Sliding Mode Based Congestion Control and Scheduling for Multi-class Traffic over Per-link Queueing Wireless Networks

Zongrui Ding and Dapeng Wu

Abstract

This paper considers the problem of joint congestion control and scheduling in wireless networks with quality of service (QoS) guarantees. Different from per-destination queueing in the existing works, which is not scalable, this work considers per-link queueing at each node, which significantly reduces the number of queues per node. Under per-link queueing, we formulate the joint congestion control and scheduling problem as a network utility maximization (NUM) problem and use a dual decomposition method to separate the NUM problem into two sub-problems, i.e., a congestion control problem and a scheduling problem. Then, we develop a sliding mode (SM) based distributed congestion control scheme, and prove its convergence and optimality property. Different from the existing schemes, our congestion control scheme is capable of providing multi-class QoS under the general scenario of multi-path and multi-hop; in addition, it is robust against network anomalies, e.g., link failures, because it can achieve multi-path load balancing.

Index Terms

Sliding mode control, per-link queue, QoS, multi-path, stochastic optimization, interactive multimedia, robustness against network anomalies.

I. Introduction

The authors are with Department of Electrical and Computer Engineering, University of Florida, Gainesville, FL 32601. Email of Zongrui Ding: dingzr@ufl.edu. Correspondence author: Prof. Dapeng Wu, wu@ece.ufl.edu, http://www.wu.ece.ufl.edu. This work was supported in part by National Science Foundation under grant CNS-0643731, CNS-1116970, Office of Naval Research under grant N000140810873, and the Joint Research Fund for Overseas Chinese Young Scholars under grant No. 61228101.

FUTURE wireless networks are expected to consist of a variety of heterogenous components and aim to support interactive applications such as voice over IP (VoIP), online multi-player games and live sharing of multimedia contents in terms of video, audio, texts or images among distributed users. In such a context, a network should be designed to provide quality of service (QoS) to inelastic traffic such as video and audio, and simultaneously provide high data rate for elastic traffic, e.g, email and web traffic. Therefore, how to perform congestion control and scheduling efficiently is crucial to achieve these key features for a wireless network.

Different from wire-line networks, the joint congestion control and scheduling in wireless networks is particularly challenging due to unreliability, time-varying channel gain, and interference among wireless channels. The joint congestion control and scheduling can be formulated as a network utility maximization (NUM) problem [1] [2] [3] [4] [5] [6] [7]. Under the setting that each node has no buffer, the NUM problem can be solved in a distributed manner by iteratively updating the link price, which is the sum of the per-hop prices [1] [5]. This mechanism requires that each link calculate and feed back the per-hop prices to the source, thus inducing a lot of overhead. In per-destination queueing networks, the optimal solution for the NUM problem can be obtained by the primal-dual algorithm within the network capacity region [8] [9], which requires each node maintains a queue for each flow. Works such as [6]and [7] optimize the end-to-end delay performance by the NUM problem.

There are two commonly used queueing structures: the per-destination queue [2] [3] [9] [10] and the per-link queue [3] [11] [12] [6] [7]. In the per-destination queueing networks, every node needs to maintain a queue [13] for each flow, thus the number of queues per node can be as large as the number of nodes in the network. This overhead is unbearable for relay nodes in a large-scale wireless network. It is shown in [11] that the per-link queue structure involves less overhead, preserves some good design features such as decomposition [14] and naturally provides multi-path awareness. Therefore, we consider the joint congestion control and scheduling problem under the setting of per-link queueing.

In wireless networks, distributed algorithms are desirable due to the unavailability of global information. Generally, the QoS constraints make it difficult to solve the congestion control problem in a distributed manner. Directly projecting the non-QoS-constrained solution to the QoS-constrained region usually results in a degraded performance. Sliding mode (SM) control theory [15] provides a powerful tool for designing distributed algorithms for solving convex optimization problems with a large number of constraints, and has been applied to congestion control problems in the Internet [16] [17] [18]. However, the SM control is not directly applicable to wireless networks because of time-varying channels. Fortunately, the decomposition between the scheduling problem and the congestion control problem may make it possible to apply the SM control to the congestion control problem which does not involve the time-varying channel states, given the Lagrangian multipliers resulted from the decomposition.

Different from the existing works, we consider the congestion control and scheduling of a per-link queueing wireless network with multi-class QoS requirements under both constant channels and time-varying channels. By "per-link-queue (per-next-hop queue)", we mean that packets destined to the same next hop node, are put into the same queue. Thus the number of queues that a node has to maintain is the number of its neighbors within one hop. Our system model shares some similarities with the route-dependent case in [3], where time-varying channels and QoS are not considered. Some works study similar problems under the per-destination queueing models, such as [19] [20] [21]. In [22], the authors consider the general framework of the time-average penalty or reward optimization, but no distributed solution is given. Different from the SM controller in [16] [17] designed for the Internet, the wireless context induces an interaction between congestion control on Layer 4 and scheduling on Layer 2, so that the convergence conditions are naturally satisfied and no network feedback is necessary, which is a good example of "layering as optimization decomposition" [14] [23]. Under the per-link queueing model, we apply the SM control technique to the congestion control problem and develop a distributed scheme. We prove that 1) the distributed control scheme converges and achieves optimality under continuous-time system parameters; and 2) it converges to a bounded neighborhood of the optimal solution under discrete time and time-varying channels.

