Network Utility Maximization over Partially Observable Markovian Channels

Chih-ping Li, *Student Member, IEEE* and Michael J. Neely, *Senior Member, IEEE*

***Abstract*—This paper considers maximizing throughput utility in a multi-user network with partially observable Markov ON/OFF channels. Instantaneous channel states are never known, and all control decisions are based on information provided by ACK/NACK feedback from past transmissions. This system can be viewed as a restless multi-armed bandit problem with a concave objective function of the time average reward vector. Such problems are generally intractable. However, we provide an approximate solution by optimizing the concave objective over a non-trivial inner bound on the network performance region, where the inner bound is constructed by randomizing well-designed stationary policies. Using a new frame-based Lyapunov drift argument, we design a policy of admission control and channel selection that stabilizes the network with throughput utility that can be made arbitrarily close to the optimal in the inner performance region. Our problem has applications in limited channel probing in wireless networks, dynamic spectrum access in cognitive radio networks, and target tracking of unmanned aerial vehicles. Our analysis generalizes the MaxWeight-type scheduling policies in stochastic network optimization theory from time-slotted systems to frame-based systems that have policy-dependent frame sizes.**

I. Introduction

This paper studies a multi-user wireless scheduling problem in a partially observable environment. We consider a base station serving N users via N independent Markov ON/OFF channels (see Fig. 1). Time is slotted with normalized slots

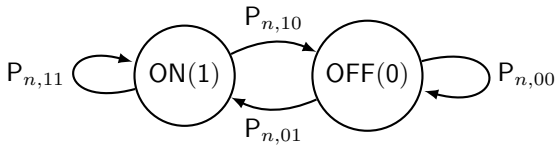

Fig. 1. The Markov ON/OFF model for channel $n \in \{1, 2, \ldots, N\}$.

$t \in \mathbb{Z}^+$. Channel states are fixed in every slot, and can only change at slot boundaries. Suppose the base station has unlimited data to send for all users. In every slot, the channel states are unknown, and the base station selects at most one user to which it blindly transmits a packet. The transmission succeeds if the used channel is ON, and fails otherwise. At the end of a slot, an error-free ACK is fed back from the served user to the base station over an independent control channel (absence of an ACK is considered as a NACK). Since channels are ON/OFF and correlated over time, the ACK/NACK feedback provides partial information of future channel states, and can improve future scheduling decisions for better performance. The goal is to design a control policy that maximizes a concave utility function of the throughput vector from all channels. Specifically, let $y_n(t)$ be the number of packets delivered to user $n \in \{1, \ldots, N\}$ in slot t. Define $\overline{y}_n \triangleq \lim_{t\to\infty} \frac{1}{t} \sum_{\tau=0}^{t-1} \mathbb{E}[y_n(\tau)]$ as the throughput of user n. Let Λ denote the *network capacity region*, defined as the closure of the set of all achievable throughput vectors $\overline{\boldsymbol{y}} \triangleq (\overline{y}_n)_{n=1}^N$. The goal is to solve:

$$\text{maximize:} \quad g(\overline{\boldsymbol{y}}), \quad \text{subject to:} \quad \overline{\boldsymbol{y}} \in \Lambda, \tag{1}$$

where $g(\cdot)$ is a concave, continuous, nonnegative, and nondecreasing function.

The interest in the above problem comes from its many applications. One application is *limited channel probing* over wireless networks. Consider the same wireless downlink as stated above, except that at most one channel is explicitly probed in every slot. A packet is served over the probed channel if the state is ON. This setup is essentially the same as our original problem, except that channels are probed differently (implicit probing by ACK/NACK feedback versus probing by explicit signaling). The motivation for studying limited channel probing is that, in a fast-changing environment where full channel probing may be infeasible due to timing overhead, we shall probe channels wisely and exploit channel memory to improve network performance. As an example of (1), we may additionally provide fairness to all users, such as a variant of rate proportional fairness [1], [2] by solving:

$$\text{maximize: } g(\overline{\boldsymbol{y}}) = \sum_{n=1}^{N} \log(1+\overline{y}_n), \quad \text{subject to: } \overline{\boldsymbol{y}} \in \Lambda. \tag{2}$$

In *cognitive radio networks* [3], [4], a secondary user has access to a collection of Markov ON/OFF channels. Every channel reflects the occupancy of a spectrum by a primary user, and the secondary user opportunistically transmits data over unused spectrums for better spectrum efficiency. In *target tracking of unmanned aerial vehicles* (UAVs) [5], a UAV detects one of the many targets in every slot. Every Markov channel reflects the movement of a target; a channel is ON if its associated target moves to a hotspot, and OFF otherwise. Delivering a packet over a channel represents gaining a reward by locating a target at its hotspot. In the above two applications, possible goals include maximizing a weighted

Chih-ping Li (web: http://www-scf.usc.edu/∼chihpinl) and Michael J. Neely (web: http://www-rcf.usc.edu/∼mjneely) are with the Department of Electrical Engineering, University of Southern California, Los Angeles, CA 90089, USA.

This material is supported in part by one or more of the following: the DARPA IT-MANET grant W911NF-07-0028, the NSF Career grant CCF-0747525, and the Network Science Collaborative Technology Alliance sponsored by the U.S. Army Research Laboratory.

sum $g(\overline{\boldsymbol{y}}) = \sum_{n=1}^{N} c_n \overline{y}_n$ of throughputs/time-average rewards, where c_n are given constants, or providing fairness to different spectrums/targets by solving (2).

The problem (1) is challenging because the current information available for each channel depends on the past transmission decisions. This problem belongs to the class of restless multi-armed bandit (RMAB) problems [6], which are generally intractable [7]. In addition, the network capacity region Λ in (1) does not seem to have a closed form expression (see [8] for more discussions). Therefore we must resort to approximation methods to solve (1). In this paper, we propose an *achievable region* approach to construct an approximate solution to (1). There are two steps: (i) we construct a good inner performance region $\Lambda_{\text{int}} \subset \Lambda$ for the original problem, then (ii) we solve the constrained problem:

$$\text{maximize:} \quad g(\overline{\boldsymbol{y}}), \quad \text{subject to:} \quad \overline{\boldsymbol{y}} \in \Lambda_{\text{int}}, \tag{3}$$

which serves as an approximation to the original problem (1).

In previous work [8], we have constructed a non-trivial inner performance region Λ_{int} using the rich structure of Markov channels (see Section III for details). The inner performance region Λ_{int} is rendered as a convex hull of performance vectors of a well-designed collection of *round robin* policies. The tightness of the inner region Λ_{int} (see Fig. 2 for an example) is analyzed in [8] when channels are statistically identical. In this special case we show that the gap between the boundary of the inner region Λ_{int} and that of the full performance region Λ decreases to zero geometrically fast as the reference direction moves closer to the 45-degree angle.[1]

The main contribution of this paper is with respect to the second step of the achievable region approach: Given an inner performance region Λ_{int}, we construct a policy that solves (3) using *Lyapunov drift theory*. Lyapunov drift theory is originally developed for throughput optimal control over time-slotted wireless networks [9], [10], later extended to optimize various performance objectives such as average power [11] or rate utility functions [1], [12], [13] in wireless networks, and recently generalized to optimize dynamic systems that have a renewal structure [14]–[16]. The intuition is the following. Since the performance region Λ_{int} is a convex hull of performance vectors of the round robin policies we design, the problem (3) is solved by some random mixture of these policies. Using Lyapunov drift theory (see more details in Section IV), we greedily construct a sequence of round robin policies whose long-term time sharing can approximate the optimal solution as close as desired, with some tradeoffs discussed later.

