j\left(\frac{\partial C_j(q_j,c_j)}{\partial q_j}+\frac{\theta}{\mu_j}\right)=0 \text{ for each } j\in\mathcal{J}, \qquad \text{(EC-26)}$$

$$\theta\left(w-\sum_{j=1}^{J}\frac{q_j}{\mu_j}\right)=0, \qquad \text{(EC-27)}$$

where $\theta\geq 0$ is the Lagrangian multiplier. Moreover, it follows from (16) that

$$\frac{\partial C_j(q_j,c_j)}{\partial q_j}=\frac{1}{\mu_j}\frac{\partial U_j(q_j,c_j)}{\partial c_j}. \qquad \text{(EC-28)}$$

Thus, based on the above facts, the claim in the first part of the lemma can be proved as follows. By condition (8), we know that $\sum_{j=1}^{J}U_j(q_j,c_j)$ is strictly concave in c for each $q>0$. Therefore, $c^*=\rho(i)$ is the unique optimal solution to the utility maximization problem in (6) for the given $q>0$ in the utility function, which satisfies (EC-24)-(EC-25). Thus, if we take

$$\theta=-\sum_{k=1}^{B}\eta_k\frac{\partial h_k(\rho(i),i)}{\partial c_j},$$

it follows from (EC-24) and (EC-28) that (EC-26) holds. From the condition (11), we know that $V(q,c)$ is strictly convex in q for each $c>0$. Hence, the cost minimization problem in (14) has a unique optimal solution $q^* = q$ when $c = c^* = \rho(i)$ is in the cost function and $w = \sum_{j=1}^J q_j^*/\mu_j$ is in the constraints.

Conversely, the claim in the second part of the lemma can be proved as follows. From the conditions (10)-(11) and the relationship (16), we know that $V(q,\rho(i))$ is strictly convex in q. Therefore, q^* is the unique optimal solution to the cost minimization problem (14) with $\Lambda(q^*,i) = \rho(i)$. Thus, we can prove $q^* > 0$ by showing a contradiction.

In fact, without loss of generality, we suppose that there is some $m \in \mathcal{J}$ with $m < J$ such that $q^* \in \mathcal{Q}(k_1,...,k_m)$ with $k_1 \neq 1$ and $k_m = J$, where $\mathcal{Q}(k_1,...,k_m)$ is defined in (12). Then, we can construct a 2-dimensional line for some constant $\epsilon \geq w$,

$$P_1: \ \frac{q_1}{\mu_1} + \frac{q_J}{\mu_J} + \sum_{j\neq 1,J,j\in\mathcal{J}} \frac{q_j^*}{\mu_j} = \epsilon \geq w \tag{EC-29}$$

such that it passes through the point q^*. Now it follows from (16) that the function $f(q_1,\rho(i))$ ($= V(q,\rho(i))$) with the constraint P_1 for all $q = (q_1, q_2^*, ..., q_{J-1}^*, q_J)' \in \mathcal{R}_+^J$) is of the following derivative function in $q_1 \in \mathcal{R}_1^+$:

$$\frac{\partial f(q_1,\rho(i))}{\partial q_1} = \frac{1}{\mu_1}\frac{\partial U_1(q_1,\rho_1(i))}{\partial c_1} - \frac{1}{\mu_1}\frac{\partial U_J((\epsilon - \frac{q_1}{\mu_1} - \sum_{j\neq 1,J,j\in\mathcal{J}}\frac{q_j^*}{\mu_j})\mu_J, \rho_J(i))}{\partial c_J}, \tag{EC-30}$$

which is strictly increasing in $q_1 \in \mathcal{R}_1^+$ from (10). Moreover, it follows from (EC-30) and (11) that

$$\frac{\partial f(0,\rho(i))}{\partial q_1} = -\frac{1}{\mu_1}\frac{\partial U_J((\epsilon - \sum_{j\neq 1,J,j\in\mathcal{J}}\frac{q_j^*}{\mu_j})\mu_J,\rho_J(i))}{\partial c_J} < 0, \tag{EC-31}$$

$$\frac{\partial f(q_1^*,\rho(i))}{\partial q_1} = \frac{1}{\mu_1}\frac{\partial U_1(q_1^*,\rho_1(i))}{\partial c_1} > 0. \tag{EC-32}$$

Then, by (EC-31) and (EC-32), we know that there is a $\tilde{q}_1 \in (0, q_1^*)$ such that

$$\frac{\partial f(\tilde{q}_1,\rho(i))}{\partial q_1} = 0, \tag{EC-33}$$

which implies that on the curve $f(q,\rho(i))$ with $q = (q_1, q_2^*, ..., q_{J-1}^*, q_J)' \in R_+^J$, there exists a minimal point $\tilde{q} \in \mathcal{R}_+^J$ with $\tilde{q} = (\tilde{q}_1, q_2^*, ..., q_{J-1}^*, \tilde{q}_J)'$ such that $V(\tilde{q},\rho(i)) < V(q^*,\rho(i))$

$$\tilde{q}_J = \left(\epsilon - \frac{\tilde{q}_1}{\mu_1} - \sum_{j\neq 1,J,j\in\mathcal{J}}\frac{q_j^*}{\mu_j}\right)\mu_J.$$

This contradicts the assumption that q^* is the optimal solution to the cost minimization problem in (14). Hence, we can conclude that $q^* > 0$.

Finally, if q^* is the optimal solution to (14) with $c = \Lambda(q^*,i) = \rho(i)$ in the cost function, we see that (EC-26)-(EC-27) hold with $q = q^*$ and $c = \Lambda(q^*,i) = \rho(i)$. Therefore, we can take $\eta_{k_U} = \theta$ and $\eta_k = 0$ when $k \neq k_U$ in (EC-24)-(EC-25) since $q^* > 0$ and ρ is on the curve $h_{k_U}(c,i) = 0$. Hence, $\Lambda^*(q^*,i) = \Lambda(q^*,i) = \rho(i)$ for each $i \in \mathcal{K}$ is an optimal solution to (6) with $q = q^*$ in the utility function. □

5. Proof of Lemma 5.

For the first part in the lemma, we have the following observations. From the condition (11), we know that $V(q,c)$ is strictly convex in q for each $c>0$. Hence, the cost minimization problem in (14) has a unique optimal solution $q^*=q$ when $c=\rho(i)$ is in the cost function. Moreover, the continuity of $q^*(w,\rho(i))$ in terms of w for each $i\in\mathcal{K}$ can be proved similarly as in Ye and Yao (2008).

For the second part of the lemma, it can be proved by showing a contradiction. In fact, if the claim is not true for some $i\in\mathcal{K}$ and some $\epsilon>0$, then for a sequence of $\sigma^l\downarrow 0$ along $l\in\mathcal{R}$, there is a sequence of states $\tilde{q}^l\in\mathcal{V}(\epsilon,\sigma^l,i)$ with $l\in\mathcal{R}$ satisfying

$$\left\|\tilde{q}^l-q^*(\tilde{w}^l,\rho(i))\right\|\to 0 \text{ as } l\to\infty \text{ along } l\in\mathcal{R}, \tag{EC-34}$$

$$\tilde{w}^l=\sum_{j=1}^{J}\frac{1}{\mu_j}\tilde{q}^l_j\geq\epsilon \text{ for all } l\in\mathcal{R} \tag{EC-35}$$

such that

$$\sum_{j=1}^{J}\Lambda_j(\tilde{q}^l,i)<\sum_{j=1}^{J}\rho_j(i) \text{ for all } l\in\mathcal{R}. \tag{EC-36}$$

Otherwise, if there is some $l_0\in\mathcal{R}$ such that $\mathcal{V}(\epsilon,\sigma^l,i)$ are empty for all $l\geq l_0$, then (52) is automatically true for the given $i\in\mathcal{K}$ and $\epsilon>0$, which is a contradiction. Now, let

$$q^l=(\epsilon/\tilde{w}^l)\tilde{q}^l \text{ so that } w^l=\sum_{j=1}^{J}\frac{q^l_j}{\mu_j}=\epsilon \text{ for all } l\in\mathcal{R}; \tag{EC-37}$$

then, it follows from (EC-34)-(EC-35), (EC-37), (13) and Lemma 4 that as $l\to\infty$ along $l\in\mathcal{R}$,

$$q^l=\frac{\epsilon q^*(\tilde{w}^l,\rho(i))}{\tilde{w}^l}+\frac{\epsilon}{\tilde{w}^l}\left(\tilde{q}^l-q^*(\tilde{w}^l,\rho(i))\right)\to\hat{q}=q^*(\epsilon,\rho(i))>0, \tag{EC-38}$$

which implies that $q^l>0$ for all sufficiently large $l\in\mathcal{R}$. Moreover, we have

$$\frac{\epsilon q^*(\tilde{w}^l,\rho(i))}{\tilde{w}^l}\to\hat{q} \text{ as } l\to\infty \text{ along } l\in\mathcal{R}. \tag{EC-39}$$

Hence, by (EC-38), (8), (12), and the similar proof as used for the second part of Lemma 4, we have for all sufficiently large $l\in\mathcal{R}$,

$$\Lambda(\hat{q},i)>0 \text{ and } \Lambda(q^l,i)>0. \tag{EC-40}$$

Furthermore, by (13) and (EC-36), we have that, for each $l\in\mathcal{R}$,

$$\sum_{j=1}^{J}\Lambda_j(q^l,i)<\sum_{j=1}^{J}\rho_j(i). \tag{EC-41}$$

Thus, it follows from (EC-41) and Lemma 3 that

$$\sum_{j=1}^{J}\Lambda_j(\hat{q},i)\leq\sum_{j=1}^{J}\rho_j(i). \tag{EC-42}$$

Note that the condition in (8) and the fact in (EC-38) imply that $\Lambda(\hat{q},i)$ and $\Lambda(q^l,i)$ for sufficiently large l can only locate on the capacity surface of $\mathcal{R}(i)$ (that is defined in (2)). Then, by combining this fact with (EC-41)-(EC-42) and Lemma 3, we can see that $\Lambda(\hat{q},i)$ cannot be in the interior of the facet corresponding to $C_U(i)$. Hence, we can conclude that there is some $j\in\mathcal{J}$, e.g., without loss of generality, take $j=1$ such that

$$\Lambda_1(\hat{q},i)<\rho_1(i). \tag{EC-43}$$

On the one hand, it follows from (EC-40) and (EC-24) that there exists a set of Lagrange multipliers $\{\eta_{jk}\geq 0, k\in\{1,...,B+J\}, j\in\mathcal{J}\}$ such that

$$\begin{aligned}\sum_{j=1}^{J}\sum_{k=1}^{B}\eta_k\frac{\partial h_k(\Lambda(\hat{q},i),i)}{\partial c_j} &= -\sum_{j=1}^{J}\frac{\partial U_j(\hat{q}_j,\Lambda_j(\hat{q},i))}{\partial c_j} \\ &= -\lim_{l\to\infty}\sum_{j=1}^{J}\frac{\partial U_j(\hat{q}_j,\Lambda_j(q^*(\tilde{w}^l,\rho(i)),i))}{\partial c_j} \\ &= -\sum_{j=1}^{J}\frac{\partial U_j(\hat{q}_j,\rho_j(i))}{\partial c_j}.\end{aligned} \tag{EC-44}$$

where B is defined in (2), the first equality of (EC-44) follows from (EC-24)-(EC-25) and (EC-40), the second equality follows from (EC-39), (13) and (50), and the third equality follows from Lemma 4.