In this paper, we first consider the congestion control and scheduling under constant wireless channels for simplicity, and then extend the results to time-varying wireless channels and evaluate the performance by the Lyapunov drift method [22] [24].

The rest of the paper is organized as follows. In Section II, we describe the system model, followed by the formulation and solution to the NUM problem under constant channels in Section III. Section IV extends the solution in Section III to wireless networks under time-varying channels. Section V presents performance analysis and Section VI presents simulation results. Section VII concludes the paper.

II. System Model

In this section, we present the model of a multi-hop multi-path per-link queueing wireless network. Section II-A describes the per-link queueing model. Section II-B presents the QoS requirements for heterogenous services and Section II-C describes the capacity region of a per-link queueing network.

A. The Queueing Model

For a per-link queueing wireless network, let $\mathcal{F}$ denote the set of flows and $\mathcal{N}$ denote the set of nodes. A link $l_{i,j}$ in the network has a transmitting node $i \in \mathcal{N}$ and a receiving node $j \in \mathcal{N}_i$, where $\mathcal{N}_i$ is the set of node i's next hop nodes. $\mathcal{L} = \{l_{i,j}\}$ denotes the set of all links. These links are directed; we assume connectivity between any two nodes is symmetric and the topology is assumed to be static. Each flow $s \in \mathcal{F}$ has a total transmission rate x^s over multiple paths. Let $x^s_{i,j}$ denote the transmission rate of flow s over $l_{i,j}$ with source node $b^s = i$. Then the total transmission rate x^s over multi-path is

$$x^s = \sum_{i=b^s} \sum_{j \in \ \mathcal{N}_i} x^s_{i,j}. \tag{1}$$

The rate power function denoted by $\mu(G, P)$ is the offered physical-layer capacity matrix under channel state matrix $G = [g_{i,j}]$ and power allocation matrix $P = [p_{i,j}] \in \mathcal{P}$, where $\mathcal{P}$ is the set of feasible power allocation matrix P. Matrix $X = [x^s_{i,j}]$ denotes the multi-path transmission rates for all flows. Throughout this paper, $A = [a_{i,j}]$ is a tensor notation for matrix A. A wireless channel is a shared medium, where links interfered with each other contend for exclusive access to the channel. A conflict graph [20] can capture the contention among links; the conflict graph depends on G and P. The convex hull of the corresponding rate vectors of the independent sets of the conflict graph determines the feasible rate region at link layer. It is assumed that a proper routing algorithm is applied to the network according to information such as distance or signal-to-interference-plus-noise-ratio (SINR), which results in a routing table denoted by $\mathcal{R}$. The source-destination pairs on link layer are considered as flows. In scheduling and routing, we consider the packet-level operation and the overhead of packetization is assumed to be zero.

Since we use the per-link queueing model, each next hop of $l_{i,j}$ has a dedicated queue (at node i) denoted by $q_{i,j}$. Therefore, the number of the queues in node i is at most the same as the number of its next hops. If $l_{i,j}$ is activated, the data in $q_{i,j}$ is transmitted. At time t, the output of $q_{i,j}$ consists of the traffic destined to node j and the transit traffic destined to nodes other than node j. Thus we have

$$\mu_{i,j}(t) = \sum_k \underline{\mu}_{i,j,k}(t) + \alpha_{i,j}(t) \tag{2}$$

where $\alpha_{i,j}(t)$ is the data rate delivered to node j through $l_{i,j}$; and $\mu_{i,j}(t)$ is a shorthand notation for $\mu_{i,j}\left(G(t), P(t)\right)$, i.e., the physical-layer capacity offered on $l_{i,j}$; and $\underline{\mu}_{i,j,k}(t)$ is the transmission rate from $q_{i,j}$ to $q_{j,k}$ through $l_{i,j}$ at time t. Here, a triplet node index (i, j, k) is used to define the two cascading queues $q_{i,j}$ and $q_{j,k}$ involved in a per-link transmission.

B. QoS Requirements

To describe the QoS requirements of heterogenous applications, define the QoS constraint functions for flow s by:

$$h^1(x^s) = \theta^s_{min} - x^s \tag{3a}$$

$$h^2(x^s) = x^s - \theta^s_{max} \tag{3b}$$

where θ^s_{min}, and θ^s_{max} are constants. The QoS requirements can be described by using inequality constraints on the functions in (3). For example, the service that guarantees minimum transmission rate for flow s can be defined as $h^1(x^s) \leq 0$. For notation simplicity, we use $h^1(x^s) \leq 0$ and $h^2(x^s) \leq 0$ to represent equality constraint, where $\theta^s_{min} = \theta^s_{max}$.