Our control policy that solves (3) has two components. To facilitate the solution to (3), we keep an infinite-capacity queue for every user at the base station, and design an admission control algorithm that admits data into the queues for eventual transmissions. In every slot, the amount of data admitted for every user is decided by solving a simple convex program.[2] In addition, the base station deploys a sequence of round robin policies implemented frame by frame, where every frame is one round of execution by a round robin policy. The round robin policy used in every frame is selected by maximizing an average "drift minus reward" ratio over the average frame size (c.f. (18)). We emphasize that this new ratio rule generalizes the MaxWeight-type policies [1], [10] for stochastic network optimization from time-slotted wireless networks to frame-based systems in which the distribution of the random frame size is policy-dependent. We prove that the above policy of admission control and channel scheduling yields a throughput vector $\overline{\boldsymbol{y}}$ satisfying

$$g(\overline{\boldsymbol{y}}) \geq g(\overline{\boldsymbol{y}}^*) - \frac{B}{V}, \tag{4}$$

where $g(\overline{\boldsymbol{y}}^*)$ is the optimal objective of problem (3), $B > 0$ is a finite constant, $V > 0$ is a predefined control parameter, and we temporarily assume that all limits exist. By choosing V sufficiently large in (4), the performance utility $g(\overline{\boldsymbol{y}})$ can be made arbitrarily close to the optimal $g(\overline{\boldsymbol{y}}^*)$, with the tradeoff that the average queue size at the base station grows linearly with V. We remark that the proof of (4) does not require the knowledge of the optimal utility $g(\overline{\boldsymbol{y}}^*)$.

In the literature, stochastic utility maximization over wireless networks is solved in [1], [12], assuming that channel states are i.i.d. over slots and are known perfectly and instantly. Limited channel probing in wireless networks is studied in different contexts in [17]–[22], also assuming that channel states are i.i.d. over time. This paper generalizes the framework in [1] to wireless networks with limited channel probing and time-correlated channels, and uses channel memory to improve performance.

RMAB problems with Markov ON/OFF projects are previously studied in [23]–[28] for the maximization of sum of time average or discounted rewards. In particular, work [23]–[25] shows that greedy round robin policies are optimal in some special cases; we modify these policies in [8] for the construction of a tractable inner performance region Λ_{int}. Index policies such Whittle's index [6] are constructed in [26], [27], and are shown to have good performance by simulations. A $(2+\epsilon)$-approximate algorithm is derived in [28] based on duality methods.

This paper provides a new mathematical programming method for optimizing nonlinear objective functions of time average rewards in RMAB problems. In the literature RMAB problems are mostly studied with linear objective functions. The two popular methods for linear RMABs — Whittle's index [6] and (partially observable) Markov decision theory [29] — seem difficult to apply to nonlinear RMABs because they are based on dynamic programming ideas. Extensions of our new method in this paper to other RMAB problems with general project state space are left for future research.

[1] We remark that the tightness of the inner region Λ_{int} is difficult to check in general cases, although the region is intuitively large by the nature of its construction. The bottomline is, constructing an intuitively good and easily achievable inner performance region is of practical interest, because satisfying performance outside the inner region may inevitably involve solving much more complicated partially observable Markov decision processes. From this view, in intractable RMAB problems, we may regard an inner performance region as an *operational* performance region, which shall be gradually improved by a deeper investigation into the problem structure.

[2] The admission control decision decouples into separable one-dimensional convex programs that are easily solved in real time when the throughput utility $g(\overline{\boldsymbol{y}})$ is a sum of one-dimensional utility functions.

In the rest of the paper, the detailed network model is in the next section. Section III introduces the inner performance region Λ_{int} constructed in [8]. Our dynamic control policy is motivated and given in Section IV, followed by performance analysis.

II. Detailed Network Model

Beside the network model introduced in the previous section, we suppose that every Markov ON/OFF channel $n \in \{1, \ldots, N\}$ changes states at slot boundaries by the transition probability matrix

$$\mathbf{P}_n = \begin{bmatrix} \mathsf{P}_{n,00} & \mathsf{P}_{n,01} \\ \mathsf{P}_{n,10} & \mathsf{P}_{n,11} \end{bmatrix},$$

where state ON is represented by 1 and OFF by 0, and $\mathsf{P}_{n,ij}$ denotes the transition probability from state i to j. Assume that the matrices $\mathbf{P}_n$ are known. We assume that every channel is *positively correlated* over time, so that an ON state is more likely followed by the same state. An equivalent mathematical definition is $\mathsf{P}_{n,01} + \mathsf{P}_{n,10} < 1$ for all n.

We suppose that every user has a higher-layer data source of unlimited packets at the base station. The base station keeps a network-layer queue $Q_n(t)$ of infinite capacity for every user $n \in \{1, \ldots, N\}$, where $Q_n(t)$ denotes the backlog for user n in slot t. In every slot, the base station admits $r_n(t) \in [0, 1]$ packets for user n from its data source into queue $Q_n(t)$. For simplicity, we assume that $r_n(t)$ takes real values in $[0, 1]$ for all n.[3] Let $\mu_n(t) \in \{0, 1\}$ denote the service rate for user n in slot t. The queueing process $\{Q_n(t)\}_{t=0}^{\infty}$ of user n evolves as

$$Q_n(t+1) = \max[Q_n(t) - \mu_n(t), 0] + r_n(t). \tag{5}$$

Initially $Q_n(0) = 0$ for all n. We say queue $Q_n(t)$ is (strongly) stable if its limiting average backlog is finite, i.e.,

$$\limsup_{t\to\infty} \frac{1}{t} \sum_{\tau=0}^{t-1} \mathbb{E}\left[Q_n(\tau)\right] < \infty.$$

The network is stable if all queues $(Q_1(t), \ldots, Q_N(t))$ are stable. Clearly a sufficient condition for stability is:

$$\limsup_{t\to\infty} \frac{1}{t} \sum_{\tau=0}^{t-1} \sum_{n=1}^{N} \mathbb{E}\left[Q_n(\tau)\right] < \infty. \tag{6}$$

Our goal is to design a policy that admits the right amount of data into the network and serves them properly by channel scheduling, so that the network is stable with throughput utility that can be made arbitrarily close to the optimal solution to the problem (3).

III. A Performance Inner Bound

This section presents an inner performance region Λ_{int} constructed in previous work [8] using randomized round robin policies; see [8] for details. For every channel $n \in \{1, \ldots, N\}$, let $\mathsf{P}_{n,ij}^{(k)}$ denote the k-step transition probability from state i to j, and $\pi_{n,\mathsf{ON}}$ be the stationary probability of state ON. We define the information state for user n in slot t, denoted by $\omega_n(t)$, as the conditional probability that channel n is ON in slot t given all past channel observations. Namely,

$$\omega_n(t) \triangleq \Pr\left[s_n(t) = \mathsf{ON} \mid \text{channel observation history}\right],$$

where $s_n(t)$ denotes the state of channel n in slot t. Conditioning on the most recent channel observation, we observe that $\omega_n(t)$ takes values in the countably infinite set $\mathcal{W}_n \triangleq \{\mathsf{P}_{n,01}^{(k)}, \mathsf{P}_{n,11}^{(k)} : k \in \mathbb{N}\} \cup \{\pi_{n,\mathsf{ON}}\}$. The information state vector $(\omega_n(t))_{n=1}^N$ is a sufficient statistic [29]; it is optimal to act based only on the $(\omega_n(t))_{n=1}^N$ information. Let $n(t)$ denote the channel observed in slot t via ACK/NACK feedback. The probability $\omega_n(t)$ on channel $n \in \{1, \ldots, N\}$ evolves as:

$$\omega_n(t+1) = \begin{cases} \mathsf{P}_{n,01}, & \text{if } n = n(t),\ s_n(t) = \mathsf{OFF} \\ \mathsf{P}_{n,11}, & \text{if } n = n(t),\ s_n(t) = \mathsf{ON} \\ \omega_n(t)\mathsf{P}_{n,11} + (1 - \omega_n(t))\mathsf{P}_{n,01}, & \text{if } n \neq n(t). \end{cases} \tag{7}$$

A. Randomized round robin policy

Let Φ denote the set of all nonzero N-dimensional binary vectors. Every vector $\boldsymbol{\phi} \triangleq (\phi_n)_{n=1}^N \in \Phi$ represents a collection of *active* channels, where we say channel n is active if $\phi_n = 1$. Let $M(\boldsymbol{\phi})$ denote the number of ones (or active channels) in $\boldsymbol{\phi}$. Consider the next dynamic round robin policy $\mathsf{RR}(\boldsymbol{\phi})$ that serves active channels in $\boldsymbol{\phi}$, possibly with different order in different rounds. This is the building block of randomized round robin policies that we will introduce shortly.