On the other hand, owing to the strict concavity of $U_j(q_j,c_j)$ in c_j for each $j\in\mathcal{J}$ as stated in (8), it follows from (EC-40)-(EC-43) that

$$\begin{aligned}&\sum_{j=1}^{J}\sum_{k=1}^{B}\eta_k\frac{\partial h_k(\Lambda(\hat{q},i),i)}{\partial c_j} \\ &< -\frac{\partial U_1(\hat{q}_1,\rho_1(i))}{\partial c_1}-\lim_{l\to\infty}\sum_{j\neq 1,j\in\mathcal{J}}\frac{\partial U_j(\hat{q}_j,\Lambda_j(q^*(\tilde{w}^l,\rho(i)),i))}{\partial c_j} \\ &= -\sum_{j=1}^{J}\frac{\partial U_j(\hat{q}_j,\rho_j(i))}{\partial c_j}.\end{aligned} \tag{EC-45}$$

Obviously, there is a contradiction between (EC-44) and (EC-45). Thus, the assumption stated in (EC-34)-(EC-36) is not true, which implies that the second claim in the lemma holds for the third case. □

6. Proof of Lemma 6

It follows from the heavy traffic condition (21), the functional central limit theorem (see, e.g., Iglehart and Whitt (1971) and Prokhorov (1956)), and the random change of time lemma (see, e.g., page 151 of Billingsley (1999)), Lemma 8.4 in Dai and Dai (1999) that, for each $n\in\{0,1,...\}$,

$$\begin{aligned}&\left(\hat{E}^r(\tau_n+t)-\hat{E}^r(\tau_n)\right)I_{\{0\leq t<\sigma_n\}} \\ =\ &\frac{1}{r}\left(A^r(r^2(\tau_n+t))-A^r(r^2\tau_n)\right)I_{\{0\leq t<\sigma_n\}}-r\left(\bar{\lambda}^r(\tau_n+t)-\bar{\lambda}^r(\tau_n)\right)I_{\{0\leq t<\sigma_n\}} \\ =\ &\frac{1}{r}\tilde{A}^r(r^2(t-\phi_n/r^2)I_{\{0\leq t<\sigma_n\}})+\frac{1}{r}e_n-r\left(\bar{\lambda}^r(\tau_n+t)-\bar{\lambda}^r(\tau_n)\right)I_{\{0\leq t<\sigma_n\}} \\ \Rightarrow\ &\left(\Gamma^E(\alpha(\tau_n))\right)^{\frac{1}{2}}I_{\{0\leq t<\sigma_n\}}B^E(t) \qquad \text{as } r\to\infty \\ =^d\ &\left(H^E(\tau_n+t)-H^E(\tau_n)\right)I_{\{0\leq t<\sigma_n\}},\end{aligned} \tag{EC-46}$$

where $\sigma_n = \tau_{n+1} - \tau_n$ is an exponentially distributed random variable independent of all other random events concerned since $\alpha(\cdot)$ is a FS-CTMC, $e_n = (e_n^1, ..., e_n^J)'$ with $e_n^j = 0$ if $\hat{E}_j^r(\cdot)$ has a jump at τ_n and 1 otherwise for each $j \in \{1, ..., J\}$, $\tilde{A}^r(\cdot)$ is a renewal process with rate vector $\lambda^r(\alpha(\tau_n)) = (\lambda_1^r(\alpha(\tau_n)), ..., \lambda_J^r(\alpha(\tau_n)))'$ and

$$\tilde{A}^r(r^2(\cdot - \phi_n/r^2)) = (\tilde{A}_1^r(r^2(\cdot - \phi_n^1/r^2)), ..., \tilde{A}_J^r(r^2(\cdot - \phi_n^J/r^2)))'$$

with $\phi_n = (\phi_n^1, ..., \phi_n^J)'$ being a J-dimensional random vector whose jth component ϕ_n^j for each $j \in \{1, ..., J\}$ denotes the remaining arrival time beginning at τ_n for a packet to the jth queue with rate $\lambda_j(\alpha(\tau_n))$ switched from $\lambda_j(\alpha(\tau_{n-1}))$ at τ_n for each $n \in \{1, 2, ...\}$. Moreover, for later convenience, we rewrite (EC-46) as follows, over each $[\tau_n, \tau_{n+1})$ and as $r \to \infty$:

$$\begin{aligned} \tilde{E}^{r,n}(\cdot) &\equiv \hat{E}^r(\tau_n + \cdot) - \hat{E}^r(\tau_n) \\ &\Rightarrow H^E(\tau_n + \cdot) - H^E(\tau_n) \\ &\equiv \tilde{H}^{E,n}(\cdot). \end{aligned} \tag{EC-47}$$

Then, by following (EC-47) and by generalizing the discussion in the proof for Theorem 3.2 in Dai (1996) or Lemma 8.2 in Dai and Dai (1999), we can prove the claim in (62).

To do so, we first establish the relative compactness for $\hat{E}^r(\cdot)$ with $r \in \mathcal{R}$. In fact, define the modulus of continuity in terms of a function $x(\cdot) : [0, \infty) \to R^d$ with some integer $d > 0$ for each given $T > 0$ and $\delta > 0$ as follows:

$$w(x, \delta, T) \equiv \inf_{t_l} \max_{l} \operatorname{Osc}(x, [t_{l-1}, t_l)), \tag{EC-48}$$

where the infimum takes over the finite sets $\{t_l\}$ of points satisfying $0 = t_0 < t_1 < ... < t_m = T$ and $t_l - t_{l-1} > \delta$ for $l = 1, ..., m$, and

$$\operatorname{Osc}(x, [t_{l-1}, t_l]) = \sup_{t_1 \le s \le t \le t_2} \|x(t) - x(s)\|_2 \tag{EC-49}$$

with $\|\cdot\|_2$ denoting the Euclidean norm in R^d. Then it follows from Corollary 7.4 in page 129 of Ethier and Kurtz (1986) that the justification of the relative compactness is equivalent to proving the following two conditions:

(a) For each $\eta > 0$ and rational $t \ge 0$, there exists a constant $c(\eta, t)$ such that

$$\liminf_{r \to \infty} P\left\{\left\|\hat{E}^r(t)\right\|_2 \le c(\eta, t)\right\} \ge 1 - \eta.$$

(b) For each $\eta > 0$ and $T > 0$, there exists a $\delta > 0$ such that

$$\limsup_{r \to \infty} P\left\{w(\hat{E}^r, \delta, T) \ge \eta\right\} \le \eta.$$

To show (a), we first define $N(t) \equiv \max\{n, \tau_n \le t\}$ for each $t \in (0, \infty)$. Then, for each rational $t > 0$, take a $T > 0$ such that $t \in (0, T]$ and define a sequence of events: $\mathcal{S}_l \equiv \{\omega : N(T, \omega) \le l\}$ for each $l \in \{1, 2, ..., \}$. Since $\alpha(\cdot)$ has at most finitely many jumps a.s. over $[0, T]$, we know that the sequence of probabilities $P\{S_l\}$ increases monotonously to the unity as $l \to \infty$. Thus, for the given $\eta > 0$, there is some large enough $L > 0$ such that

$$P\{\mathcal{S}_L\} \ge 1 - \frac{\eta}{2}. \tag{EC-50}$$

Moreover, it follows from (EC-47) and Remark 7.3 in page 129 of Ethier and Kurtz (1986) that $\tilde{E}^r(\cdot)$ satisfies the following compact containment condition, i.e., for each $\eta > 0$ and $T > 0$, there is a constant $K_n > 0$ for each $n \in \{0, 1, ..., \}$ such that

$$\inf_r P\{\mathcal{T}^{r,n}\} \geq 1 - \frac{\eta}{2L} \text{ with } \mathcal{T}^{r,n} \equiv \{\omega : \|\tilde{E}^r(t)\|_2 \leq K_n, t \in [0,T] \cap [0, \sigma_n)\}. \qquad \text{(EC-51)}$$

In addition, for each $n \in \{1, 2, ..., \}$, let $\Delta_n = (\delta_n^1, ..., \delta_n^J)'$ with $\delta_n^j = 1$ if $\hat{E}_j^r(\cdot)$ has a jump at τ_n and zero otherwise for each $j \in \{1, ..., J\}$. Then, for each $t \in [\tau_{N(t)}, \tau_{N(t)+1})$, we have

$$\hat{E}^r(t) = \hat{E}^r(\tau_{N(t)}) + \tilde{E}^{r,N(t)}(t - \tau_{N(t)}), \qquad \text{(EC-52)}$$

$$\hat{E}^r(\tau_n) - \hat{E}^r(\tau_n^-) = \frac{1}{r}\Delta_n. \qquad \text{(EC-53)}$$

Therefore, it follows from (EC-52)-(EC-53) that along each sample path and for any $t_1, t_2 \in [0, T]$,

$$\text{Osc}\left(\hat{E}^r, [t_1, t_2]\right) \leq \sum_{n=0}^{N(t_2)} \text{Osc}\left(\tilde{E}^{r,n}, [t_1 - \tau_n, t_2 - \tau_n] \cap [0, \sigma_n)\right) + \frac{1}{r}(N(t_2) - N(t_1)). \text{ (EC-54)}$$

Thus, it follows from (EC-54) that along each sample path in $\mathcal{S}_L \cap \mathcal{T}^{r,n}$ with $r, n \in \{1, 2, ...\}$,

$$\begin{aligned} \left\|\hat{E}^r(t)\right\|_2 &\leq \left\|\hat{E}^r(0)\right\|_2 + \text{Osc}\left(\hat{E}^r, [0, t]\right) \qquad \text{(EC-55)} \\ &\leq 2\sum_{n=0}^{L} \sup_{t \in [0,T] \cap [0, \sigma_n} \left\|\tilde{E}^{r,n}(t)\right\|_2 + \frac{L}{r}. \end{aligned}$$

Hence, for the above arbitrarily given $\eta > 0$, each rational $t \in [0, T]$, and sufficiently large $r \in \mathcal{R}$, we know that

$$\begin{aligned} & P\left\{\left\|\hat{E}^r(t)\right\|_2 \leq 2^{L+1}\sum_{n=0}^{L} K_n\right\} \qquad \text{(EC-56)} \\ \geq\ & P\left\{\left\{\left\|\hat{E}^r(t)\right\|_2 \leq 2^{L+1}\sum_{n=0}^{L} K_n\right\} \bigcap \mathcal{S}_L\right\} \\ \geq\ & P\{\mathcal{S}_L\} - \sum_{n=0}^{L} P\left\{\left\{\left\|\tilde{E}^{r,n}(t)\right\|_2 > \left(2K_n - \frac{1}{r2^L}\right)\right\} \bigcap \mathcal{S}_L \text{ for some } t \in [0,T] \cap [0, \sigma_n)\right\} \\ \geq\ & P\{\mathcal{S}_L\} - \sum_{n=0}^{L} P\left\{\left\{\left\|\tilde{E}^{r,n}(t)\right\|_2 > K_n\right\} \bigcap \mathcal{S}_L \text{ for some } t \in [0,T] \cap [0, \sigma_n)\right\} \\ >\ & 1 - \eta, \end{aligned}$$

where the second inequality follows from (EC-55) and the fact that

$$P\{\|aX + bY\|_2 \geq K_1 + K_2\} \leq P\left\{\|X\|_2 \geq \frac{K_1}{2|a|}\right\} + P\left\{\|Y\|_2 \geq \frac{K_2}{2|b|}\right\}$$

for any real number a, b and random vectors X, Y. Moreover, the last inequality in (EC-56) follows from (EC-50) and (EC-51). Thus, condition (a) holds.