C. Capacity Region of a Per-link Queueing Network

In queueing networks, the joint congestion and scheduling problem should be solved under the constraint of network stability, which is defined as the stability of the queues. Let the exogenous arrival at $q_{i,j}$ be defined by

$$x_{i,j} = \sum_{s:b^s=i} x^s_{i,j}. \tag{4}$$

Let $F^i_{i,j}(t)$ and $F^o_{i,j}(t)$ denote the steady state inflow and outflow rate of $q_{i,j}$ at time t. Intuitively, we have

$$\begin{aligned} F^i_{i,j}(t) &= \frac{\sum_{\tau=0}^{t-1} \sum_{m\in\mathcal{N}_i} \underline{\mu}_{m,i,j}(\tau)}{t} \\ F^o_{i,j}(t) &= \frac{\sum_{\tau=0}^{t-1} \sum_{k\in\mathcal{N}_j} \mu_{i,j,k}(\tau)}{t}. \end{aligned} \tag{5}$$

Then the average inflow and outflow rate, denoted by $\bar{F}^i_{i,j}$ and $\bar{F}^o_{i,j}$ can be written as

$$\bar{F}^i_{i,j} = \lim_{t\to\infty} F^i_{i,j}(t)$$

$$\bar{F}^o_{i,j} = \lim_{t\to\infty} F^o_{i,j}(t)$$

Define $\mathcal{X}$ as the capacity region of a per-link queueing network under time-varying channels, then the necessary and sufficient conditions are derived as

$$x_{i,j} \leq \bar{F}^o_{i,j} - \bar{F}^i_{i,j}, \forall (i,j) \in \mathcal{L} \tag{6}$$

under the assumptions of convergent and bounded arrivals and convergent time-varying channel states similar to [**?**]. Note that a network with constant arrivals and constant wireless channels is a special case for this assumption. Usually, (6) is also known as the flow conservation constraint for queueing networks [2] [3], which is recognized as the network stability condition [9].

Note that the capacity region of a network under per-link queueing may be different from that under per-destination queueing. Intuitively, the per-destination queueing yields smaller granularity in scheduling than per-link networks. In [25], we studied the capacity region of the per-link queueing network with per-destination backlog information. Please refer to [9] [25] for details about the derivation and dynamic control in queueing networks.

III. The NUM Problem under Constant Channels

In this section, the joint congestion control and scheduling under constant wireless channels is formulated as an NUM problem. By dual decomposition, the NUM problem is decomposed into a congestion control problem and a scheduling problem.

We consider an NUM problem with QoS constraints as below

$$\max_{\{X\in\mathcal{X}\}} \sum_{s\in\mathcal{F}} U^s(x^s) \tag{7a}$$

$$s.t.: \quad \mathbf{h} \leq 0 \tag{7b}$$

$$x_{i,j} - \bar{F}^o_{i,j} + \bar{F}^i_{i,j} \leq 0, \forall (i,j) \in \mathcal{L} \tag{7c}$$

$$\sum_{i=b^s} \sum_{j\in\ \mathcal{N}_i} x^s_{i,j} = x^s \tag{7d}$$

where $U^s(x^s)$ is the utility of flow s with transmission rate x^s $(s \in S)$; (7c) is the flow conservation constraint; (7b) specifies the QoS constraints and vector $\mathbf{h} = [h_1(x^{s_1}), h_2(x^{s_2}), \ldots, h_\Omega(x^{s_\Omega})]^T$, where $h_\omega(x^{s_\omega})$ $(s_\omega \in S, \omega \in \{1, 2, \ldots, \Omega\})$ can be any function defined in (3); and $H^s = \{\omega | h_\omega(x^s)$ is an element of $\mathbf{h}\}$ is the set of the indices of the constraint functions for flow s $(s \in S)$. For example, in a network where only Flow 1 has a transmission rate requirement $10 \leq x^1 \leq 20$, then $\mathbf{h} = [x^1 - 20, 10 - x^1]^T$ and $H^1 = \{1, 2\}, H^s = \emptyset$ $(s \neq 1)$. Note that $\bar{F}^o_{i,j}$, $\bar{F}^i_{i,j}$ in (7) are the steady state outflow and inflow for link $l_{i,j}$, which are constants.

The solution to the NUM problem is organized as below: in Section III-A, (7) is decomposed into the congestion control problem (9) and the scheduling problem (10), given the Lagrangian multiplier Λ, which is updated iteratively according to Section III-B. Then (9) and (10) are solved in Section III-C and Section III-D, respectively. The optimality and convergence of the continuous time control laws are stated in Theorem 1 in Section III-E.

A. Decomposition

To solve problem (7), we use the dual decomposition method [3]. The Lagrangian dual function of (7) with respect to (w.r.t.) constraint (7c) is $D(\Lambda)$:

$$
\begin{aligned}
D(\Lambda) = \max_{\{X\in\mathcal{X}\}} \sum_{s\in S} U^s(x^s) - \\
\sum_{(i,j)\in\mathcal{L}} \lambda_{i,j}\left(x_{i,j} - \bar{F}^o_{i,j} + \bar{F}^i_{i,j}\right) \\
= \max_{\{X\in\mathcal{X}\}} \left(\sum_{s\in S} U^s(x^s) - \sum_{(i,j)\in\mathcal{L}} \lambda_{i,j} x_{i,j}\right) + \\
\max \sum_{(i,j)\in\mathcal{L}} \lambda_{i,j}\left(\bar{F}^o_{i,j} - \bar{F}^i_{i,j}\right),
\end{aligned} \tag{8}
$$

where $\Lambda = [\lambda_{i,j}]$ is the Lagrangian price matrix for constraint (7c) with $\lambda_{i,j}$ corresponding to each node pair $(i,j) \in \mathcal{L}$. It is implied by (8) that given the optimal link price $[\lambda_{i,j}]$, the first and the second term in (8) can be solved separately. Define the congestion control problem as

$$
\max_{\{X\in\mathcal{X},\mathbf{h}\leq 0\}} \left(\sum_{s\in\mathcal{F}} U^s(x^s) - \sum_{(i,j)\in\mathcal{L}} \lambda_{i,j} x_{i,j}\right), \tag{9}
$$

where (8) only depends on X given Λ, and the scheduling problem is given by

$$
\max \sum_{(i,j)\in\mathcal{L}} \lambda_{i,j}\left(\bar{F}^o_{i,j} - \bar{F}^i_{i,j}\right), \tag{10}
$$

which is independent of X with Λ known. To solve (10), a set of links are chosen to transmit under some interference constraints and proper routing algorithm to get the minimum.