Dynamic Round Robin Policy $\mathsf{RR}(\boldsymbol{\phi})$:

1) In every round, we assume an ordering of active channels in $\boldsymbol{\phi}$ is given.
2) When switching to an active channel n,
   - With probability $\mathsf{P}_{n,01}^{(M(\boldsymbol{\phi}))}/\omega_n(t)$, we keep transmitting packets over channel n until a NACK is received, after which we switch to the next active channel according to the predefined ordering.
   - Otherwise, we transmit over channel n a dummy packet with no information content for one slot (used for channel sensing), then switch to the next active channel.
3) Update probabilities $(\omega_n(t))_{n=1}^N$ by (7) in every slot.

These $\mathsf{RR}(\boldsymbol{\phi})$ policies are carefully designed to have good and, more importantly, tractable performance.

Work [24] shows that, when channels are statistically identical, serving all channels by greedy round robin policies (different from the above) maximizes the sum throughput of the network. Thus, intuitively, we get a good achievable throughput region Λ_{int} by randomly mixing round robin policies each of which serves a different subset of channels.

Randomized Round Robin Policy RandRR:

1) In every round, pick a binary vector $\boldsymbol{\phi} \in \Phi \cup \{\mathbf{0}\}$ with some probability $\alpha_{\boldsymbol{\phi}}$, where $\alpha_{\mathbf{0}} + \sum_{\boldsymbol{\phi}\in\Phi} \alpha_{\boldsymbol{\phi}} = 1$.
2) If a vector $\boldsymbol{\phi} \in \Phi$ is selected, run policy $\mathsf{RR}(\boldsymbol{\phi})$ for one round using the channel ordering of *least recently used*

[3] We can accommodate the integer-value assumption of $r_n(t)$ by introducing *auxiliary queues*; see [1] for an example.

first. Otherwise, $\boldsymbol{\phi} = \mathbf{0}$, idle the system for one slot. At the end of either case, go to Step 1.

We note that, in every round of a RandRR policy, a RR($\boldsymbol{\phi}$) policy is feasible only if $\mathsf{P}_{n,01}^{(M(\boldsymbol{\phi}))} \leq \omega_n(t)$ whenever an active channel n starts service (see Step 2 of the RR($\boldsymbol{\phi}$) policy). This condition is guaranteed by serving active channels in the order of least recently used first [8, Lemma 6]. Thus all RandRR policies are feasible.[4]

The following results present the amount of service opportunities provided by a RandRR policy to every user n.

Theorem 1 ([8]). *(i) In every round of a* RandRR *policy, when policy* RR($\boldsymbol{\phi}$) *is randomly chosen for service, let $L_n^{\boldsymbol{\phi}}$ denote the time duration an active channel n is accessed. The duration $L_n^{\boldsymbol{\phi}}$ has the probability distribution:*

$$L_n^{\boldsymbol{\phi}} = \begin{cases} 1 & \text{with prob. } 1 - \mathsf{P}_{n,01}^{(M(\boldsymbol{\phi}))} \\ j \geq 2 & \text{with prob. } \mathsf{P}_{n,01}^{(M(\boldsymbol{\phi}))} \left(\mathsf{P}_{n,11}\right)^{(j-2)} \mathsf{P}_{n,10} \end{cases}$$

and

$$\mathbb{E}\left[L_n^{\boldsymbol{\phi}}\right] = 1 + \frac{\mathsf{P}_{n,01}^{(M(\boldsymbol{\phi}))}}{\mathsf{P}_{n,10}}. \tag{8}$$

(ii) In the duration $L_n^{\boldsymbol{\phi}}$, channel n serves $(L_n^{\boldsymbol{\phi}} - 1)$ packets.

Theorem 1 shows that the distribution of $L_n^{\boldsymbol{\phi}}$ is independent of the information state vector $(\omega_n(t))_{n=1}^N$ at the start of a transmission round; it only depends on the number of channels, $M(\boldsymbol{\phi})$, chosen for service in a round. This observation implies that the transmission rounds in a RandRR policy have i.i.d. durations. Moreover, for every fixed user n, the number of user-n packets served in a round is also i.i.d. over different rounds. This leads to the following corollary.

Corollary 1. *(i) Let T_k denote the duration of the kth transmission round in a* RandRR *policy. The random variables T_k are i.i.d. over different k with*

$$\mathbb{E}\left[T_k\right] = \alpha_{\mathbf{0}} + \sum_{\boldsymbol{\phi} \in \Phi} \alpha_{\boldsymbol{\phi}} \left(\sum_{n:\phi_n = 1} \mathbb{E}\left[L_n^{\boldsymbol{\phi}}\right] \right),$$

which is computed by conditioning on the policy chosen in a round.

(ii) Let $N_{n,k}$ denote the number of packets served for user n in round T_k. For each fixed n, the random variables $N_{n,k}$ are i.i.d. over different k with $\mathbb{E}\left[N_{n,k}\right] = \sum_{\boldsymbol{\phi}:\phi_n=1} \alpha_{\boldsymbol{\phi}} \mathbb{E}\left[L_n^{\boldsymbol{\phi}} - 1\right]$, which is computed by conditioning on the RR($\boldsymbol{\phi}$) *policy that is chosen and uses channel n.*

(iii) Because $N_{n,k}$ and T_k are i.i.d. over k, the throughput of user n under a RandRR *policy is equal to $\mathbb{E}\left[N_{n,k}\right] / \mathbb{E}\left[T_k\right]$.*

B. The inner performance region Λ_{int}

In this paper, we define the inner performance region Λ_{int} in (3) as the set of all throughput vectors achieved by the class of RandRR policies. Equivalently, the inner throughput region Λ_{int} can be viewed as a convex hull of the zero vector and the performance vectors of the subset of RandRR policies, each of which executes a fixed RR($\boldsymbol{\phi}$) policy in every round. A closed form expression of the inner region Λ_{int} is given in [8, Theorem 1]. An example is given next.

Consider a two-user system with statistically identical channels with $\mathsf{P}_{01} = \mathsf{P}_{10} = 0.2$. Fig. 2 shows the tightness of the inner throughput region Λ_{int} compared to the (unknown) full network capacity region Λ. We note that points B, A, and C in Fig. 2 maximize the sum throughput of the network in directions $(0, 1)$, $(1, 1)$, and $(1, 0)$, respectively [24]. Thus the boundary of Λ is a concave curve connecting these points.

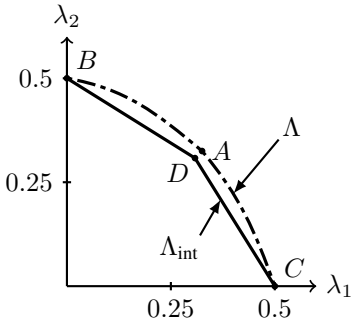

Fig. 2. The closeness of the inner throughput region Λ_{int} and the network capacity region Λ in a two-user network with statistically identical channels.