Next, we prove the condition (b) to be true. From (EC-47), we know that for each $\eta > 0$ and $T > 0$, there exists a $\delta_n > 0$ for each $n \in \{0, 1, ..., \}$ such that

$$\limsup_{r \to \infty} P\left\{w(\tilde{E}^{r,n}, \delta_n, [0,T] \cap [0, \sigma_n)) \geq \frac{\eta}{2^L L}\right\} \leq \frac{\eta}{2^L L}. \qquad \text{(EC-57)}$$

Now, take $\delta = \min\{\delta_0, ..., \delta_L\} > 0$; then, for each $r \in \{1, 2, ..., \}$ and each sample path in $\mathcal{S}_L$,

$$\begin{aligned} w(\hat{E}^r, \delta, T) &\le \sum_{n=0}^{J} w(\tilde{E}^{r,n}, \delta, [0,T] \cap [0, \sigma_n)) + \frac{L}{r} \\ &\le \sum_{n=0}^{J} w(\tilde{E}^{r,n}, \delta_n, [0,T] \cap [0, \sigma_n)) + \frac{L}{r}, \end{aligned} \tag{EC-58}$$

where the first inequality follows from (EC-48) and (EC-54), and the second inequality follows from 1.9 in page 326 of Jacod and Shiryaev (2003). Therefore, for each sufficiently large $r \in \mathcal{R}$, it follows from (EC-57)-(EC-58) that

$$\begin{aligned} & P\left\{ w(\hat{E}^r, \delta, T) \ge \eta \right\} \\ < \ & \frac{\eta}{2} + P\left\{ \left\{ w(\hat{E}^r, \delta, T) \ge \eta \right\} \bigcap \mathcal{S}_L \right\} \\ < \ & \frac{\eta}{2} + \sum_{n=0}^{L} P\left\{ \left\{ w(\tilde{E}^{r,n}, \delta_n, [0,T] \cap [0, \sigma_n)) \ge \frac{1}{2^{L-1}} \left(\frac{\eta}{L} - \frac{1}{r} \right) \right\} \bigcap \mathcal{S}_L \right\} \\ < \ & \frac{\eta}{2} + \sum_{n=0}^{L} P\left\{ \left\{ w(\tilde{E}^{r,n}, \delta_n, [0,T] \cap [0, \sigma_n)) \ge \frac{\eta}{2^L L} \right\} \bigcap \mathcal{S}_L \right\} \\ \le \ & \eta. \end{aligned}$$

Hence, the condition (b) is true and hence we know that $\hat{E}^r(\cdot)$ is relatively compact for $r \in \mathcal{R}$.

Finally, consider any subsequence $\mathcal{R}_1 \subseteq \mathcal{R}$ such that, along $r \in \mathcal{R}_1$, we have

$$\hat{E}^r(\cdot) \Rightarrow \hat{E}(\cdot) \text{ (a process to be identified).} \tag{EC-59}$$

Then, it follows from the Skorohod representation theorem (see, e.g., Theorem 3.1.8 in page 102 of Ethier and Kurtz (1986)) and the random change of time lemma (see, e.g., page 151 of Billingsley (1999)) that for each $n \in \{0, 1, ..., \}$ and along $r \in \mathcal{R}_1$,

$$\left(\hat{E}^r(\cdot) I_{\{\cdot \le \tau_{n+1}\}}, \hat{E}^r(\cdot) I_{\{\cdot \le \tau_n\}} \right) \Rightarrow \left(\hat{E}(\cdot) I_{\{\cdot \le \tau_{n+1}\}}, \hat{E}(\cdot) I_{\{\cdot \le \tau_n\}} \right).$$

Then, by the method of induction in terms of $n \in \{0, 1, ..., \}$, (EC-47), and the continuous-mapping theorem (see, e.g., Theorem 3.4.1 in page 85 of Whitt (2002)), we can conclude that along $r \in \mathcal{R}_1$, the limit in (EC-59) is $H^E(\cdot)$. Moreover, since $\mathcal{R}_1$ is arbitrarily chosen, we know that $\hat{E}^r \Rightarrow H^E(\cdot)$ along $r \in \mathcal{R}$. Furthermore, by the independence assumptions and the functional central limit theorem, we know that the claim in Lemma 6 is true. □

7. Proof of Lemma 7

It follows from (55) and (18) that $\bar{T}^r(\cdot)$ is a.s. uniformly Lipschitz continuous with Lipschitz constant $\max_{i \in \mathcal{K}}(\sum_{j=1}^{J} \rho_j(i))$ for each $r > 0$, which implies that it is absolutely continuous and differentiable at almost every $t \in (0, \infty)$ (in other words, almost every $t \in (0, \infty)$ is a *regular point* of $\bar{T}^r(\cdot)$). Thus, the sequence of stochastic processes $\{\bar{T}^r(\cdot), r \in \mathcal{R}\}$ is C-tight, that is, it is tight and each weak limit point is in $C[0, \infty)^J$ a.s., where $C[0, \infty)^J$ is the space of all J-dimensional continuous functions over $[0, \infty)$ and is endowed with the Skorohod J_1-topology (see, e.g, Page 116 of Ethier and Kurtz (1986)). Moreover, it follows from Lemma 6 that $\left(\bar{E}^r(\cdot), \bar{S}^r(\cdot)\right)$ is also C-tight. In addition, by (57)-(55) and (63), we know that

$$\bar{Q}_j^r(t) = \bar{E}_j^r(t) - \bar{S}_j^r(\bar{T}_j^r(t)). \tag{EC-60}$$

Hence, it follows from (EC-60), (19) and the random time change lemma in page 151 of Billingsley (1999) that the following sequence is C-tight as well,

$$\left(\bar{E}^r(\cdot), \bar{S}^r(\cdot), \bar{T}^r(\cdot), \bar{Q}^r(\cdot), \bar{W}^r(\cdot), \bar{Y}^r(\cdot)\right), \tag{EC-61}$$

where we have used the independent assumption related to $\{\alpha^r(\cdot), r \in \mathcal{R}\}$, the second condition in (21) and the fact that

$$\frac{1}{r^2}\int_0^{r^2\cdot} \rho_j(\alpha^r(s))ds = \int_0^{\cdot} \rho_j(\alpha(s))ds. \tag{EC-62}$$

Therefore, any subsequence of the processes in (EC-61) has a further subsequence convergent in distribution. Now, suppose $(\bar{E}(\cdot), \bar{S}(\cdot), \bar{T}(\cdot), \bar{Q}(\cdot), \bar{W}(\cdot), \bar{Y}(\cdot))$ is a weak limit point corresponding to the further subsequence indexed by $\{r_l, l = 1, 2, ...\}$. Then, by the Skorohod representation theorem (see, e.g., Theorem 3.1.8 in page 102 of Ethier and Kurtz (1986)), there is a common supporting probability space such that

$$\left(\bar{E}^{r_l}(\cdot), \bar{S}^{r_l}(\cdot), \bar{T}^{r_l}(\cdot), \bar{Q}^{r_l}(\cdot), \bar{W}^{r_l}(\cdot), \bar{Y}^{r_l}(\cdot)\right) \to (\bar{\lambda}(\cdot), \mu(\cdot), \bar{T}(\cdot), \bar{Q}(\cdot), \bar{W}(\cdot), \bar{Y}(\cdot)) \tag{EC-63}$$

u.o.c. a.s. as $l \to \infty$ and the limiting processes in (EC-63) satisfy (64)-(69). Here we only need to justify (69) to be true and other equations hold obviously.

In fact, from (EC-63), we know that the limit processes in (EC-63) are uniformly Lipschitz continuous a.s. Hence, our discussion will be based on a fix sample path and each regular point $t > 0$ over an interval (τ_{n-1}, τ_n) with $n \in \{1, 2, ...\}$ for $\bar{T}_j$ with $j \in \mathcal{J}$. It follows from (64) that $\bar{Q}$ is differential at t and satisfies

$$\frac{d\bar{Q}_j(t)}{dt} = \lambda_j(\alpha(t)) - \mu_j \frac{d\bar{T}_j(t)}{dt} \tag{EC-64}$$

for each $j \in \mathcal{J}$. If $\bar{Q}_j(t) = 0$ for some $j \in \mathcal{J}$, then it follows from $\bar{Q}_j(\cdot) \geq 0$ that

$$\frac{d\bar{Q}_j(t)}{dt} = 0 \text{ which implies that } \frac{d\bar{T}_j(t)}{dt} = \frac{\lambda_j(\alpha(t))}{\mu_j} = \rho_j(\alpha(t)). \tag{EC-65}$$

If $\bar{Q}_j(t) > 0$ for the $j \in \mathcal{J}$, then there exists a finite interval $(a,b) \in [0, \infty)$ containing t in it such that $\bar{Q}_j(s) > 0$ for all $s \in (a,b)$ and hence we can take small enough $\delta > 0$ such that $\bar{Q}_j(t+s) > 0$ with $s \in (0, \delta)$. Thus, it follows from (55) and (50) that

$$\begin{aligned}
&\left|\frac{1}{\delta}\left(\bar{T}_j^{r_l}(t+\delta) - \bar{T}_j^{r_l}(t)\right) - \Lambda_j(\bar{Q}(t), \alpha(t))\right| \qquad \text{(EC-66)}\\
\leq\ & \frac{1}{\delta}\int_0^{\delta} \left|\Lambda_j(\bar{Q}^{r_l}(t+s), \alpha(t+s)) - \Lambda_j(\bar{Q}(t+s), \alpha(t+s))\right| ds \\
&+ \frac{1}{\delta}\int_0^{\delta} \left|\Lambda_j(\bar{Q}(t+s), \alpha(t+s)) - \Lambda_j(\bar{Q}(t), \alpha(t))\right| ds \\
\to\ & \frac{1}{\delta}\int_0^{\delta} \left|\Lambda_j(\bar{Q}(t+s), \alpha(t+s)) - \Lambda_j(\bar{Q}(t), \alpha(t))\right| ds \quad \text{as } l \to \infty
\end{aligned}$$

where we have used the Lebesgue dominated convergence theorem for the last claim in (EC-66). From the right-continuity of $\alpha(\cdot)$, the Lipschitz continuity of $\bar{Q}(\cdot)$ and (50), the last expression in (EC-66) tends to zero as $\delta \to 0^+$. Hence, we have

$$\frac{d\bar{T}_j(t)}{dt} = \frac{d\bar{T}_j(t^+)}{dt} = \bar{\Lambda}_j(\bar{Q}(t), \alpha(t)) \text{ for each } j \in \mathcal{J}, \tag{EC-67}$$

which implies that the claims in (68)-(69) are true.