Thus, by dual decomposition, the joint problem is decomposed into two separate problems corresponding to different protocol layers, and they interact with each other through Lagrangian price. Next, we explain how to determine the Lagrangian prices $[\lambda_{ij}]$.

B. Lagrangian Price

To obtain $[\lambda_{i,j}]$, consider the dual problem of (7)

$$
\min_{\Lambda} D(\Lambda), \tag{11}
$$

Since $D(\Lambda)$ is a convex function w.r.t. $[\lambda_{i,j}]$ [26], taking the partial derivative w.r.t. $\lambda_{i,j}$, we obtain:

$$
d_{i,j} \triangleq \frac{\partial D(\Lambda)}{\partial \lambda_{i,j}} = F^o_{i,j} - F^i_{i,j} - x_{i,j}. \tag{12}
$$

which is assumed to be uniformly bounded by constant D. According to Slater's theorem [26], there is no duality gap between (7) and (11). Thus, the Lagrangian price can be updated iteratively by:

$$\dot{\lambda}_{i,j}(t) \quad = \quad \left[y_{i,j}(X,t)\left(x_{i,j}(t) - F^o_{i,j}(t) + F^i_{i,j}(t)\right)\right]^+_{\lambda_{i,j}(t)}, \ \forall (i,j) \in \mathcal{L} \tag{13}$$

where $y_{i,j}(X,t) \geq 0$ is a user-specified function and if $y_{i,j}(X,t)$ is a constant, the Lagrangian price is proportional to the per-link queue length; and the projection $[a]^+_b$ is defined by:

$$[a]^+_b = \begin{cases} a, & \text{if } b \geq 0 \text{ or } a \geq 0 \\ 0, & \text{otherwise} \end{cases} .$$

C. SM Based Solution to the Congestion Control Problem

In this section, the SM controller for the congestion control problem (9) is developed for given $[\lambda_{i,j}]$. The SM control theory [15] [27] provides an optimal distributed method to iteratively solve convex optimization problems with a large number of constraints. Furthermore, the SM controller involves some adjustable parameters which correspond to the performance of the controller such as steady-state error and rate of convergence.

Here, we give a brief introduction to the sliding mode control theory. We illustrate the technique of sliding mode control by an optimization problem example. Consider a convex optimization problem

$$\begin{aligned} & \min_{Y} \quad s(Y) \\ s.t. \quad & f_1(Y) \quad \geq \quad 0 \qquad (14) \\ & f_2(Y) \quad \geq \quad 0 \qquad (15) \\ & f_3(Y) \quad \geq \quad 0 \qquad (16) \end{aligned}$$

with $s(Y)$ as the objective function and $Y \in \mathbb{R}^n$ as the optimization variables. The constraints (14) to (16) represent regions in $\mathbb{R}^n$ space and the intersection of these regions is the feasible region. The sliding mode controller produces a distributed solution whose trajectory is like sliding along the curve that is at the intersection of surfaces $f_1(Y) = 0$ and/or $f_2(Y) = 0$ and/or $f_3(Y) = 0$; the trajectory finally reaches the optimal point. For more details, we refer the readers to [15] and [17].

This motivates the design of our SM based controller for the congestion control problem (9) as

$$\dot{x^s_{i,j}}(t) = \left[z^s_{i,j}(X,t) \left(u^s_{i,j}(t) - \lambda_{i,j}(t) + \sum_{m \in H^s} v_m(t) \left. \frac{\partial h_m(x^s)}{\partial x^s_{i,j}} \right|_{x^s(t)} \right) \right]^+_{x^s_{i,j}(t)}, \forall (i,j) \in \mathcal{L} \tag{17}$$

where $z^s_{i,j}(X,t) \geq 0$ is a user-specified function; $u^s_{i,j}(t) \triangleq \left.\frac{\partial U^s(x^s)}{\partial x^s_{i,j}}\right|_{x^s(t)}$ is the partial derivative of $U^s(\cdot)$ w.r.t. $x^s_{i,j}(t)$ and evaluated at $x^s(t)$; and $\left.\frac{\partial h_m(x^s)}{\partial x^s_{i,j}}\right|_{x^s(t)} \triangleq b^s_{i,j}(t)$ is the partial derivative of h_m w.r.t. $x^s_{i,j}$ and $v_m(t)$ is defined as

$$v_m(x^s(t)) = \begin{cases} \gamma_m(t) & ,\text{if } h_m(x^s(t)) > 0 \\ 0 & ,\text{if } h_m(x^s(t)) \leq 0 \end{cases} \tag{18}$$

where $\gamma_m(t) > 0$ is a user-specified functions corresponding to $h_m(x^s)$ and we use $\gamma_m(t) = v > 0$ as an example in our following design, where v is a constant. The functions $[z^s_{i,j}(X,t)]$ and $[y_{i,j}(X,t)]$ are called iteration functions of the system. Here, we do not discuss the parameter design for SM control. Relevant works can be found in [16] and [17].