IV. Network Utility Maximization

A. The QRRNUM policy

Following the above discussions, the problem (3) is a well-defined convex program. Yet, solving (3) is difficult because the performance region Λ_{int} is represented as a convex hull of 2^N performance vectors. The following admission control and channel scheduling policy solves (3) in a dynamic manner with low complexity.

Queue-dependent Round Robin for Network Utility Maximization (QRRNUM):

- (Admission control) At the start of every round, observe the current queue backlog $\boldsymbol{Q}(t) = (Q_1(t), \ldots, Q_N(t))$ and solve the convex program

$$\text{maximize:} \quad V\, g(\boldsymbol{r}) - \sum_{n=1}^{N} Q_n(t)\, r_n \tag{9}$$

$$\text{subject to:} \quad r_n \in [0, 1],\ \forall n \in \{1, \ldots, N\}, \tag{10}$$

  where $V > 0$ is a predefined control parameter, and vector $\boldsymbol{r} \triangleq (r_n)_{n=1}^N$. Let $(r_n^{\mathsf{QRR}})_{n=1}^N$ denote the solution to (9)-(10). In every slot of the current round, admit r_n^{QRR} packets into queue $Q_n(t)$ for every user $n \in \{1, \ldots, N\}$.
- (Channel scheduling) At the start of every round, over all nonzero binary vectors $\boldsymbol{\phi} = (\phi_n)_{n=1}^N \in \Phi$, let $\boldsymbol{\phi}^{\mathsf{QRR}}$ be the maximizer of the ratio

$$\frac{\sum_{n=1}^N Q_n(t) \mathbb{E}\left[L_n^{\boldsymbol{\phi}} - 1\right] \phi_n}{\sum_{n=1}^N \mathbb{E}\left[L_n^{\boldsymbol{\phi}}\right] \phi_n}, \tag{11}$$

[4]The feasibility of RandRR policies is proved in [8] under the special case that there are no idle operations ($\alpha_{\mathbf{0}} = 0$). Using the monotonicity of the k-step transition probabilities $\{\mathsf{P}_{n,01}^{(k)}, \mathsf{P}_{n,11}^{(k)}\}$, the feasibility can be similarly proved for the extended RandRR policies considered here.

where $\mathbb{E}\left[L_n^{\boldsymbol{\phi}}\right]$ is given in (8). If the maximum of (11) is positive, run policy $\mathsf{RR}(\boldsymbol{\phi}^{\mathsf{QRR}})$ for one round using the channel ordering of least recently used first. Otherwise, idle the system for one slot. At the end of either case, start a new round of admission control and channel scheduling.

When the utility function $g(\cdot)$ is a sum of individual utilities, i.e., $g(\boldsymbol{r}) = \sum_{n=1}^{N} g_n(r_n)$, problem (9)-(10) decouples into N one-dimensional convex programs, each of which maximizes the weighted difference $[Vg_n(r_n) - Q_n(t)r_n]$ over $r_n \in [0,1]$, which can be solved efficiently in real time.

The most complex part of the QRRNUM policy is to maximize the ratio (11). In the following we present a bisection algorithm [16, Section 7.3.1] that searches for the maximum of (11) with exponentially fast speed. This algorithm is motivated by the next lemma.

Lemma 1. *([16, Lemma 7.5]) Let $a(\boldsymbol{\phi})$ and $b(\boldsymbol{\phi})$ denote the numerator and denominator of* (11)*, respectively. Define*

$$\theta^* \triangleq \max_{\boldsymbol{\phi}\in\Phi}\left\{\frac{a(\boldsymbol{\phi})}{b(\boldsymbol{\phi})}\right\},\ c(\theta) \triangleq \max_{\boldsymbol{\phi}\in\Phi}\left[a(\boldsymbol{\phi}) - \theta b(\boldsymbol{\phi})\right].$$

Then the following is true: (1) If $\theta = \theta^*$, then $c(\theta) = 0$. (2) If $\theta < \theta^*$, then $c(\theta) > 0$. (3) If $\theta > \theta^*$, then $c(\theta) < 0$.

The value $c(\theta)$ can be easily computed by noticing

$$c(\theta) = \max_{k\in\{1,\ldots,N\}}\left\{\max_{\boldsymbol{\phi}\in\Phi_k}\left[a(\boldsymbol{\phi}) - \theta b(\boldsymbol{\phi})\right]\right\}, \tag{12}$$

where $\Phi_k \subset \Phi$ denotes the set of binary vectors having k ones. The inner maximum of (12) is equal to

$$\max_{\boldsymbol{\phi}\in\Phi_k}\left\{\sum_{n=1}^{N}\left[\frac{\mathsf{P}_{n,01}^{(k)}}{\mathsf{P}_{n,10}}\left(Q_n(t)-\theta\right)-\theta\right]\phi_n\right\},$$

which is solved by sorting the values $\left[\frac{\mathsf{P}_{n,01}^{(k)}}{\mathsf{P}_{n,10}}\left(Q_n(t)-\theta\right)-\theta\right]$.

Intuition from Lemma 1: To search for the optimal ratio θ^*, suppose initially we know $\theta^* \in [\theta_{\min}, \theta_{\max}]$ for some $\theta_{\min}$ and $\theta_{\max}$. We compute the midpoint $\theta_{\mathrm{mid}} = \frac{1}{2}(\theta_{\min} + \theta_{\max})$ and evaluate $c(\theta_{\mathrm{mid}})$. If $c(\theta_{\mathrm{mid}}) > 0$, we have $\theta_{\mathrm{mid}} < \theta^*$ and thus $\theta^* \in [\theta_{\mathrm{mid}}, \theta_{\max}]$; one such bisection operation reduces the feasible region of θ^* by half. By iterating the bisection, we can find θ^* quickly. Notice that the maximizer of $c(\theta^*)$ is the desired policy $\boldsymbol{\phi}^{\mathsf{QRR}}$, since by definition we have $a(\boldsymbol{\phi}) - \theta^* b(\boldsymbol{\phi}) \le 0$ for all $\boldsymbol{\phi} \in \Phi$ and $a(\boldsymbol{\phi}^{\mathsf{QRR}}) - \theta^* b(\boldsymbol{\phi}^{\mathsf{QRR}}) = 0$.

The bisection algorithm that maximizes (11)**:**

- Initially, define $\theta_{\min} \triangleq 0$ and

$$\theta_{\max} \triangleq \frac{\left(\sum_{n=1}^{N} Q_n(t)\right)\max_{n\in\{1,\ldots,N\}}\left\{\frac{\pi_{n,\mathsf{ON}}}{\mathsf{P}_{n,10}}\right\}}{1+\min_{n\in\{1,\ldots,N\}}\left\{\frac{\mathsf{P}_{n,01}}{\mathsf{P}_{n,10}}\right\}}$$

  so that $\theta_{\min} \le a(\boldsymbol{\phi})/b(\boldsymbol{\phi}) \le \theta_{\max}$ for all $\boldsymbol{\phi} \in \Phi$. It follows that $\theta^* \in [\theta_{\min}, \theta_{\max}]$.[5]
- Compute $\theta_{\mathrm{mid}} = \frac{1}{2}(\theta_{\min} + \theta_{\max})$ and $c(\theta_{\mathrm{mid}})$. If $c(\theta_{\mathrm{mid}}) = 0$, we have $\theta^* = \theta_{\mathrm{mid}}$ and $\boldsymbol{\phi}^{\mathsf{QRR}}$ is the maximizer of $c(\theta^*)$. When $c(\theta_{\mathrm{mid}}) < 0$, update the feasible region of θ^* as $[\theta_{\min}, \theta_{\mathrm{mid}}]$. If $c(\theta_{\mathrm{mid}}) > 0$, update the feasible region of θ^* as $[\theta_{\mathrm{mid}}, \theta_{\max}]$. In either case, repeat the bisection process.