Next, we introduce the following cost objective with $c(i)=\rho(i)$ in (16) for each $i\in\mathcal{K}$,

$$\psi(q,i)\equiv\sum_{j=1}^{J}C_j(q_j,\rho_j(i)). \tag{EC-68}$$

Then, for each regular time $t\geq 0$ of $\bar{Q}(t)$ over time interval (τ_{n-1},τ_n) with a given $n\in\{1,2,...\}$, we have

$$\begin{aligned}\frac{d\psi(\bar{Q}(t),\alpha(t))}{dt} &= \sum_{j=1}^{J}\left(\frac{d\bar{Q}_j(t)}{dt}\frac{\partial C_j(\bar{Q}_j(t),\rho_j(\alpha(t))}{\partial\bar{Q}_j(t)}+\frac{d\rho_j(\alpha(t))}{dt}\frac{\partial C_j(\bar{Q}_j(t),\rho_j(\alpha(t))}{\partial\rho_j(\alpha(t))}\right) \\ &= \sum_{j=1}^{J}\left(\rho_j(\alpha(t))-\Lambda_j(\bar{Q}(t),\alpha(t))\right)\frac{\partial U_j(\bar{Q}_j(t),\rho_j(\alpha(t)))}{\partial\rho_j(\alpha(t))}I_{\{\bar{Q}_j(t)>0\}} \\ &\leq 0\end{aligned} \tag{EC-69}$$

where we have used (EC-64), (EC-67), (16) and the fact that the sample paths of $\alpha(\cdot)$ are piecewise constants for the second equality of (EC-69), and we have used the concavity of the utility functions, the fact that $\Lambda_j(\bar{Q}(t),\alpha(t))$ is the optimal solution to (6), and similar arguments as used in Ye and Yao (2008) for the last inequality of (EC-69). Therefore, for any given $n\in\{0,1,2,...\}$ and each $t\in[\tau_n,\tau_{n+1})$, we have

$$\begin{aligned}0 &\leq \psi(\bar{Q}(t),\alpha(t)) \\ &\leq \psi(\bar{Q}(\tau_n),\alpha(\tau_n)) \\ &= \sum_{j=1}^{J}\frac{1}{\mu_j}\int_0^{\bar{Q}_j(\tau_n)}\frac{\partial U_j(u,\rho_j(\alpha(\tau_n)))}{\partial C_j}du \\ &= \left(\frac{d\Psi(\bar{\rho}_1(\alpha(\tau_n)))}{dc_1}\right)\left(\frac{d\Psi(\bar{\rho}_1(\alpha(\tau_{n-1})))}{dc_1}\right)^{-1}\psi(\bar{Q}(\tau_n),\alpha(\tau_{n-1})) \\ &\leq \left(\frac{d\Psi(\bar{\rho}_1(\alpha(\tau_n)))}{dc_1}\right)\left(\frac{d\Psi(\bar{\rho}_1(\alpha(\tau_{n-1})))}{dc_1}\right)^{-1}\psi(\bar{Q}(\tau_{n-1}),\alpha(\tau_{n-1})) \\ &\leq \left(\frac{d\Psi(\bar{\rho}_1(\alpha(\tau_n)))}{dc_1}\right)\left(\frac{d\Psi(\bar{\rho}_1(\alpha(\tau_{n-2})))}{dc_1}\right)^{-1}\psi(\bar{Q}(\tau_{n-2}),\alpha(\tau_{n-2})) \\ &\cdots \\ &\leq \left(\frac{d\Psi(\bar{\rho}_1(\alpha(\tau_n)))}{dc_1}\right)\left(\frac{d\Psi(\bar{\rho}_1(\alpha(\tau_0)))}{dc_1}\right)^{-1}\psi(\bar{Q}(0),\alpha(0)) \\ &\leq \kappa\psi(\bar{Q}(0),\alpha(0)),\end{aligned} \tag{EC-70}$$

where the second inequality in (EC-70) follows from (EC-69); the first equality in (EC-70) follows from (EC-68), (16), (8)-(9); the second equality in (EC-70) follows from (22)-(24) and the continuity of $\bar{Q}(t)$ at τ_n with $n\in\{1,2,...,\}$; the third inequality in (EC-70) follows from (EC-69). Moreover, the κ in the last inequality of (EC-70) is a positive constant given by

$$\kappa=\max_{i,j\in\mathcal{K}}\left(\frac{d\Psi(\bar{\rho}_1(i))}{dc_1}\right)\left(\frac{d\Psi(\bar{\rho}_1(j))}{dc_1}\right)^{-1}.$$

Finally, if $\bar{Q}(0)=0$, it follows from (EC-70) that $\bar{Q}(t)=0$ for all $t\geq 0$. Thus, by (65), we know that $\bar{W}(t)=\bar{Y}(t)=0$ for all $t\geq 0$. Moreover, it follows from(EC-65) that the third claim in (71) is true. Hence, under the assumption that $\bar{Q}(0)=0$, all the claims stated in the lemma are true. □

8. Proof of Lemma 8

If $\|\bar{Q}(0)\| \leq \chi$, it follows from (EC-70) that $\bar{Q}(t)$ is bounded for all $t \geq 0$ since $C_j(q_j, \rho_j(i))$ for each $j \in \mathcal{J}$ and $i \in \mathcal{K}$ is strictly increasing and unbounded function in q_j. Moreover, it follows from (74) that $g(\kappa)$ is increasing in κ with $g(0) = 0$, and hence, it follows from (EC-70) and (EC-68) that

$$C_j(\bar{Q}_j(t), \rho_j(\alpha(t))) \leq \tilde{g}(\chi) \text{ and } \bar{Q}_j(t) \leq C_j^{-1}(\tilde{g}(\chi), \rho_j(\alpha(t))) \tag{EC-71}$$

for each $t \geq 0$ and $j \in \mathcal{J}$, which implies that (75) is true and $\bar{W}(t)$ increases to some finite number as t increases due to (65)-(66) and (75), i.e.,

$$\bar{W}(t) \uparrow \bar{W}(\infty) < \infty \text{ as } t \to \infty. \tag{EC-72}$$

Thus, we can define the following Lyapunov function with at most countably many jumps:

$$L(\bar{Q}(t), \alpha(t)) \equiv \psi(\bar{Q}(t), \alpha(t)) - \psi(q^*(\bar{W}(t), \rho(\alpha(t))), \alpha(t)), \tag{EC-73}$$

which is nonnegative and bounded over $t \in [0, \infty)$ due to Lemma 5, (65), Lemma 7, (EC-72) and the fact that $\bar{Q}(t)$ and $\rho(\alpha(t))$ are bounded over $[0, \infty)$. Then, for any given regular time $t > 0$ over an interval (τ_{n-1}, τ_n) with $n \in \{1, 2, ...\}$ and for any $\delta > 0$, we can show that there exists a $\sigma > 0$ such that

$$\frac{dL(\bar{Q}(t), \alpha(t))}{dt} \leq -\sigma \text{ if } \left\|\bar{Q}(t) - q^*(\bar{W}(t), \rho(\alpha(t)))\right\| \geq \delta. \tag{EC-74}$$

In fact, it follows from (EC-69) and (EC-72) that $\psi(\bar{Q}(t), \alpha(t))$ is non-increasing and $\psi(q^*(\bar{W}(t), \rho(\alpha(t))), \alpha(t))$ is non-decreasing in $t \in (\tau_{n-1}, \tau_n)$ since $\alpha(t)$ keeps flat over the time interval (τ_{n-1}, τ_n). Hence, we only need to show (EC-74) to be true with respect to $\psi(\bar{Q}(t), \alpha(t))$. By (EC-69), we define

$$\begin{aligned} h(\bar{Q}(t), \alpha(t)) &\equiv \frac{d\psi(\bar{Q}(t), \alpha(t))}{dt} \\ &= \sum_{j=1}^{J} \left(\rho_j(\alpha(t)) - \Lambda_j(\bar{Q}(t), \alpha(t))\right) \frac{\partial U_j(\bar{Q}_j(t), \rho_j(\alpha(t)))}{\partial \rho_j(\alpha(t))} I_{\{\bar{Q}_j(t) > 0\}}, \end{aligned} \tag{EC-75}$$

which is continuous in terms of $\bar{Q}(t) = q \neq 0$ with $q \in R_+^J$ from (50), (11), and the second-order differentiability of $U_j(q_j, c_j)$. Next, let

$$\mathcal{C}(i) \equiv \{q \in R_+^J : \|q - q^*(w(q), \rho(i))\| \geq \delta\} \subset \{q \in R_+^J : q \neq 0\}, \tag{EC-76}$$

where the workload $w(q)$ corresponding to each $q \in R_+^J$ is defined as in (65) and the set $\mathcal{C}(i)$ is a closed subset of R_+^J from the first part of Lemma 5. Moreover, similar to (EC-69), we know that $h(q, i) \leq 0$ and the equality is true if and only if $q = q^*(w(q), \rho(i))$.

In fact, suppose that the *if part* is true with some $q \in \mathcal{Q}(k_1, ..., k_m)$ that is defined in (12). Then, it follows from (EC-75) and the last equality in (EC-69) that $\{\rho_l(i), l \neq k_1, ..., k_m\}$ is the solution to the corresponding optimization problem in (6) with $\{q_l, l \neq k_1, ..., k_m\}$ in the associated $(J-m)$-dimensional utility function. Thus, it follows from Lemma 4 that $\{q_l, l \neq k_1, ..., k_m\} = \{q_l^*(w(q), \rho(i)), l \neq k_1, ..., k_m\}$. Moreover, since $q^*(w(q), \rho(i)) > 0$ owing to $w(q) > 0$ and Lemma 4, we know that $\psi(\bar{Q}(t), \alpha(t)) - \psi(q^*(\bar{W}(t), \rho(\alpha(t))), \alpha(t)) < 0$ for $\bar{Q}(t) = q$, which contradicts the fact that $q^*(w(q), \rho(i))$ is the solution to the cost minimization problem in (14). Conversely, the *only if* part is the direct conclusion of the second part in Lemma 4. Therefore, $h(q, i) < 0$ over $\mathcal{C}(i)$. Since $h(q, i)$ is continuous in $q \neq 0$, we know that there exists a $\sigma > 0$ such that

$$h(q, i) \leq -\sigma \text{ in } \mathcal{C}(i) \tag{EC-77}$$

Moreover, since the state space of $\alpha(\cdot)$ is finite, we can consider σ as the common constant such that (EC-77) is true for all $i \in \mathcal{K}$. So the claim in (EC-74) is proved.