D. Solution to the Scheduling Problem

Similar to [2] [20], our solution to (10) at time t is to select a set of interference-free links $\mathcal{L}_o$ that maximize

$$\sum_{(i,j)\in\mathcal{L}_o} \lambda_{i,j}(t)\left(F^o_{i,j}(t) - F^i_{i,j}(t)\right), \tag{19}$$

where $F^i_{i,j}(t)$ and $F^o_{i,j}(t)$ are the time-averaged inflow rate and outflow rate at time t. According to the definition of $F^i_{i,j}(t)$ and $F^o_{i,j}(t)$, each node can estimate these parameters by tracking the arrival rate, transmission rate and receive rate, which are local information. The Lagrangian price $\lambda_{i,j}(t)$ is proportional to the queueing length in the node. Therefore, the solution to the scheduling problem turns out to be an maximum weighted independent set problem with weight $\lambda_{i,j}(t)\left(F^o_{i,j}(t) - F^i_{i,j}(t)\right)$. Alternatively, this solution can be simplified to yield a practical scheduling and routing algorithm, which will be presented in Section IV under time-varying channels.

E. Optimality and Convergence of the Joint Scheme

In summary, we solve (7) by iteratively computing the Lagrangian price for each node pair $(i,j) \in \mathcal{L}$, adjusting the transmission rates of each source and scheduling the links according to the network states. In this section, we prove the optimality and convergence property of the joint scheme, i.e., (13), (17) and (19), for the NUM problem (7), as stated below:

Theorem 1. *Assume that:*

1) $X(0)$ is a feasible rate matrix;

2) each utility function $U^s(\cdot)$ is concave w.r.t. the transmission rate x^s;

3) (13), (17) and (19) are the governing laws of the system;

4) $y_{i,j}(X,t) = y_{i,k}(X,t) = z_{i,j}^{s_1}(X,t) = z_{i,j}^{s_2}(X,t)$ $(\forall (i,j) \in \mathcal{L}, s_1 \neq s_2, k \neq j)$*, i.e., the iteration functions are the same for node* i*.*

Then the system state $(X(t), \Lambda(t))$ *will converge to the optimal solution, denoted by* $(\hat{X}, \hat{\Lambda})$*, where* $\hat{X}$ *is the optimal rate matrix for (7), and* $\hat{\Lambda}$ *is the optimal price matrix.*

The proof is sketched in Appendix A. Furthermore, if the optimal solution exists, and the set of all optimal solutions is bounded, the SM controller can make the rate matrix converge to an optimal point from any initial rate matrix, which is not necessarily within the capacity region [15] [16].

Intuitively, for a given Λ, (17) makes the transmission rates converge to their optimality similar to the case in the Internet [17]. Meantime, the scheduler selects an optimal set of links to transmit based on the given Λ as in (19). These decisions produce an updated Λ, and finally the transmission rates and the Lagrangian price converge simultaneously to an optimal solution of (7). It is pointed out in Theorem 1 that $y_{i,j}(X,t) = y_{i,k}(X,t) = z_{i,j}^{s_1}(X,t) = z_{i,j}^{s_2}(X,t)$ $(\forall (i,j) \in \mathcal{L}, s_1 \neq s_2, j \neq k)$ is a sufficient condition for the convergence of the joint iteration (13) and (17) under constant channels. Thus, the iteration functions $z_{i,j}^s, \forall s \in S$ are the same for all flows originating from node i, which are also the same as the iteration functions $y_{i,j}$ over multipaths of these flows. This naturally allows the co-existence of heterogeneous components in a network. We can see that the control law given in [11] has $z_{i,j}^s = y_{i,j} = 1, \forall s \in S, \forall (i,j) \in \mathcal{L}$, which is a special case for the design in (13) and (17).

IV. Extension to Time-varying Channels

In this section, the NUM problem of a wireless network is considered under slotted time-varying channels. Based on the control laws of (13), (17) and (19), we design Algorithm 1 and Algorithm 2, which asymptotically solve the NUM problem under time-varying channels.

For simplicity, we consider a wireless network with slotted time-varying channels, the channel gains of which are i.i.d. convergent [24]. Different from the transmission rate under constant channels, the instantaneous transmission rate under time-varying channels changes from slot to slot, thus the optimized variables are $[\bar{x}_{i,j}^s]$, defined in a time-average sense. Accordingly, the QoS constraints are also defined on the time-averaged transmission rate. The network utility for a flow is a function of the average transmission rate and an optimal dynamic control policy is defined over all possible power control, routing and scheduling that can realize the optimal average transmission rate $\hat{X}$.

Similarly, the NUM problem under time-varying channels is formulated as

$$\max_{\{X\in\mathcal{X},P\}} \sum_{s\in\mathcal{F}} U^s(\bar{x}^s) \tag{20a}$$

$$s.t.: \quad P\in\mathcal{P} \tag{20b}$$

$$\bar{x}_{i,j} - \bar{F}^o_{i,j} + \bar{F}^i_{i,j} \le 0, \forall(i,j)\in\mathcal{L} \tag{20c}$$

$$\mathbf{h}\le 0 \tag{20d}$$

$$\bar{x}^s = \sum_{i=b^s}\sum_{j\in\ \mathcal{N}_i} \bar{x}^s_{i,j}, \tag{20e}$$

where (20b) is the constraint on power and each element of which is subject to constraints on minimum SINR and maximum transmission power. (20) allows many types of traffic such as constant bit rate (CBR), variable bit rate (VBR) and elastic traffic.