[5]The value $\theta_{\max}$ is created by noting that, in a positively correlated channel, the k-step transition probabilities $P_{n,01}^{(k)}$ and $P_{n,10}^{(k)}$ increase and decrease with k, respectively; both sequences have the same limit $\pi_{n,\mathsf{ON}}$.

B. Lyapunov drift inequality

The construction of the QRRNUM policy follows a new Lyapunov drift argument. We start with constructing a frame-based Lyapunov drift inequality over a frame of size T, where T is possibly random but has a finite second moment bounded by a constant C so that $C \ge \mathbb{E}\left[T^2 \mid \boldsymbol{Q}(t)\right]$ for all t and all possible $\boldsymbol{Q}(t)$. Intuition for constructing such an inequality is shown later. By iteratively applying (5), it is not hard to show

$$Q_n(t+T) \le \max\left[Q_n(t) - \sum_{\tau=0}^{T-1}\mu_n(t+\tau), 0\right] + \sum_{\tau=0}^{T-1} r_n(t+\tau) \tag{13}$$

for every $n \in \{1,\ldots,N\}$. We define the *quadratic Lyapunov function* $L(\boldsymbol{Q}(t)) \triangleq \frac{1}{2}\sum_{n=1}^{N} Q_n^2(t)$ as a scalar measure of the queue size vector $\boldsymbol{Q}(t)$. Define the *T-slot Lyapunov drift*

$$\Delta_T(\boldsymbol{Q}(t)) \triangleq \mathbb{E}\left[L(\boldsymbol{Q}(t+T)) - L(\boldsymbol{Q}(t)) \mid \boldsymbol{Q}(t)\right]$$

as a conditional expected change of queue sizes across T slots, where the expectation is with respect to the randomness of the network within the T slots, including the randomness of T. By taking square of (13) for every n, using inequalities

$$\max[a-b,0] \le a \quad \forall a, b \ge 0,$$
$$(\max[a-b,0])^2 \le (a-b)^2, \quad \mu_n(t) \le 1, \quad r_n(t) \le 1,$$

to simplify terms, summing all resulting inequalities, and taking conditional expectation on $\boldsymbol{Q}(t)$, we can show

$$\begin{aligned}\Delta_T(\boldsymbol{Q}(t)) &\le B \\ &- \mathbb{E}\left[\sum_{n=1}^{N} Q_n(t)\left[\sum_{\tau=0}^{T-1}\mu_n(t+\tau) - r_n(t+\tau)\right] \mid \boldsymbol{Q}(t)\right]\end{aligned} \tag{14}$$

where $B \triangleq NC > 0$ is a constant. Subtracting from both sides of (14) the weighted sum $V\mathbb{E}\left[\sum_{\tau=0}^{T-1} g(\boldsymbol{r}(t+\tau)) \mid \boldsymbol{Q}(t)\right]$, where $V > 0$ is a predefined control parameter and $\boldsymbol{r}(t+\tau)$ an admitted data vector, we get the Lyapunov drift inequality

$$\begin{aligned}&\Delta_T(\boldsymbol{Q}(t)) - V\mathbb{E}\left[\sum_{\tau=0}^{T-1} g(\boldsymbol{r}(t+\tau)) \mid \boldsymbol{Q}(t)\right] \\ &\quad \le B - f(\boldsymbol{Q}(t)) - h(\boldsymbol{Q}(t)),\end{aligned} \tag{15}$$

where

$$f(\boldsymbol{Q}(t)) \triangleq \sum_{n=1}^{N} Q_n(t)\mathbb{E}\left[\sum_{\tau=0}^{T-1}\mu_n(t+\tau) \mid \boldsymbol{Q}(t)\right] \tag{16}$$

$$\begin{aligned}h(\boldsymbol{Q}(t)) \triangleq \mathbb{E}\Bigg[\sum_{\tau=0}^{T-1}\Bigg[&V\, g(\boldsymbol{r}(t+\tau)) \\ &- \sum_{n=1}^{N} Q_n(t) r_n(t+\tau)\Bigg] \mid \boldsymbol{Q}(t)\Bigg].\end{aligned} \tag{17}$$

The inequality (15) holds for any scheduling policy over a frame of any size T.

C. Intuition behind the Lyapunov drift inequality

The desired network control policy shall stabilize all queues $(Q_1(t), \ldots, Q_N(t))$ and maximize the throughput utility $g(\cdot)$. For queue stability, we want to minimize the Lyapunov drift $\Delta_T(\boldsymbol{Q}(t))$, because it captures the expected growth of queue sizes over a duration of time. On the other hand, to increase throughput utility, we want to admit more data into the system for service, and maximize the expected sum utility $\mathbb{E}\left[\sum_{\tau=0}^{T-1} g(\boldsymbol{r}(t+\tau)) \mid \boldsymbol{Q}(t)\right]$. Minimizing Lyapunov drift and maximizing throughput utility, however, conflict with each other, because queue sizes increase with more data admitted into the system. To capture this tradeoff, it is natural to minimize a weighted difference of Lyapunov drift and throughput utility, which is the left side of (15). The tradeoff is controlled by the positive parameter V. Intuitively, a large V value puts more weights on throughput utility, thus throughput utility is improved, at the expense of the growth of the queue size reflected in $\Delta_T(\boldsymbol{Q}(t))$. The construction of the inequality (15) provides a useful bound on the weighted difference of Lyapunov drift and throughput utility.

The QRRNUM policy that we will construct in the next section uses the above ideas with two modifications. First, it suffices to minimize a bound on the weighted difference of Lyapunov drift and throughput utility, i.e., the right side of (15). Second, since the weighted difference of Lyapunov drift and throughput utility in (15) is made over a frame of T slots, where the value T is random and depends on the policy implemented within the frame, it is natural to normalize the weighted difference by the average frame size, and we will minimize the resulting ratio (see (18)). This new ratio rule is a generalization of the MaxWeight policies for stochastic network optimization over frame-based systems.

D. Construction of the QRRNUM policy

We consider the policy that, at the start of every round, observes the current queue backlog vector $\boldsymbol{Q}(t)$ and maximizes over all feasible policies the expression:

$$\frac{f(\boldsymbol{Q}(t)) + h(\boldsymbol{Q}(t))}{\mathbb{E}\left[T \mid \boldsymbol{Q}(t)\right]} \tag{18}$$

over a frame of size T, where the numerator is defined in (16) and (17). Every feasible policy here consists of: (1) an admission policy that admits packets into queues $\boldsymbol{Q}(t)$ for all users in every slot; (2) a randomized round robin policy RandRR (given in Section III-A) for data delivery. The frame size T in (18) is the length of one transmission round under the candidate RandRR policy, and its distribution depends on the backlog vector $\boldsymbol{Q}(t)$ via the queue-dependent choice of policy RandRR. When the feasible policy that maximizes (18) is chosen, it is executed for one round of transmission, after which a new policy is chosen for the next round based on the updated ratio of (18), and so on.

Next we simplify the maximization of (18); the result is the QRRNUM policy in Section IV-A. In $h(\boldsymbol{Q}(t))$, the optimal admitted data vector $\boldsymbol{r}(t+\tau)$ in every slot is independent of the frame size T and of the rate allocations $\mu_n(t+\tau)$ in $f(\boldsymbol{Q}(t))$. In addition, it should be the same for all $\tau \in \{0, \ldots, T-1\}$, and is the solution to (9)-(10). These observations lead to the admission control subroutine in the QRRNUM policy.