Next, we prove that there exists a time $T_{\chi,\epsilon} > 0$ for any given $\epsilon > 0$ such that (76) is true. To do so, we first show that

$$L(\bar{Q}(t), \alpha(t)) \to 0 \text{ as } t \to \infty. \tag{EC-78}$$

In fact, define

$$L_1(\bar{Q}(t), \alpha(t)) \equiv L(\bar{Q}(t), \alpha(t)) - e(t) \tag{EC-79}$$

where $e(t)$ is a step function given by

$$\begin{aligned} e(t) \equiv & \sum_{n:\tau_n \le t} \left(\psi(\bar{Q}(\tau_n), \alpha(\tau_n))) - \psi(\bar{Q}(\tau_n^-), \alpha(\tau_n^-)) \right) \\ & - \sum_{n:\tau_n \le t} \left(\psi(q^*(\bar{W}(\tau_n), \rho(\alpha(\tau_n)), \alpha(\tau_n)) - \psi(q^*(\bar{W}(\tau_n^-), \rho(\alpha(\tau_n^-)), \alpha(\tau_n^-)) \right). \end{aligned} \tag{EC-80}$$

Therefore, we can see that $L_1(\bar{Q}(t), \alpha(t))$ is continuous and bounded over $t \in [0, \infty)$ since $\bar{Q}(t)$ and $\rho(\alpha(t))$ are bounded. Thus we know that $e(t)$ is also bounded over $t \in [0, \infty)$ because $L(\bar{Q}(t), \alpha(t))$ is bounded. Moreover, since

$$\frac{dL_1(\bar{Q}(t), \alpha(t))}{dt} = \frac{dL(\bar{Q}(t), \alpha(t))}{dt} \le 0 \text{ for a.a. } t \in [0, \infty), \tag{EC-81}$$

we know that $L_1(\bar{Q}(t), \alpha(t))$ converges to some constant as $t \to \infty$.

Now, since $e(t)$ is a step function and is bounded, any convergent subsequence of $e(t)$ in terms of t corresponds to a sequence of holding time intervals as $t \to \infty$ such that the convergence of $e(t)$ is true for all t along the sequence of holding time intervals. Moreover, since the state space of $\alpha(\cdot)$ is finite, there exists at least one $i \in \mathcal{K}$ such that the holding time intervals corresponding to this particular state i appear infinitely many times. For convenience, we use $[\tau_{n_l}, \tau_{n_{l+1}})$ with $l \in \{1, 2, ..., \}$ to denote such a sequence of holding time intervals, where τ_{n_l} is the jump time of $\alpha(\cdot)$ corresponding to the index n_l. Note that $[\tau_{n_l}, \tau_{n_{l+1}})$ with $l \in \{1, 2, ..., \}$ are sampled from a sequence of *i.i.d* random variables (actually exponentially distributed). Therefore, due to the strong law of large numbers and without loss of generality, we can assume that

$$\sum_{l=1}^{\infty} \left(\tau_{n_{l+1}} - \tau_{n_l} \right) = \infty. \tag{EC-82}$$

Therefore, for an arbitrarily given convergent subsequence of $e(t)$, we can obtain a sequence of holding time intervals $[\tau_{n_l}, \tau_{n_{l+1}})$ ($l \in \{1, 2, ..., \}$) with the property (EC-82) associated with a particular state $i \in \mathcal{K}$. Then, it follows from the convergence of $L_1(\bar{Q}(t), \alpha(t))$ that

$$L(\bar{Q}(t), \alpha(t)) \to L_\infty \ge 0 \text{ as } t \to \infty \text{ over } t \in \cup_{l=1}^{\infty} [\tau_{n_l}, \tau_{n_{l+1}}). \tag{EC-83}$$

Furthermore, we can claim that $L_\infty = 0$ by showing a contradiction. In fact, if we assume that $L_\infty > 0$, then for any given constant ϵ satisfying $0 < \epsilon < L_\infty$, there exists some sufficiently large time $T_1 > 0$ such that

$$L(\bar{Q}(t), \alpha(t)) > L_\infty - \epsilon > 0 \text{ for all } t \in [T_1, \infty) \cap \left(\cup_{l=1}^{\infty} [\tau_{n_l}, \tau_{n_{l+1}}) \right). \tag{EC-84}$$

Since $\psi(q,i)$ is continuous and strictly increasing in $q \in R_+^J$ for each $i \in \mathcal{K}$, it follows from (EC-84) that there exist some $\delta > 0$ and $\sigma > 0$ such that (EC-74) is true for all $t \in [T_1, \infty) \cap \left(\cup_{l=1}^{\infty} [\tau_{n_l}, \tau_{n_{l+1}})\right)$. Thus, it follows from (EC-79) and (EC-81)-(EC-82) that

$$\begin{aligned} L(\bar{Q}(t), \alpha(t)) &= L(\bar{Q}(0), \alpha(0)) + e(t) + \int_0^t \frac{dL_1(\bar{Q}(t), \alpha(t))}{dt} dt \\ &\leq C - \sigma \sum_{l=1}^{N(t)-1} (\tau_{n_{l+1}} - \tau_{n_l}) \\ &< 0 \end{aligned} \tag{EC-85}$$

for all sufficiently large $t \in [T_1, \infty) \cap \left(\cup_{l=1}^{\infty} [\tau_{n_l}, \tau_{n_{l+1}})\right)$, where $N(t) = \max\{l : \tau_{n_l} \leq t\}$ and C is a positive constant since $e(t)$ is bounded. However, the derived result in (EC-85) contradicts the fact that $L(\bar{Q}(t), \alpha(t)) \geq 0$. Therefore, the assumption that $L_\infty > 0$ is not true, which implies that $L_\infty = 0$. Since the convergent subsequence of $\alpha(\cdot)$ is arbitrarily chosen, we know that the convergence in (EC-78) is true (readers are also referred to Dai (1995) for related discussion concerning a continuous Lyapunov function with no jumps.) Hence, it follows from (EC-78), the continuity, and strict monotonicity of $\psi(q,i)$ in $q \in R_+^J$ for each $i \in \mathcal{K}$ that there exists a time $T_{\chi,\epsilon} > 0$ for any given $\epsilon > 0$ such that (76) is true.

Finally, if $\bar{Q}(0) = q^*(\bar{W}(0), \rho(\alpha(0)))$, then it follows from (EC-73) that the claim that $\bar{Q}(t) = \bar{Q}(0)$ a.s. for all $t \in [\tau_0, \tau_1)$ is true. □

9. Proof of Lemma 10

For convenience, besides (91), we will prove the following stronger claims instead of showing (92) and (93) directly, that is, for sufficiently large $r \in \{1, 2, ...\}$ and all nonnegative integers $l \in \{0, 1, ..., \lceil r\delta/T \rceil - 1\}$,

$$\begin{aligned} &\text{if } \bar{W}^{r,l}(u) \leq \epsilon < C \text{ for some } u \in [0,T], && \text{(EC-86)} \\ &\text{then } \bar{W}^{r,l}(u) \leq b_{2,1},\ \left\|\bar{Q}^{r,l}(u)\right\| \leq b_{1,1} \text{ for all } u \in [0,T]; \\ &\text{if } \bar{W}^{r,l}(u) > \epsilon \text{ for all } u \in [0,T], && \text{(EC-87)} \\ &\text{then } \bar{W}^{r,l}(u) \leq b_{2,2},\ \bar{Q}^{r,l}(u) \leq b_{1,2},\ \bar{Y}^{r,l}(u) - \bar{Y}^{r,l}(0) = 0 \text{ for all } u \in [0,T]. \end{aligned}$$

Thus, the remaining proof of the lemma can be divided into the following two parts.

Part One: We justify the claims stated in the lemma to be true when $l = 0$. In fact, it follows from (81) and (89) that

$$(\bar{W}^{r,0}(0), \bar{Q}^{r,0}(0)) = (\hat{W}^r(\tau), \hat{Q}^r(\tau)) \to (\nu, q^*(\nu, \rho(\alpha(\tau))) \text{ as } r \to \infty. \tag{EC-88}$$

Moreover, from the definition of τ_n defined in (4), we know that $[\tau, \tau + T/r] \subset [\tau_{n-1}, \tau_n)$ with some $n \in \{1, 2, ...\}$ for all sufficiently large $r \in \mathcal{R}$. Thus, $\alpha(\eta^{r,0}(u))$ keeps some constant $\alpha(\tau)$ for all $u \in [0,T]$ when r is sufficiently large along the given sample path. Hence, it follows from Lemma 9 that

$$(\bar{W}^{r,0}(u), \bar{Q}^{r,0}(u)) \to (\bar{W}(u), \bar{Q}(u)) = (\nu, q^*(\nu, \rho(\alpha(\tau))) \text{ u.o.c. for all } u \in [0,T] \tag{EC-89}$$

as $r \to \infty$ from (81) and the uniqueness of the limit.

Therefore, it follows from the first part of Lemma 5, (EC-89), and the similar argument as used in Ye and Yao (2008) that for all sufficiently large $r \in \{1, 2, ...\}$ and for all $u \in [0,T]$,

$$\left\|\bar{Q}^{r,0}(u) - q^*(\bar{W}^{r,0}(u), \rho(\alpha(\eta^{r,0}(u))))\right\| \leq \epsilon. \tag{EC-90}$$

Thus, (91) presented in the lemma holds when $l=0$. Moreover, it follows from (EC-89) and (90) that the bound estimations in (EC-86) and (EC-87) are true for all $u\in[0,T]$ and all sufficiently large r when $l=0$. In addition, the complementarity in (EC-87) can be shown as follows. For the given $\epsilon>0$ in the current lemma, it follows from the first part of Lemma 5 and (EC-89) that a $\sigma>0$ can be chosen such that for sufficiently large $r\in\mathcal{R}$ and all $u\in[0,T]$,

$$\left\|\bar{Q}^{r,0}(u)-q^*(\bar{W}^{r,0}(u),\rho(\alpha(\eta^{r,0}(u)))\right\|\leq\sigma \tag{EC-91}$$

since $\alpha(\eta^{r,0}(u))=\alpha(\tau)$ for all $u\in[0,T]$ when r is sufficiently large. Thus, if $\bar{W}^{r,0}(u)>\epsilon$ for all $u\in[0,T]$, then

$$\begin{aligned}\bar{Y}^{r,0}(u)-\bar{Y}^{r,0}(0)&=\sum_{j=1}^{J}\int_0^u\left(\rho(\alpha(\eta^{r,0}(s)))-\Lambda_j(\bar{Q}^r(\eta^{r,0}(s)),\alpha(\eta^{r,0}(s)))\right)ds \\ &=\sum_{j=1}^{J}\int_0^u\left(\rho(\alpha(\eta^{r,0}(s)))-\Lambda_j(\bar{Q}^{r,0}(s),\alpha(\eta^{r,0}(s)))\right)ds \\ &=0,\end{aligned} \tag{EC-92}$$

for any $u\in[0,T]$, where the first equality of (EC-92) follows from (78), (55) and the fact that $\bar{Q}^{r,0}(s)\neq 0$ for all $s\in[0,u]\subset[0,T]$ from the assumption imposed in (93). Furthermore, the second equality of (EC-92) follows from (13), and the last equality of (EC-92) follows from (52) in the second part of Lemma 5.