Although (20) seems similar to (7), it cannot be directly solved by dual decomposition, which may result in an unfeasible link assignment or transmission rate. Instead, we directly extend the control law, i.e., (13), (17) and (19), to the time-varying system settings and evaluate its performance in Section V.

The SM based joint congestion control and scheduling algorithm under time-varying channels is given in Algorithm 1 where Δ is the sampling interval for the discrete time system.

1 *Initialization*: $X(0)\in\mathcal{X}$, $F^o_{i,j}(0)=0$, $F^i_{i,j}(0)=0$, $\Lambda(0)=0$;

2 At time $t+1$, update the system states according to:

3 *Price*: the Lagrangian price is updated according to

$$\lambda_{i,j}(t+1) = \left\lfloor \left[y_{i,j}(X,t)\Delta\left(x_{i,j}(t) - F^o_{i,j}(t) + F^i_{i,j}(t)\right)\lambda_{i,j}(t)\right]^+ \right\rfloor; \tag{21}$$

4 *Congestion Control*: each source updates its multi-path transmission rate by

$$x^s_{i,j}(t+1) = \left\lfloor \left[x^s_{i,j}(t) + z^s_{i,j}(t)\Delta\left(u^s_{i,j}(t) - \lambda_{i,j}(t) + vb^s_{i,j}(t)\right)\right]^+ \right\rfloor; \tag{22}$$

5 *Scheduling*: perform Alg. 2 to get the optimal power control.

Algorithm 1: SM based congestion control and scheduling

The time average inflow rate $F^i_{i,j}(t)$ and outflow rate $F^o_{i,j}(t)$ at time t can be represented as

$$F^i_{i,j}(t) = \frac{\sum_m R_{m,i,j}(t-1) + \sum_m \underline{\mu}_{m,i,j}(t)}{t} \tag{23a}$$

$$F^o_{i,j}(t) = \frac{\sum_k D_{i,j,k}(t-1) + \mu_{i,j}(t)}{t}, \tag{23b}$$

where $R_{m,i,j}(t-1)$ is the total transmitted data from slot 0 to slot $t-1$ over $q_{m,i}$ to $q_{i,j}$ and $D_{i,j,k}(t-1)$ is the total transmitted data from $q_{i,j}$ to $q_{j,k}$; $\underline{\mu}_{m,i,j}(t)$ and $\mu_{i,j}(t)$ satisfy (2).

Substituting (23) into (19), we obtain the optimal power control as below

$$\begin{aligned} P(t) &= \arg\max_{P\in\mathcal{P}} \sum_{i,j} \lambda_{i,j}(t) \left(\mu_{i,j}(t) - \sum_{m} \underline{\mu}_{m,i,j}(t) \right) \\ &= \arg\max_{P\in\mathcal{P}} \sum_{i,j} \left(\lambda_{i,j}(t)\mu_{i,j}(t) - \sum_{k} \underline{\mu}_{i,j,k}(t)\lambda_{j,k}(t) \right), \end{aligned} \tag{24}$$

where (24) is due to the fact that the decision at time t does not affect $\sum_m R_{m,i,j}(t-1)$ and $\sum_k D_{i,j,k}(t-1)$. It is recognized that (24) is a maximum weighted independent set (MWIS) problem with $\lambda_{i,j}(t)\mu_{i,j}(t) - \sum_k \underline{\mu}_{i,j,k}(t)\lambda_{j,k}(t)$ assigned as the weight $W_{i,j}$ for each $l_{i,j}$. Then (24) can be solved in three steps based on a predefined routing algorithm $\mathcal{R}$, as stated in Algorithm 2, where $L_m^k(i,j)(t)$ is the length of the mth packet in $q_{i,j}$ at time t, which will be transmitted from node i to node j and put into $q_{j,k}$; and $\Gamma_{i,j}(t) = \{\gamma | \sum_{\forall k, m=1}^{m=\gamma} L_m^k(i,j) \leq \mu_{i,j}(t)\}$ is the set of packet index for transmission, which also depends on P and G.

1 At time t for each $q_{i,j}(\forall (i,j) \in \mathcal{L})$, decide the next hop node k of each packet that not destined to node j by

$$k = \arg\min_{n:(j,n)\in\mathcal{R}} \lambda_{j,n}(t),$$

then get $\underline{\mu}_{i,j,k}(t) = \sum_{m\in\Gamma_{i,j}(t)} L_m^k(i,j)(t)$;

2 The weight $W_{i,j}(t)$ for any link $l_{i,j}$ is calculated by:

$$W_{i,j}(t) = \mu_{i,j}(t)\lambda_{i,j}(t) - \sum_k \underline{\mu}_{i,j,k}(t)\lambda_{j,k}(t);$$

3 Allocate transmission power P according to:

$$P(t) = \arg\max_{P\in\mathcal{P}} \sum_{i,j} W_{i,j}(G(t), P); \tag{25}$$

Algorithm 2: The scheduling and multi-path routing algorithm under per-link queues

It is easy to justify that Algorithm 2 gives an optimal solution to (24) based on routing scheme $\mathcal{R}$.