Let $\Psi^*(\boldsymbol{Q}(t))$ denote the optimal objective of (9)-(10). Since the optimal admitted data vector is independent of the frame size T, we get $h(\boldsymbol{Q}(t)) = \mathbb{E}\left[T \mid \boldsymbol{Q}(t)\right] \Psi^*(\boldsymbol{Q}(t))$, and (18) is equal to

$$\frac{f(\boldsymbol{Q}(t))}{\mathbb{E}\left[T \mid \boldsymbol{Q}(t)\right]} + \Psi^*(\boldsymbol{Q}(t)). \tag{19}$$

It indicates that finding the optimal admission policy is independent of finding the optimal RandRR policy that maximizes the first term of (19).

Next we evaluate the first term of (19) under a fixed RandRR policy with parameters $\{\alpha_{\boldsymbol{\phi}}\}_{\boldsymbol{\phi}\in\Phi\cup\{\mathbf{0}\}}$. In the rest of the section, when we use a $\mathsf{RR}(\boldsymbol{\phi})$ policy for one round, the channel ordering of least recently used first is always adopted. Conditioning on the choice of $\boldsymbol{\phi}$, we get

$$f(\boldsymbol{Q}(t)) = \sum_{\boldsymbol{\phi}\in\Phi\cup\{\mathbf{0}\}} \alpha_{\boldsymbol{\phi}}\, f(\boldsymbol{Q}(t), \mathsf{RR}(\boldsymbol{\phi})),$$

where $f(\boldsymbol{Q}(t), \mathsf{RR}(\boldsymbol{\phi}))$ denotes the term $f(\boldsymbol{Q}(t))$ in (16) evaluated under the policy $\mathsf{RR}(\boldsymbol{\phi})$ for one round; for convenience we have denoted by $\mathsf{RR}(\mathbf{0})$ the policy of idling the system for one slot. Similarly, by conditioning we can show [6]

$$\mathbb{E}\left[T\right] = \mathbb{E}\left[T \mid \boldsymbol{Q}(t)\right] = \sum_{\boldsymbol{\phi}\in\Phi\cup\{\mathbf{0}\}} \alpha_{\boldsymbol{\phi}}\, \mathbb{E}\left[T_{\mathsf{RR}(\boldsymbol{\phi})}\right],$$

where $T_{\mathsf{RR}(\boldsymbol{\phi})}$ denotes the duration of one transmission round under the $\mathsf{RR}(\boldsymbol{\phi})$ policy. It follows that

$$\frac{f(\boldsymbol{Q}(t))}{\mathbb{E}\left[T \mid \boldsymbol{Q}(t)\right]} = \frac{\sum_{\boldsymbol{\phi}\in\Phi\cup\{\mathbf{0}\}} \alpha_{\boldsymbol{\phi}}\, f(\boldsymbol{Q}(t), \mathsf{RR}(\boldsymbol{\phi}))}{\sum_{\boldsymbol{\phi}\in\Phi\cup\{\mathbf{0}\}} \alpha_{\boldsymbol{\phi}}\, \mathbb{E}\left[T_{\mathsf{RR}(\boldsymbol{\phi})}\right]}. \tag{20}$$

The next lemma shows that there always exists a $\mathsf{RR}(\boldsymbol{\phi})$ policy maximizing (20) over all RandRR policies for one round of transmission. Therefore it suffices to focus on the class of $\mathsf{RR}(\boldsymbol{\phi})$ policies in every round of transmission.

Lemma 2. *We index $\mathsf{RR}(\boldsymbol{\phi})$ policies for all $\boldsymbol{\phi} \in \Phi\cup\{\mathbf{0}\}$. For the $\mathsf{RR}(\boldsymbol{\phi})$ policy with index k, define*

$$f_k \triangleq f(\boldsymbol{Q}(t), \mathsf{RR}(\boldsymbol{\phi})), \quad D_k \triangleq \mathbb{E}\left[T_{\mathsf{RR}(\boldsymbol{\phi})}\right].$$

Without loss of generality, assume

$$\frac{f_1}{D_1} \geq \frac{f_k}{D_k}, \quad \forall k \in \{2, 3, \ldots, 2^N\}.$$

Then for any probability distribution $\{\alpha_k\}_{k\in\{1,\ldots,2^N\}}$ with $\alpha_k \geq 0$ and $\sum_{k=1}^{2^N} \alpha_k = 1$, we have

$$\frac{f_1}{D_1} \geq \frac{\sum_{k=1}^{2^N} \alpha_k\, f_k}{\sum_{k=1}^{2^N} \alpha_k\, D_k}.$$

Proof of Lemma 2: Omitted due to space constraint. ■

By Lemma 2, next we evaluate the first term of (19) under a given $\mathsf{RR}(\boldsymbol{\phi})$ policy. When $\boldsymbol{\phi} = \mathbf{0}$, we get $f(\boldsymbol{Q}(t))/\mathbb{E}\left[T \mid \boldsymbol{Q}(t)\right] = 0$. Otherwise, fix some $\boldsymbol{\phi} \in \Phi$. For each active channel n in $\boldsymbol{\phi}$, we denote by $L_n^{\boldsymbol{\phi}}$ the amount

[6] Given a fixed RandRR policy, the frame size T no longer depends on the backlog vector $\boldsymbol{Q}(t)$, and $\mathbb{E}\left[T\right] = \mathbb{E}\left[T \mid \boldsymbol{Q}(t)\right]$.

of time the network stays with channel n in one round of transmission under policy RR($\boldsymbol{\phi}$). The probability distribution and the mean of $L_n^{\boldsymbol{\phi}}$ are given in Theorem 1. It follows that under the RR($\boldsymbol{\phi}$) policy we have

$$\mathbb{E}[T] = \mathbb{E}[T \mid \boldsymbol{Q}(t)] = \sum_{n:\phi_n=1} \mathbb{E}\left[L_n^{\boldsymbol{\phi}}\right],$$

$$\mathbb{E}\left[\sum_{\tau=0}^{T-1} \mu_n(t+\tau) \mid \boldsymbol{Q}(t)\right] = \begin{cases} \mathbb{E}\left[L_n^{\boldsymbol{\phi}}\right] - 1 & \text{if } \phi_n = 1 \\ 0 & \text{if } \phi_n = 0 \end{cases}.$$

As a result,

$$\frac{f(\boldsymbol{Q}(t))}{\mathbb{E}[T \mid \boldsymbol{Q}(t)]} = \frac{\sum_{n=1}^N Q_n(t)\mathbb{E}\left[L_n^{\boldsymbol{\phi}} - 1\right]\phi_n}{\sum_{n=1}^N \mathbb{E}\left[L_n^{\boldsymbol{\phi}}\right]\phi_n}. \qquad (21)$$

The above simplifications lead to the channel scheduling subroutine of the QRRNUM policy.

V. Performance Analysis

Theorem 2. *Let $y_n(t) = \min[Q_n(t), \mu_n(t)]$ be the number of packets delivered to user n in slot t; define $\boldsymbol{y}(t) \triangleq (y_n(t))_{n=1}^N$. For any control parameter $V > 0$, the* QRRNUM *policy stabilizes all queues $(Q_1(t), \ldots, Q_N(t))$ and yields throughput utility satisfying*

$$\liminf_{t\to\infty} g\left(\frac{1}{t}\sum_{\tau=0}^{t-1} \mathbb{E}[\boldsymbol{y}(\tau)]\right) \geq g(\overline{\boldsymbol{y}}^*) - \frac{B}{V}, \qquad (22)$$

where $g(\overline{\boldsymbol{y}}^*)$ is the optimal objective of the constrained problem (3) *and $B > 0$ is a finite constant.*

Proof of Theorem 2: In Appendix A. ∎

Theorem 2 shows that the throughput utility under the QRRNUM policy is at most B/V away from the optimal $g(\overline{\boldsymbol{y}}^*)$. By choosing V sufficiently large, the throughput utility can be made arbitrarily close to the optimal $g(\overline{\boldsymbol{y}}^*)$ and the constrained problem (3) is solved. The tradeoff of choosing a large V value is that the average queue size in the network grows linearly with V. Such tradeoff agrees with the design principle of the QRRNUM policy discussed in Section IV-C.