Part Two: We prove the claims in the lemma for the case that $l\in\{1,...,\lceil r\delta/T\rceil-1\}$ by showing a contradiction. In fact, suppose that there is a subsequence $\mathcal{R}_1$ of r such that at least one of the claims stated in (91) and (EC-86)-(EC-87) does not hold for any $r\in\mathcal{R}_1$ and some integer $l\in\{1,...,\lceil r\delta/T\rceil-1\}$, where for later reference, we use $l_r\in\{1,...,\lceil r\delta/T\rceil-1\}$ with $r\in\mathcal{R}_1$ to denote the smallest integer to have such property. However, we can show that there is a subsequence $\mathcal{R}_2\subset\mathcal{R}_1$ such that all the claims stated in (91) and (EC-86)-(EC-87) are true for $l=l_r$ and all sufficiently large $r\in\mathcal{R}_2$. To do so, we first construct a subsequence $\mathcal{R}_3$ such that (91) is true for $l=l_r$ and all sufficiently large $r\in\mathcal{R}_3$ as follows.

From the proof in the first part, we know that the claims stated in (91) and (EC-86)-(EC-87) are true for all $l\in\{0,1,...,l_r-1\}$ and all sufficiently large $r\in\mathcal{R}_1$. Hence, for $l=l_r-1$, we have

$$\left\|\bar{Q}^{r,l_r-1}(0)\right\|\leq\max\{b_{1,1},b_{1,2}\}\text{ for all }r\in\mathcal{R}_1. \tag{EC-93}$$

Then, we know that there exists a subsequence $\bar{\mathcal{R}}_3\subset\mathcal{R}_1$ such that $\{\bar{Q}^{r,l_r-1}(0)\}$ converges along $r\in\bar{\mathcal{R}}_3$. Moreover, since $0\leq(l_r-1)/r\leq\delta/T$, the infimum exists for the number sequence $(l_r-1)/r$ over $r\in\bar{\mathcal{R}}_3$, i.e.,

$$0\leq\inf_{h\in\bar{\mathcal{R}}_3}\left(\frac{l_h-1}{h}\right)<\infty.$$

Thus, we can find a further subsequence $\mathcal{R}_3'\subset\bar{\mathcal{R}}_3\subset\mathcal{R}_1$ such that

$$0\leq\frac{l_r-1}{r}\downarrow l_\infty\equiv\inf_{h\in\bar{\mathcal{R}}_3}\left(\frac{l_h-1}{h}\right)<\infty\text{ along }r\in\mathcal{R}_3'. \tag{EC-94}$$

Then, it follows from (80) and (EC-94) that, for $u\in[0,2T]$,

$$\eta^{r,l_r-1}(u)\downarrow\tau+l_\infty T\equiv\eta_\infty\text{ as }r\to\infty\text{ along }\mathcal{R}_3'. \tag{EC-95}$$

Thus, from the definition of τ_n defined in (4), we know that $[\eta_\infty, \eta^{r,l_r-1}(u)] \subset [\tau_{n-1}, \tau_n)$ with some $n \in \{1, 2, ...\}$ for all $u \in [0, 2T]$ and sufficiently large $r \in \mathcal{R}_3'$. Moreover, from Lemma 9, there is a subsequence $\mathcal{R}_3 \subset \mathcal{R}_3'$ such that

$$(\bar{W}^{r,l_r-1}(u), \bar{Q}^{r,l_r-1}(u)) \to (\bar{W}(u), \bar{Q}(u)) \text{ with } \left\|\bar{Q}(0)\right\| \leq \max\{b_{1,1}, b_{1,2}\} \tag{EC-96}$$

u.o.c. over $u \in [0, 2T]$ along $\mathcal{R}_3$. Hence,

$$\begin{aligned} &\left\|\bar{Q}^{r,l_r-1}(u) - q^*(\bar{W}^{r,l_r-1}(u), \rho(\alpha(\eta^{r,l_r-1}(u))))\right\| \\ \leq\ & \frac{\epsilon}{3} + \left\|\bar{Q}(u) - q^*(\bar{W}(u), \rho(\alpha(\eta^{r,l_r-1}(u))))\right\| + \frac{\epsilon}{3} \end{aligned} \tag{EC-97}$$

holds over $u \in [0, 2T]$ when $r \in \mathcal{R}_3$ is sufficiently large, where we have used (EC-96) and the first part of Lemma 5 for (EC-97). Then, by (76) in Lemma 7 and (EC-95), we know that, for all $u \in [T, 2T]$ and sufficiently large $r \in \mathcal{R}_3$,

$$\left\|\bar{Q}(u) - q^*(\bar{W}(u), \rho(\alpha(\eta^{r,l_r-1}(u))))\right\| < \frac{\epsilon}{3} \tag{EC-98}$$

since $\alpha(\eta^{r,l_r-1}(u))$ keeps a constant $i \in \mathcal{K}$ for all $u \in [0, 2T]$ and sufficiently large $r \in \mathcal{R}_3$. Moreover, since

$$T \geq T_1 \geq T_{\max\{b_{1,1}, b_{1,2}\}, \epsilon/2}$$

where T_1 is defined in (86), for sufficiently large $r \in \mathcal{R}_3$ and $u \in [0, T]$, it follows from (EC-97) and (EC-98) that

$$\begin{aligned} &\left\|\bar{Q}^{r,l_r}(u) - q^*(\bar{W}^{r,l_r}(u), \rho(\alpha(\eta^{r,l_r}(u))))\right\| \\ =\ & \left\|\bar{Q}^{r,l_r-1}(T+u) - q^*(\bar{W}^{r,l_r-1}(T+u), \rho(\alpha(\eta^{r,l_r-1}(T+u))))\right\| \\ <\ & \epsilon, \end{aligned} \tag{EC-99}$$

where we have used (EC-97) for the inequality in (EC-99). Then, we know that the claim in (91) is true with $l = l_r$ for sufficiently large $r \in \mathcal{R}_3$.

Next, we divide $\mathcal{R}_3$ into the union of the following two sets, that is, $\mathcal{R}_3 = \mathcal{R}_4 \cup \mathcal{R}_5$, where

$$\mathcal{R}_4 \equiv \left\{r \in \mathcal{R}_3 : \bar{W}^{r,l_r}(u) \leq \epsilon \text{ for some } u \in [0, T]\right\}, \tag{EC-100}$$

$$\mathcal{R}_5 \equiv \left\{r \in \mathcal{R}_3 : \bar{W}^{r,l_r}(u) > \epsilon \text{ for all } u \in [0, T]\right\}. \tag{EC-101}$$

Here, we remark that at least one of $\mathcal{R}_4$ and $\mathcal{R}_5$ must contain infinite numbers. Hence, the remaining proof can be divided into the following two parts.

First, if $\mathcal{R}_4$ is infinite, then there is a fixed $u_r \in [0, T]$ for each $r \in \mathcal{R}_4$ such that

$$\bar{W}^{r,l_r}(u_r) \leq \epsilon. \tag{EC-102}$$

Moreover, there is a subset $\mathcal{R}_4' \subset \mathcal{R}_4$ such that $u_r \to u'$ as $r \to \infty$ for $r \in \mathcal{R}_4'$ and some $u' \in [0, T]$. Therefore, we have

$$\bar{W}(0) \leq \bar{W}(u') = \lim_{r \to \infty} \bar{W}^{r,l_r}(u_r) \leq \epsilon, \tag{EC-103}$$

where the first inequality in (EC-103) follows from the increasing property of $\bar{W}(\cdot)$, the equality in (EC-103) follows from (EC-96) since $\bar{W}^{r,l_r}(u_r) = \bar{W}^{r,l_r-1}(T + u_r)$, and the second inequality in (EC-103) follows from (EC-102). Thus, we have

$$\left\|\bar{Q}(0) - q^*(\bar{W}(0), \rho(\alpha(\eta_\infty)))\right\| < \epsilon, \tag{EC-104}$$

where η_∞ is defined in (EC-95) and the inequality in (EC-104) follows from (EC-96), the first part of Lemma 5, and the fact that (91) is true with $l = l_r$ for all sufficiently large $r \in \mathcal{R}'_4 \subset \mathcal{R}_3$ as discussed above. Therefore, it follows from (EC-104), (85), and (EC-103) that

$$\left\|\bar{Q}(0)\right\| \leq \left\|q^*(\bar{W}(0), \rho(\alpha(\eta_\infty)))\right\| + \epsilon \leq (c_2 + 1)\epsilon. \tag{EC-105}$$

Then, for all sufficiently large $r \in \mathcal{R}'_4$ and all $u \in [0,T]$, we have

$$\left\|\bar{Q}^{r,l_r}(u)\right\| \leq \left\|\bar{Q}(u)\right\| + \epsilon \leq b_{1,1} \tag{EC-106}$$

where $b_{1,1}$ is defined in (87) and the two inequalities in (EC-106) follow from (EC-96), the similar argument as in (EC-99), and Lemma 8 respectively. Similarly, for sufficiently large $r \in \mathcal{R}'_4$ and all $u \in [0,T]$, we have

$$\bar{W}^{r,l_r}(u) \leq \bar{W}(u) + \epsilon \leq c_1 \left\|q^*(\bar{W}(u), \rho(\alpha(\eta_\infty)))\right\| + \epsilon \leq b_{2,1}, \tag{EC-107}$$

where the two inequalities in (EC-107) follow from (EC-96) and the similar argument as in (EC-99). Then it follows from (EC-106)-(EC-107) that (EC-86) is true for $l = l_r$ for sufficiently large $r \in \mathcal{R}_4$.

Second, if $\mathcal{R}_5$ is infinite, we can choose $\sigma = \sigma(\epsilon)$ as in Lemma 5. Then, it follows from Lemma 8 that for all $u \in [0,T]$,

$$\left\|\bar{Q}(T+u) - q^*(\bar{W}(T+u), \rho(\alpha(\eta_\infty)))\right\| \leq \frac{\sigma}{2} \tag{EC-108}$$

where $\alpha(\eta_\infty) = \alpha(\eta^{r,l_r-1}(T+u))$ keeps a constant $i \in \mathcal{K}$ for all $u \in [0, 2T]$ and all sufficiently large $r \in \mathcal{R}_5$. Moreover, the chosen time T satisfies

$$T \geq T_1 \geq T_{\max\{b_{1,1}, b_{1,2}\}, \sigma/2}$$

with T_1 defined in (86). Thus, for all sufficiently large $r \in \mathcal{R}_5$ and all $u \in [0,T]$, we have

$$\left\|\bar{Q}^{r,l_r-1}(T+u) - q^*(\bar{W}^{r,l_r-1}(T+u), \rho(\alpha(\eta^{r,l_r-1}(T+u))))\right\| < \sigma, \tag{EC-109}$$

where the inequality follows from the similar explanations as used for (EC-104). Therefore, by (EC-109), (52) in the second part of Lemma 5, and the fact that

$$\bar{W}^{r,l_r-1}(T+u) = \bar{W}^{r,l_r}(u) > \epsilon \text{ for all } u \in [0,T],$$

we know that $\bar{Y}^{r,l_r-1}(T+u)$ does not increase over $u \in [0,T]$ for all sufficiently large $r \in \mathcal{R}_5$, i.e.,

$$\bar{Y}^{r,l_r}(u) - \bar{Y}^{r,l_r}(0) = 0 \text{ for all } u \in [0,T]. \tag{EC-110}$$

To finish the remaining proof based on (EC-110), we need to consider the following two mutually exclusive cases for a given large enough $r \in \mathcal{R}_5$.