The Lagrangian price (21) and the transmission rate (22) at each node are updated in a distributed manner with low computational complexity and the information used in Algorithm 1 is local rather than global. Given the link price $[\lambda_{i,j}]$, (25) is generally $\mathcal{NP}$ hard, even if the power control is on-off with two values. If suboptimal performance is allowed, some algorithms such as column generation [28] can be used to solve the MWIS problem. The MWIS problem can also be solved in a distributed manner with low complexity by the algorithm mentioned in [20], which usually achieves a performance within about 4/5 of the optimal performance. There are also some works studying the impact of the imperfect scheduling and its bound such as [29] [30].

Algorithm 1 cannot be derived directly from the dual decomposition of the NUM problem (20). However, (20) can be used as a reference system to characterize the performance of Algorithm 1. The performance of Algorithm 1 is studied in the next section.

V. Performance Analysis

In this section, Section V-A shows the performance of Algorithm 1. To enhance the performance of Algorithm 1 in terms of delay, we propose a Very Important Packet (VIP) queueing structure in Section V-B.

A. Performance of Algorithm 1

The system states under the discrete-time control policy evolve as a Markov chain and we need to show that this Markov chain is stable. Because we apply the floor operation to the Lagrangian price and the transmission rate, it is easy to check that the Markov chain has a countable state space, but is not necessarily irreducible. Thus, we consider the partition of the state space as the transit state set and the recurrent state set. It is defined to be stable if all recurrent states are positive recurrent and the Markov process hits the recurrent states with probability one [8] as t approaches infinity. This can guarantee that the Markov chain is absorbed/reduced into some recurrent class, and the positive recurrence ensures the periodicity of the Markov chain over this class. Then we evaluate its stability by the following theorem:

Theorem 2. *The Markov chain described by (21) (22) is stable.*

Theorem 3. *The network utility produced by Algorithm 1 converges statistically to the neighborhood of the optimal utility with radius* $\left(\frac{K_2|S|\Delta(V_0^2+D_0^2+Z_1+Z_2)}{2K_1}\right)^{\frac{1}{2}}$, *i.e.,*

$$\left[\sum_s \mathbb{E}\left\{U^s\left(x^s(\infty)\right)\right\} - \sum_s \hat{U}^s\right]^2 \leq \frac{K_2|S|\Delta(V_0^2 + D_0^2 + Z_1 + Z_2)}{2K_1} \tag{26}$$

where (26) is the result under $y_{i,j}(X,t) = y_{i,k}(X,t) = z_{i,j}^{s_1}(X,t) = z_{i,j}^{s_2}(X,t) = 1$ $(\forall (i,j) \in \mathcal{L}, s_1 \neq s_2, k \neq j)$ *and similar result holds for* $y_{i,j}(X,t) = y_{i,k}(X,t) = z_{i,j}^{s_1}(X,t) = z_{i,j}^{s_2}(X,t) \neq 1$ $(\forall (i,j) \in \mathcal{L}, s_1 \neq s_2, k \neq j)$*;* $x^s(\infty)$ *denotes the transmission rate of the Markov chain in steady state, and* V *and*

D are constant bounds of function $(u^s_{i,j} - \lambda_{i,j} + vb^s_{i,j})^+$ and the partial derivative of the dual function $d_{i,j}$; and $K_2 = \max_s |u^s(x^s(\infty))|^2$ and $K_1 > 0$ is the largest lower bound of $\left|\frac{u^s_{i,j} - \hat{u}^s_{i,j}}{x_{i,j}(t) - \hat{x}_{i,j}}\right|$.

The proofs of Theorem 2 and 3 are given in Appendix B.

Theorems 2 and 3 imply that the expectation of the network utility converges to a small neighborhood of the optimal utility. Thus, the iterations of transmission rates and the Lagrangian prices converge simultaneously. There may be other decomposition possibilities as studied in [23].

Here, we provide some intuition about our sliding mode based joint congestion control and scheduling design. In queueing networks, the NUM problem can be decomposed into a scheduling problem and a congestion control problem, which produces the classic primal-dual algorithm [31]. The dual algorithm turns out to be the back-pressure routing/scheduling, which can stabilize the network whenever the source arrival rate is within the capacity region of the wireless network [8] [9]. Since the QoS requirements in our problem are functions of source transmission rate, these constraints will only affect the solution to the congestion control problem rather than the stability of the network within the capacity region. Since the sliding mode control theory provides a distributed optimal solution to convex optimization problems, it can be applied to solving the congestion control problem.

Note the the capacity region under per-link queueing may be different from that under per-destination queueing. Intuitively, the per-destination queueing provides smaller granularity for scheduling and it can be easily proved that the capacity region under per-destination queueing is not less than that under per-link queueing. Little work is done about this comparison. In Sec. VI, we present the simulation results about the throughput comparison of a network under both per-link queueing and per-destination queueing.