VI. Conclusion

This paper provides a theoretical framework for utility maximization over a wireless network with partially observable Markov ON/OFF channels. The performance and control in this network are constrained by limiting channel probing and delayed/uncertain channel state information, but can be improved by exploiting channel memory. Overall, attacking such problems requires solving (at least approximately) high-dimensional restless multi-armed bandit problems with non-linear objective functions of time average rewards, which are difficult to solve by existing tools such as Whittle's index or Markov decision theory. This paper provides a new achievable region method for these problems. The idea is to first identify a good inner performance region rendered by randomizing stationary policies, and then solve the problem over the inner region, serving as an approximation to the original problem. In this paper, with an inner performance region constructed in [8], we provide a novel frame-based Lyapunov drift argument that solves the approximation problem with provably near-optimal performance. We generalize the classical MaxWeight policies from time-slotted wireless networks to frame-based ones that have policy-dependent random frame sizes. Extensions of this new achievable region method to other open stochastic optimization problems are interesting future research.

References

[1] M. J. Neely, E. Modiano, and C.-P. Li, "Fairness and optimal stochastic control for heterogeneous networks," *IEEE/ACM Trans. Netw.*, vol. 16, no. 2, pp. 396–409, Apr. 2008.

[2] F. P. Kelly, "Charging and rate control for elastic traffic," *European Trans. Telecommunications*, vol. 8, pp. 33–37, 1997. [Online]. Available: http://www.statslab.cam.ac.uk/∼frank/elastic.html

[3] Q. Zhao and B. M. Sadler, "A survey of dynamic spectrum access," *IEEE Signal Process. Mag.*, vol. 24, no. 3, pp. 79–89, May 2007.

[4] Q. Zhao and A. Swami, "A decision-theoretic framework for opportunistic spectrum access," *IEEE Wireless Commun. Mag.*, vol. 14, no. 4, pp. 14–20, Aug. 2007.

[5] J. L. Ny, M. Dahleh, and E. Feron, "Multi-uav dynamic routing with partial observations using restless bandit allocation indices," in *American Control Conference*, Seattle, WA, USA, Jun. 2008.

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

[7] C. H. Papadimitriou and J. N. Tsitsiklis, "The complexity of optimal queueing network control," *Math. of Oper. Res.*, vol. 24, pp. 293–305, May 1999.

[8] C.-P. Li and M. J. Neely, "Exploiting channel memory for multiuser wireless scheduling without channel measurement: Capacity regions and algorithms," *Performance Evaluation*, 2011, accepted for publication.

[9] L. Tassiulas and A. Ephremides, "Stability properties of constrained queueing systems and scheduling policies for maximum throughput in multihop radio networks," *IEEE Trans. Autom. Control*, vol. 37, no. 12, pp. 1936–1948, Dec. 1992.

[10] ——, "Dynamic server allocation to parallel queues with randomly varying connectivity," *IEEE Trans. Inf. Theory*, vol. 39, no. 2, pp. 466–478, Mar. 1993.

[11] M. J. Neely, "Energy optimal control for time varying wireless networks," *IEEE Trans. Inf. Theory*, vol. 52, no. 7, pp. 2915–2934, Jul. 2006.

[12] ——, "Dynamic power allocation and routing for satellite and wireless networks with time varying channels," Ph.D. dissertation, Massachusetts Institute of Technology, November 2003.

[13] A. Eryilmaz and R. Srikant, "Fair resource allocation in wireless networks using queue-length-based scheduling and congestion control," *IEEE/ACM Trans. Netw.*, vol. 15, no. 6, pp. 1333–1344, Dec. 2007.

[14] M. J. Neely, "Stochastic optimization for markov modulated networks with application to delay constrained wireless scheduling," in *IEEE Conf. Decision and Control (CDC)*, 2009.

[15] ——, "Dynamic optimization and learning for renewal systems," in *Asilomar Conf. Signals, Systems, and Computers*, Nov. 2010, invited paper.

[16] ——, *Stochastic Network Optimization with Application to Communication and Queueing Systems*. Morgan & Claypool, 2010.

[17] C.-P. Li and M. J. Neely, "Energy-optimal scheduling with dynamic channel acquisition in wireless downlinks," *IEEE Trans. Mobile Comput.*, vol. 9, no. 4, pp. 527 –539, Apr. 2010.

[18] P. Chaporkar, A. Proutiere, H. Asnani, and A. Karandikar, "Scheduling with limited information in wireless systems," in *ACM Int. Symp. Mobile Ad Hoc Networking and Computing (MobiHoc)*, New Orleans, LA, May 2009.

[19] N. B. Chang and M. Liu, "Optimal channel probing and transmission scheduling for opportunistic spectrum access," in *ACM Int. Conf. Mobile Computing and Networking (MobiCom)*, New York, NY, 2007, pp. 27–38.

[20] P. Chaporkar, A. Proutiere, and H. Asnani, "Learning to optimally exploit multi-channel diversity in wireless systems," in *IEEE Proc. INFOCOM*, 2010.

[21] P. Chaporkar and A. Proutiere, "Optimal joint probing and transmission strategy for maximizing throughput in wireless systems," *IEEE J. Sel. Areas Commun.*, vol. 26, no. 8, pp. 1546–1555, Oct. 2008.

[22] S. Guha, K. Munagala, and S. Sarkar, "Jointly optimal transmission and probing strategies for multichannel wireless systems," in *Conf. Information Sciences and Systems*, Mar. 2006.

[23] Q. Zhao, B. Krishnamachari, and K. Liu, "On myopic sensing for multi-channel opportunistic access: Structure, optimality, and preformance," *IEEE Trans. Wireless Commun.*, vol. 7, no. 12, pp. 5431–5440, Dec. 2008.
[24] S. H. A. Ahmad, M. Liu, T. Javidi, Q. Zhao, and B. Krishnamachari, "Optimality of myopic sensing in multichannel opportunistic access," *IEEE Trans. Inf. Theory*, vol. 55, no. 9, pp. 4040–4050, Sep. 2009.
[25] S. H. A. Ahmad and M. Liu, "Multi-channel opportunistic access: A case of restless bandits with multiple plays," in *Allerton Conf. Communication, Control, and Computing*, 2009, pp. 1361–1368.
[26] K. Liu and Q. Zhao, "Indexability of restless bandit problems and optimality of whittle's index for dynamic multichannel access," *IEEE Trans. Inf. Theory*, vol. 56, no. 11, pp. 5547–5567, Nov. 2010.
[27] J. Nino-Mora, "An index policy for dynamic fading-channel allocation to heterogeneous mobile users with partial observations," in *Next Generation Internet Networks*, 2008, pp. 231–238.
[28] S. Guha, K. Munagala, and P. Shi, "Approximation algorithms for restless bandit problems," Tech. Rep., Feb. 2009.
[29] D. P. Bertsekas, *Dynamic Programming and Optimal Control*, 3rd ed. Athena Scientific, 2005, vol. I.

APPENDIX A

Proof of Theorem 2: We need to show that all queues $(Q_n(t))_{n=1}^N$ are stable and that (22) is achieved. Due to space constraint, we only prove (22) here. Under policy QRRNUM, let t_{k-1} and T_k be the beginning and the duration of the kth transmission round, respectively. We have $T_k = t_k - t_{k-1}$ and $t_k = \sum_{i=1}^k T_i$ for all $k \in \mathbb{N}$. Every T_k is the length of a transmission round of some $\mathsf{RR}(\boldsymbol{\phi})$ policy. Assume $t_0 = 0$.