Case One: the condition in (EC-87) is true for all $l \in \{0, 1, ..., l_r\}$. Then, we know that $\bar{Y}^{r,l_r}(u)$ does not increase over $u \in [0,T]$ for all $l \in \{0, 1, ..., l_r\}$ owing to the induction assumption and (EC-110). Hence, for sufficiently large $r \in \mathcal{R}_5$ and all $u \in [0,T]$, we have

$$\begin{aligned}
\bar{W}^{r,l_r}(u) &= \bar{W}^{r,0}(0) + \sum_{l=0}^{l_r-1}\left(\bar{W}^{r,l}(T) - \bar{W}^{r,l}(0)\right) + \left(\bar{W}^{r,l_r}(u) - \bar{W}^{r,l_r}(0)\right) \\
&= \hat{W}^r(\tau) + \left(\hat{X}^r(\eta^{r,l_r}(u)) - \hat{X}^r(\eta^{r,0}(0))\right) \\
&\leq (\nu + \epsilon) + (C + \epsilon),
\end{aligned} \tag{EC-111}$$

where the second equality in (EC-111) follows from (77)-(78), and the inequality in (EC-111) follows from (89), (EC-95), (79) and (90).

Case Two: the condition in (EC-86) is true for some $l \in \{0,1,...,l_r-1\}$ and use l_r^m to denote the largest such integer. Then, both the condition and the claim in (EC-87) are true for all $l \in \{l_r^m+1,...,l_r\}$, and therefore, the corresponding $\bar{Y}^{r,l}(u)$ does not increase over $u \in [0,T]$ owing to the induction assumption and the discussion as in (EC-109)-(EC-110). Moreover, by the same discussion as used in (EC-95), there is a subsequence $\mathcal{R}_5' \subset \mathcal{R}_5$ such that $\eta^{r,l_r^m}(T)$ converges along $r \in \mathcal{R}_5'$. Thus, similar to (EC-111), for sufficiently large $r \in \mathcal{R}_5'$ and all $u \in [0,T]$, we have

$$\begin{aligned} \bar{W}^{r,l_r}(u) &= \bar{W}^{r,l_r^m}(T) + \Big(\hat{X}^r(\eta^{r,l_r}(u)) - \hat{X}^r(\eta^{r,l_r^m}(T))\Big) \\ &\leq b_{2,1} + (C+\epsilon), \end{aligned} \tag{EC-112}$$

where the inequality in (EC-112) follows from (EC-96), the induction assumption (since $l_r^m < l_r$), (79), and (90).

Therefore, it follows from both of the discussions in Case One and Case Two that for sufficiently large $r \in \mathcal{R}_5'$ and all $u \in [0,T]$, we have,

$$\begin{aligned} \bar{W}^{r,l_r}(u) &\leq \max\{(\nu+\epsilon)+(C+\epsilon), b_{2,1}+(C+\epsilon)\} \\ &= b_{2,2}, \end{aligned} \tag{EC-113}$$

$$\begin{aligned} \left\|\bar{Q}^{r,l_r}(u)\right\| &\leq \left\|q^*(\bar{W}^{r,l_r}(u), \rho(\alpha(\eta^{r,l_r}(u))))\right\| + \epsilon \\ &\leq c_2\bar{W}^{r,l_r}(u) + \epsilon \\ &\leq b_{1,2}, \end{aligned} \tag{EC-114}$$

where the first inequality in (EC-114) follows from (EC-99), the second inequality in (EC-114) follows from (85), and the third inequality in (EC-114) follows from (EC-113). Thus, by (EC-110) and (EC-113)-(EC-114), we know that (EC-87) is true with $l = l_r$ for large enough $r \in \mathcal{R}_5'$.

Finally, take $\mathcal{R}_2 = \mathcal{R}_4' \cup \mathcal{R}_5'$, and then, (91) and (EC-86)-(EC-87) are true in terms of $l = l_r$ for sufficiently large $r \in \mathcal{R}_2 \subset \mathcal{R}_1$. This is a contradiction. □

10. The Remainder of Proof of Theorem 1

As in the proof of Lemma 10, our discussion will be based on each particular sample path. For convenience, we divide the proof into two parts.

Part One. In this part, we prove the convergence in distribution as stated in (35) and the related properties (33)-(36). First, it may be not true that any subsequence of $\{\hat{Y}^r(t), r \in \mathcal{R}\}$ exists a further subsequence that converges to a continuous and nondecreasing limit $\hat{Y}(t)$ when $\hat{Y}^r(t)$ are unbounded (e.g., $\hat{Y}^r(t) = (logr)t$). Hence, we employ Lemmas 7-10 to provide a justification in terms of u.o.c. convergence for $\{\hat{Y}^r(t), r \in \mathcal{R}\}$, which can be considered as a supplementary illustration to the corresponding claims used in Ye and Yao (2008), Stolyar (2004), etc. In fact, since $Q^r(0) = 0$ for all $r \in \mathcal{R}$, we can conclude that the conditions stated in (89) of Lemma 10 are satisfied. Moreover, from (79), we know that (90) is true for an arbitrarily chosen constant $C > 0$ over any given interval $[0,T] \supset [0,T_1]$, where T_1 is defined in (86). Hence, by (77), (81), and Lemmas 7-10, we know that for any $t \in [0,T]$ and each sufficiently large $r \in \mathcal{R}$, there is a $u \in [0,T]$ and $l \in \{0,1,...,\lceil r\delta/T\rceil - 1\}$ such that for any given sufficiently small $\epsilon > 0$,

$$0 \leq \hat{Y}^r(t) = \bar{W}^{r,l}(u) - \hat{X}^r(t) \leq C + O(\epsilon) + K, \tag{EC-115}$$

where K is some positive constant from (79) and the continuity of $\hat{X}(t)$. Therefore, we know that $\hat{Y}^r(t)$ is uniformly bounded over the given interval $[0,T]$ for all $r \in \mathcal{R}$. Moreover, since $\hat{Y}^r(t)$ for each $r \in \mathcal{R}$ is nondecreasing and continuous with $\hat{Y}^r(0) = 0$, it follows from Helly's Theorem (e.g., Theorem 2 in page 319 of Shiryaev (1996)) that for any subsequence of these processes, there is a further subsequence $\mathcal{R}_1 \subset \mathcal{R}$ such that

$$\hat{Y}^r(t) \to \hat{Y}(t) \text{ for every } t \in [0, T] \text{ along } r \in \mathcal{R}_1, \tag{EC-116}$$

where $\hat{Y}(t)$ is also nondecreasing and continuous with $\hat{Y}(t) = 0$ over $[0, T]$.

Next, take $T \in \{1, 2, ...\}$ and let $T \to \infty$ since $T \geq T_1$ is arbitrarily taken. We know that there is a further subsequence $\mathcal{R}_2 \subset \mathcal{R}_1$ such that the convergence in (EC-116) is extended to the whole interval $[0, \infty)$ along $r \in \mathcal{R}_2$, and $\hat{Y}(t)$ is nondecreasing and continuous over $[0, \infty)$. Thus, it follows from Theorem 2.15 on page 342, Corollary 2.24 on page 345, and Proposition 1.17(b) of Jacod and Shiryaev (2003) that the convergence in (EC-116) is u.o.c. over $[0, \infty)$. Consequently, it follows from (77) and (79) that along $r \in \mathcal{R}_2$,

$$\hat{W}^r(t) \to \hat{W}(t) = \hat{X}(t) + \hat{Y}(t) \geq 0 \text{ u.o.c. over } t \in [0, \infty), \tag{EC-117}$$

which is continuous in $t \in [0, \infty)$.

Thus, it follows from (EC-116)-(EC-117), Lemma 7, and the similar argument used in Ye and Yao (2008), that the complementary property as stated in Theorem 1 is true. Furthermore, take a number $\delta > 0$; then, for a given $\epsilon > 0$, it follows from (91) in Lemma 10 that for sufficiently large $r \in \mathcal{R}_2$,

$$\sup_{t \in [0, \delta]} \left\| \hat{Q}^r(t) - q^*(\hat{W}^r(t), \rho(\alpha(t))) \right\| \leq \epsilon, \tag{EC-118}$$

or equivalently, for each $j \in \{1, 2\}$, $t \in [0, \delta]$ and sufficiently large $r \in \mathcal{R}_2$, we have

$$q_j^*(\hat{W}^r(t), \rho(\alpha(t))) - \epsilon \leq \hat{Q}_j^r(t) \leq q_j^*(\hat{W}^r(t), \rho(\alpha(t))) + \epsilon. \tag{EC-119}$$

Then, it follows from Lemma 5 and (EC-117) that the following convergence is true (e.g., let $r \to \infty$ first and let $\epsilon \to 0$ later in (EC-119)),

$$\hat{Q}^r(t) \to \hat{Q}(t) \equiv q^*(\hat{W}(t), \rho(\alpha(t))) \text{ uniformly over } t \in [0, \delta]. \tag{EC-120}$$

Since δ is arbitrarily taken, the convergence stated in (EC-120) can be considered true u.o.c. over $[0, \infty)$. Therefore, we have

$$(\hat{Q}^r(t), \hat{W}^r(t), \hat{Y}^r(t)) \to (\hat{Q}(t), \hat{W}(t), \hat{Y}(t)) \text{ u.o.c. over } [0, \infty) \text{ along } r \in \mathcal{R}_2 \tag{EC-121}$$

with the limit satisfying all the requirements as stated in Theorem 1. Consequently, owing to the uniqueness of solution to the associated Skorohod problem (see, e.g., Chen and Yao (2001), or Dai (1996) and Dai and Dai (1999)), we know that the convergence in (EC-121) is true along $r \in \mathcal{R}$.

Part Two. In this part, we prove the optimality claims stated in (37)-(38) along the line of Ye and Yao (2008); however, the justification logic and technical treatment are somewhat different. First, suppose that all the processes related to an arbitrarily given feasible allocation scheme G will be superscripted by an additional G. Then, for each $t \in [0, \infty)$, we define

$$\hat{W}^G(t) \equiv \liminf_{r \to \infty} \hat{W}^{r,G}(t), \tag{EC-122}$$

which may be infinitely valued. In other words, for any particularly given $t \in [0, \infty)$, there exists a subsequence $\mathcal{T} \subset \mathcal{R}$ such that

$$\hat{W}^G(t) = \lim_{r \to \infty} \hat{W}^{r,G}(t) \text{ along } r \in \mathcal{T}. \tag{EC-123}$$

Moreover, let $\mathcal{Q}$ denote the set of all nonnegative rational numbers. Thus, there exists a subsequence $\mathcal{T}_1 \in \mathcal{T}$ such that

$$\hat{W}^{r,G}(s) \to \hat{W}^G(s) \text{ along } r \in \mathcal{T}_1 \text{ for each } s \in \mathcal{Q}. \tag{EC-124}$$

In addition, by applying a similar discussion as in Lemma 7, we can select a subsequence $\mathcal{T}_2 \subset \mathcal{T}_1$ such that

$$\bar{T}^{r,G}(s) \to \bar{T}^G(s) \text{ u.o.c. over } s \in [0,\infty) \text{ as } r \to \infty \text{ along } r \in \mathcal{T}_2, \tag{EC-125}$$

where $\bar{T}^G(s)$ is Lipschitz continuous and increasing with $\bar{T}^G(0) = 0$. Furthermore, we can see that $\bar{Q}^{r,G}(s)$, $\bar{W}^{r,G}(s)$, and $\bar{Y}^{r,G}(s)$ also converge u.o.c. to $\bar{Q}^G(s)$, $\bar{W}^G(s)$, and $\bar{Y}^G(s)$ along $r \in \mathcal{T}_2$, which are Lipschitz continuous and satisfy the relationships

$$\bar{Q}_j^G(s) = \bar{\lambda}_j(s) - \mu_j \bar{T}_j^G(s) \geq 0 \text{ for each } j \in \mathcal{J}, \tag{EC-126}$$

$$\bar{W}^G(s) = \sum_{j=1}^{J} \frac{\bar{Q}_j^G(s)}{\mu_j} = \bar{Y}^G(s), \tag{EC-127}$$

$$\bar{Y}^G(s) = \sum_{j=1}^{J} \left(\int_0^s \rho_j(\alpha(u))du - \bar{T}_j^G(s) \right), \tag{EC-128}$$

where $\bar{Y}^G(s)$ is nondecreasing with $\bar{Y}^G(0) = 0$. To further investigate this system, we define

$$\zeta = \inf\left\{ s \geq 0 : \bar{T}_j^G(s) \neq \bar{c}_j(s) \text{ for some } j \in \mathcal{J} \right\}, \tag{EC-129}$$

where $\bar{c}_j(s)$ is defined in (73). Then, under the policy G, it follows from a similar discussion as in (79) that

$$\hat{X}^{r,G}(s) \to \hat{X}^G(s) \text{ u.o.c. over } s \in [0,\zeta) \text{ along } r \in \mathcal{T}_2. \tag{EC-130}$$

Hence, it follows from (EC-124) that

$$\hat{Y}^{r,G}(s) \to \hat{\gamma}^G(s) \text{ along } r \in \mathcal{T}_2 \text{ for each } s \in \mathcal{Q}, \tag{EC-131}$$

where $\hat{\gamma}^G(s)$ is some discrete function in $s \in \mathcal{Q}$ and is nondecreasing since $\hat{Y}^{r,G}(s)$ is nondecreasing for each $r \in \mathcal{T}_2$. Moreover, define

$$\zeta_1 = \inf\left\{ s \geq 0 : \hat{\gamma}^G(s) = +\infty, s \in \mathcal{Q} \right\}; \tag{EC-132}$$

then, we know that $\{\hat{Y}^{r,G}(s), r \in \mathcal{T}_2\}$ is uniformly bounded over any compact set of $[0, \zeta \wedge \zeta_1)$. Thus, it follows from the similar explanation as used for (EC-116) that there is a subsequence $\mathcal{T}_3 \subset \mathcal{T}_2$ such that

$$\hat{Y}^{r,G}(s) \to \hat{Y}^G(s) \text{ for each } s \in [0, \zeta \wedge \zeta_1) \text{ along } r \in \mathcal{T}_3, \tag{EC-133}$$

where $\hat{Y}^G(s)$ is continuous and nondecreasing with $\hat{Y}^G(0) = 0$. Moreover, it satisfies

$$\hat{Y}^G(s) = \hat{\gamma}^G(s) \text{ for all } s \in \mathcal{Q} \cap [0, \zeta \wedge \zeta_1).$$

Then, it follows from (EC-130), (EC-133), and the similar expression as in (77) that along $r \in \mathcal{T}_3$ and for each $s \in [0, \zeta \wedge \zeta_1)$,

$$\hat{\beta}^G(s) \equiv \lim_{r \to \infty} \hat{W}^{r,G}(s) = \hat{X}^G(s) + \hat{Y}^G(s) \geq 0. \tag{EC-134}$$

However, the complementarity may not be true for $(\hat{W}^G(t), \hat{Y}^G(t))$. Therefore, it follows from (EC-133)-(EC-134) and the minimality of the Skorohod problem (see, e.g., Chen and Yao (2001), Dai (1996), Dai and Dai (1999), and Harrison and Reiman (1981)) that

$$\hat{\beta}^G(s) \geq \hat{W}(s) \text{ for all } s \in [0, \zeta \wedge \zeta_1). \tag{EC-135}$$

Hence, if $t \in [0, \zeta \wedge \zeta_1)$, we know that, along $r \in \mathcal{T}_3$,

$$\hat{W}^G(t) = \lim_{r\to\infty,\, r\in\mathcal{T}} \hat{W}^{r,G}(t) = \lim_{r\to\infty,\, r\in\mathcal{T}_3} \hat{W}^{r,G}(t) = \hat{\beta}^G(t) \geq \hat{W}(t), \tag{EC-136}$$

which is always true if $\zeta = \zeta_1 = \infty$.

Furthermore, if $\zeta < \zeta_1$ or $\zeta = \zeta_1 < \infty$, and $t \in [\zeta, \infty)$, we can take time $\tau \in [\zeta, t]$ such that $\bar{T}_j^G(\tau) \neq \bar{c}_j(\tau)$ for some $j \in \mathcal{J}$. Hence, it follows from (EC-126) that $\bar{T}_j^G(\tau) < \bar{c}_j(\tau)$ and $\bar{Q}_j^G(\tau) > 0$ for the j. Then, it follows from (EC-127)-(EC-128) that $\bar{W}^G(t) \geq \bar{W}^G(\tau) > 0$. Therefore, along $r \in \mathcal{T}_3$, we have

$$\hat{W}^G(t) = \lim_{r\to\infty,\, r\in\mathcal{R}_2} \hat{W}^{r,G}(t) = \lim_{r\to\infty,\, r\in\mathcal{R}_2} r\bar{W}^{r,G}(t) = +\infty \geq \hat{W}(t). \tag{EC-137}$$

In addition, if $\zeta > \zeta_1$ and $t \in [\zeta_1, \infty)$, it follows from (EC-129) that

$$\liminf_{r\to\infty,\, r\in\mathcal{T}_3} \hat{Y}^{r,G}(t) \geq \lim_{r\to\infty,\, r\in\mathcal{T}_3} \hat{Y}^{r,G}(\zeta_1) = \hat{\gamma}^G(\zeta_1) = +\infty. \tag{EC-138}$$

Thus, by (EC-130), we know that

$$\hat{W}^G(t) = \lim_{r\to\infty,\, r\in\mathcal{T}_3} \hat{W}^{r,G}(t) = +\infty \geq \hat{W}(t). \tag{EC-139}$$

Since the given time $t \in [0, \infty)$ is arbitrarily taken, it follows from (EC-136), (EC-137), and (EC-139) that the claim (37) in the theorem is true for any $t \geq 0$.

Finally, it follows from (37) and (36) that (38) is true. $\square$

Acknowledgments

The author thanks the editors and two anonymous referees for helpful comments and suggestions. The author also gratefully acknowledges the financial support from National Natural Science Foundation of China with Grant No. 10971249 and Grant No. 11371010.

References

Bhardwaj, S., Williams, R. J., and Acampora, A. S. 2007. On the performance of a two-user MIMO downlink system in heavy traffic. *IEEE Transactions on Information Theory* **53**(5) 1851-1859.

Bhardwaj, S. and Williams, R. J. 2009. Diffusion approximation for a heavily loaded multi-user wireless communication system with coorperation. *Queueing Systems* **62** 345-382.

Billingsley, P. 1999. Convergence of Probability Measures, 2nd ed., John Wiley & Sons, New York.

Bramson, M. and Dai, J. G. 2001. Heavy traffic limits for some queueing networks. *Annals of Applied Probability* **11** (1) 49-90.

Chen, H. and Yao, D. D. 2001. Fundamentals of Queueing Networks: Performance, Asymptotics, and Optimization. Springer-Verlag, New York.

Dai, J. G. 1995. On positive harris recurrence of multiclass queueing networks: a unified approach via fluid limit models. *Annals of Applied Probability* **5**(1) 49-77.

Dai, J. G. and Dai, W. 1999. A heavy traffic limit theorem for a class of open queueing networks with finite buffers. *Queueing Systems* **32** 5-40.

Dai, W. 1996. Brownian Approximations for Queueing Networks with Finite Buffers: Modeling, Heavy Traffic Analysis and Numerical Implementations. *Ph.D Thesis*, Georgia Institute of Technology, 1996, Also published in UMI Dissertation Services, A Bell & Howell Company, Michigan, U.S.A., 1997.

Dai, W. 2007. Diffusion approximations for multiclass queueing networks under preemptive priority service discipline. *Applied Mathematics and Mechanics* **28**(10) 1331-1342.

Ethier, S. N. and Kurtz, T. G. 1986. Markov Processes: Characterization and Convergence, John Wiley & Sons Inc., New York.

Goldsmith, A., Jafar, S. A., Jindal, and Vishwanath, N. 2003. Capacity limits of MIMO Channels. *IEEE Journal on Selected Areas in Communications* **21**(5) 684-702.

Harrison, J. M. and Reiman, M. I. 1981. Reflected Brownian motion on an orthant. *Annals of Probability* **9**(2) 302-308.

Hiriart-Urruty, J. B. and C. Lemaréchal, C. 2001. Fundamentals of Convex Analysis, Springer-Verlag, Berlin.

Iglehart, D. L. and Whitt, W. 1971. The equivalence of functional central limit theorems for counting processes and associated partial sums. *Ann. Math. Statist.* **42**(4) 1372-1378.

Jacod, J. and Shiryaev, A. N. 2003. Limit Theorems for Stochastic Processes, Second Edition, Springer-Verlag, Berlin.

Liu, J. and Hou, Y. T. 2008. Weighted proportional fairness capacity of Gaussian MIMO broadcast channels. *Proceedings of IEEE INFOCOM 2008* 1058-1066, IEEE Computer Society Press.

Luenberger, D. G. 1984. Linear and Nonlinear Programming, Second Edition, Addison-Wesley Publishing Company, Reading, Massachusetts.

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

Prokhorov, Y. V. 1956. Convergence of random processes and limit theorems in probability theory. *Theory Probab. Appl.* **1**(2) 157-214.

Shakkotai, S., Srikant, R., and Stolyar, A. L. 2004. Pathwise optimality of the exponential scheduling rule for wireless channels. *Advances in Appl. Probability* **36**(4) 1021-1045.

Shiryaev, A. N. 1996. Probaility, Second Edition, Springer-Verlag, New York.

Stolyar, A. L. 2004. MaxWeight scheduling in a generalized switch: state space collapse and workload minimization in heavy traffic. *The Annals of Applied Probability* **14**(1) 1-53.

Vishwanath, S., Jindal, N., and Goldsmith, A., Duality, achievable rates and sum-rate capacity of Gaussian MIMO broadcast channels, IEEE Trans. Inform. Theory, vol. 49, pp. 2658-2668, Oct. 2003.

Whitt, W. 2002. Stochastic-Process Limits. Springer, New York, 2002.

Ye, H. and Yao, D. D. 2008. Heavy traffic optimality of a stochastic network under utility-maximizing resource control. *Operations Research* **56**(2) 453-470.

Yu, W., Rhee, W., Boyd, S., and Cioffi, J. 2004. Iterative water-filling for vector multiple access channels. *IEEE Transactions on Information Thoery* **50**(1) 145-151.