To show (22), we compare the QRRNUM policy with the (unknown) solution to problem (3). By definition of the throughput region Λ_{int} in Section III-B, there exists a randomized round robin policy RandRR^* that solves (3) and yields the optimal throughput vector $\overline{\boldsymbol{y}}^* = (\overline{y}_n^*)_{n=1}^N$. Let T^* denote the length of one transmission round under policy RandRR^*. From Corollary 1, we have for every user $n \in \{1, \ldots, N\}$:

$$\mathbb{E}\left[\sum_{\tau=0}^{T^*-1} \mu_n(t+\tau) \mid \boldsymbol{Q}(t)\right] = \mathbb{E}\left[\sum_{\tau=0}^{T^*-1} \mu_n(t+\tau)\right] = \overline{y}_n^* \mathbb{E}\left[T^*\right].$$

Combining RandRR^* with the admission policy σ^* that sets $r_n(t+\tau) = \overline{y}_n^*$ for all users n and $\tau \in \{0, \ldots, T^*-1\}$ yields [7]

$$f^*(\boldsymbol{Q}(t)) = \mathbb{E}\left[T^*\right] \sum_{n=1}^N Q_n(t)\, \overline{y}_n^* \tag{23}$$

$$h^*(\boldsymbol{Q}(t)) = \mathbb{E}\left[T^*\right] \left[V\, g(\overline{\boldsymbol{y}}^*) - \sum_{n=1}^N Q_n(t)\, \overline{y}_n^*\right] \tag{24}$$

where (23) and (24) are $f(\boldsymbol{Q}(t))$ and $h(\boldsymbol{Q}(t))$ (see (16), (17)) evaluated under policies RandRR^* and σ^*, respectively.

Since the QRRNUM policy maximizes (18), comparing (18) under QRRNUM and the joint policy $(\mathsf{RandRR}^*, \sigma^*)$ yields

$$\begin{aligned} &f_{\mathsf{QRRNUM}}(\boldsymbol{Q}(t_k)) + h_{\mathsf{QRRNUM}}(\boldsymbol{Q}(t_k)) \\ &\geq \mathbb{E}\left[T_{k+1} \mid \boldsymbol{Q}(t_k)\right] \frac{f^*(\boldsymbol{Q}(t_k)) + h^*(\boldsymbol{Q}(t_k))}{\mathbb{E}\left[T^*\right]} \\ &\overset{(a)}{=} \mathbb{E}\left[T_{k+1} \mid \boldsymbol{Q}(t_k)\right] V\, g(\overline{\boldsymbol{y}}^*), \end{aligned} \tag{25}$$

where (a) is from (23) and (24). The inequality (15) under the QRRNUM policy in the $(k+1)$th round of transmission yields

$$\begin{aligned} &\Delta_{T_{k+1}}(\boldsymbol{Q}(t_k)) - V\mathbb{E}\left[\sum_{\tau=0}^{T_{k+1}-1} g(\boldsymbol{r}(t_k+\tau)) \mid \boldsymbol{Q}(t_k)\right] \\ &\leq B - f_{\mathsf{QRRNUM}}(\boldsymbol{Q}(t_k)) - h_{\mathsf{QRRNUM}}(\boldsymbol{Q}(t_k)) \\ &\overset{(b)}{\leq} B - \mathbb{E}\left[T_{k+1} \mid \boldsymbol{Q}(t_k)\right] V\, g(\overline{\boldsymbol{y}}^*), \end{aligned} \tag{26}$$

where (b) uses (25). Taking expectation over $\boldsymbol{Q}(t_k)$ in (26) and summing it over $k \in \{0, \ldots, K-1\}$, we get

$$\begin{aligned} &\mathbb{E}\left[L(\boldsymbol{Q}(t_K))\right] - \mathbb{E}\left[L(\boldsymbol{Q}(t_0))\right] - V\mathbb{E}\left[\sum_{\tau=0}^{t_K-1} g(\boldsymbol{r}(\tau))\right] \\ &\leq BK - Vg(\overline{\boldsymbol{y}}^*)\, \mathbb{E}\left[t_K\right] \leq \left[B - Vg(\overline{\boldsymbol{y}}^*)\right] \mathbb{E}\left[t_K\right] \end{aligned} \tag{27}$$

where the last inequality uses $t_K = \sum_{k=1}^K T_k \geq K$. Ignoring the first term, noting $\boldsymbol{Q}(t_0) = \mathbf{0}$, and dividing by V yields

$$\mathbb{E}\left[\sum_{\tau=0}^{t_K-1} g(\boldsymbol{r}(\tau))\right] \geq \left(g(\overline{\boldsymbol{y}}^*) - \frac{B}{V}\right) \mathbb{E}\left[t_K\right]. \tag{28}$$

Let $K(t)$ denote the number of transmission rounds ending by time t; we have $t_{K(t)} \leq t < t_{K(t)+1}$. It follows that

$$\begin{aligned} &\sum_{\tau=0}^{t-1} \mathbb{E}\left[g(\boldsymbol{r}(\tau))\right] \overset{(c)}{\geq} \mathbb{E}\left[\sum_{\tau=0}^{t_{K(t)}-1} g(\boldsymbol{r}(\tau))\right] \\ &\overset{(d)}{\geq} \left(g(\overline{\boldsymbol{y}}^*) - \frac{B}{V}\right) \mathbb{E}\left[t_{K(t)}\right] \\ &= \left[g(\overline{\boldsymbol{y}}^*) - \frac{B}{V}\right] t - \left[g(\overline{\boldsymbol{y}}^*) - \frac{B}{V}\right] \left(t - \mathbb{E}\left[t_{K(t)}\right]\right), \end{aligned} \tag{29}$$

where (c) uses the nonnegativity of $g(\cdot)$ and $t \geq t_{K(t)}$, and (d) follows (28). Dividing (29) by t, taking a $\liminf$ as $t \to \infty$, and noting $t - \mathbb{E}\left[t_{K(t)}\right] \leq \mathbb{E}\left[T_{K(t)+1}\right] < \infty$ yields

$$\liminf_{t\to\infty} \frac{1}{t} \sum_{\tau=0}^{t-1} \mathbb{E}\left[g(\boldsymbol{r}(\tau))\right] \geq g(\overline{\boldsymbol{y}}^*) - \frac{B}{V}. \tag{30}$$

Using Jensen's inequality and the concavity of $g(\cdot)$, we get

$$\liminf_{t\to\infty} g\left(\overline{\boldsymbol{r}}^{(t)}\right) \geq g(\overline{\boldsymbol{y}}^*) - \frac{B}{V}, \tag{31}$$

where $\overline{\boldsymbol{r}}^{(t)} \triangleq (\overline{r}_n^{(t)})_{n=1}^N$ and $\overline{r}_n^{(t)} \triangleq \frac{1}{t}\sum_{\tau=0}^{t-1} \mathbb{E}\left[r_n(\tau)\right]$. Since all queues $(Q_n(t))_{n=1}^N$ are stable, we can show that the time average throughput vector $\overline{\boldsymbol{y}}^{(t)} = (\overline{y}_n^{(t)})_{n=1}^N$, where $\overline{y}_n^{(t)} \triangleq \frac{1}{t}\sum_{\tau=0}^{t-1} \mathbb{E}\left[y_n(\tau)\right]$, satisfies

$$\liminf_{t\to\infty} g\left(\overline{\boldsymbol{y}}^{(t)}\right) \geq \liminf_{t\to\infty} g\left(\overline{\boldsymbol{r}}^{(t)}\right). \tag{32}$$

Combining (31) and (32) finishes the proof. ■

[7] The throughput $\overline{y}_n^*$ is less than or equal to one, thus is a feasible choice of $r_n(t+\tau)$.