B. Performance Enhancement Mechanism: VIP Queues

In the above system model, we do not consider the services with stringent delay requirements. In this section, we propose a framework for the co-existence of the per-link queues (non-VIP queues) and the VIP queues so that delay-sensitive services can be supported. The VIP queues and non-VIP queues are maintained separately along the routes of delay-sensitive applications and the VIP queues will be scheduled by some delay-aware algorithms. Since these applications may have much less number of destinations, the per-destination queues can also be used. One application is for interactive multi-class such as VoIP, online multiplayer games, which have stringent delay requirements. The packets from interactive multimedia applications are marked as VIPs and will be put into VIP queues and transmitted with higher priority over non-VIPs. Another application is the transmission of layered video, where

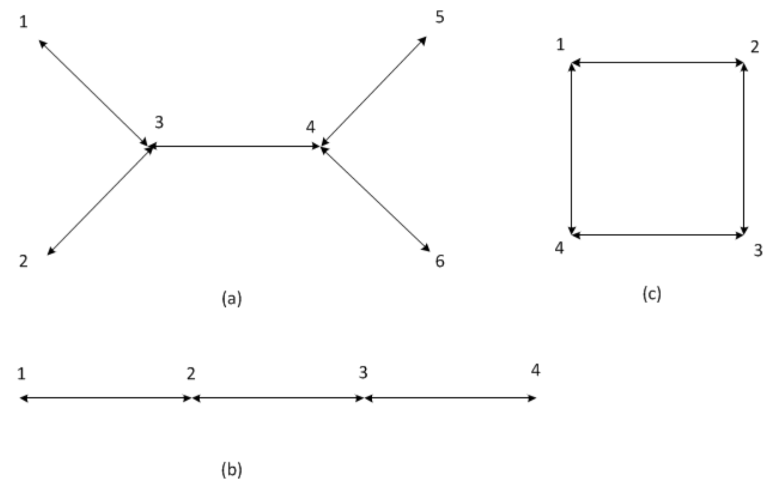

Fig. 1. Network topology: (a) Dumbbell topology. (b) Chain topology. (c) Loop topology

the base-layer packets can be placed into the VIP queues and transmitted with high priority, and the enhancement layer packets are placed into the non-VIP queues and transmitted with low priority. There are some theoretical analysis of the performance regarding priority queues [32].

VI. Simulation Results

In this section, we demonstrate the performance of Algorithm 1 by simulating ad hoc networks in three topologies with typical traffic patterns. Section VI-A describes the simulation settings, and Section VI-B presents the simulation results of Algorithm 1 under constant and time-varying time division multiple access (TDMA) channels.

A. Simulation Settings

The ad hoc network topology that we simulate is depicted in Fig. 1, where Figs. 1(a), 1(b) and 1(c) show a dumbbell topology, a chain topology and a loop topology respectively. The lines in Fig. 1 indicate bi-directional links and the distance between any two neighboring nodes is the same. Each constant wireless channel has a capacity of 1.4Mb/s. The time-varying channel state can be one of the five states/rates, i.e., $\{0.28, 0.84, 1.40, 1.96, 2.52\}$Mb/s, with probability $1/5$. Thus, the expectation of the transmission rate for a time-varying channel is 1.4Mb/s. Without loss of generality, only one-hop interference is considered, i.e., each node cannot transmit and receive simultaneously.

TABLE I
PARAMETERS

Topology	Services	QoS	Parameters
Dumbbell	s1:$\{l_{1,3}\}$ s2:$\{l_{1,3},\ l_{3,2}\}$ s3:$\{l_{1,3},\ l_{3,4}\}$ s4:$\{l_{1,3},\ l_{3,4},\ l_{4,5}\}$ s5:$\{l_{1,3},\ l_{3,4},\ l_{4,6}\}$	No QoS	$\Delta = 0.01$(TVC)
Chain (CC)	s1:$\{l_{1,2},\ l_{2,3}\}$ s2:$\{l_{2,3},\ l_{3,4}\}$ s3:$\{l_{3,4}\}$ s4:$\{l_{1,2},\ l_{2,3},\ l_{3,4}\}$	$x^2 \geq 350kbps$ $x^4 = 100kbps$	$\Delta = 1$ $v = 200$
Chain (TVC)	s1:$\{l_{1,2},\ l_{2,3}\}$ s2:$\{l_{2,3},\ l_{3,4}\}$ s3:$\{l_{3,4}\}$ s4:$\{l_{1,2},\ l_{2,3},\ l_{3,4}\}$	$x^2 \geq 350kbps$ $x^4 = 100kbps$	$\Delta = 0.01$ $v = 200$
Loop	s1:$\{l_{1,2},\ l_{2,3}\}$ and $\{l_{1,4},\ l_{4,3}\}$ s2:$\{l_{2,3}\}$	$x^1 \geq 350kbps$	$\Delta = 0.01$(TVC) $v = 300$

The utility function of each flow is:

$$U^s(x^s) = w_s \ln(x^s + 1) \tag{27}$$

where w_s is a user specified parameter to indicate different weight of the service, and we choose identically $w_s = 2000$ here. In Algorithm 1 and 2, design parameter v and the scalar iterative step-size Δ are also user-specified parameters and listed in Table I, where CC is shorthand notation for constant channels and TVC is shorthand notation for time-varying channels.

B. Performance of Algorithm 1

1) Throughput Performance: In this section, we study the throughput performance of Algorithm 1. The set up of the simulation is the following: in a 6-node ad hoc network in dumbbell topology, as shown in Fig. 1(a), there are five best effort (BE) services from node 1 to node 2, 3, 4, 5, 6 respectively. The time-varying channel states are i.i.d. uniformly distributed with transmission rate $\{40, 120, 200, 280, 360\}$kb/s. For comparison, it is simulated under Algorithm 1 with per-link queueing and under the MaxWeight algorithm [8] with per-destination queueing. It is known that the MaxWeight achieves the throughput

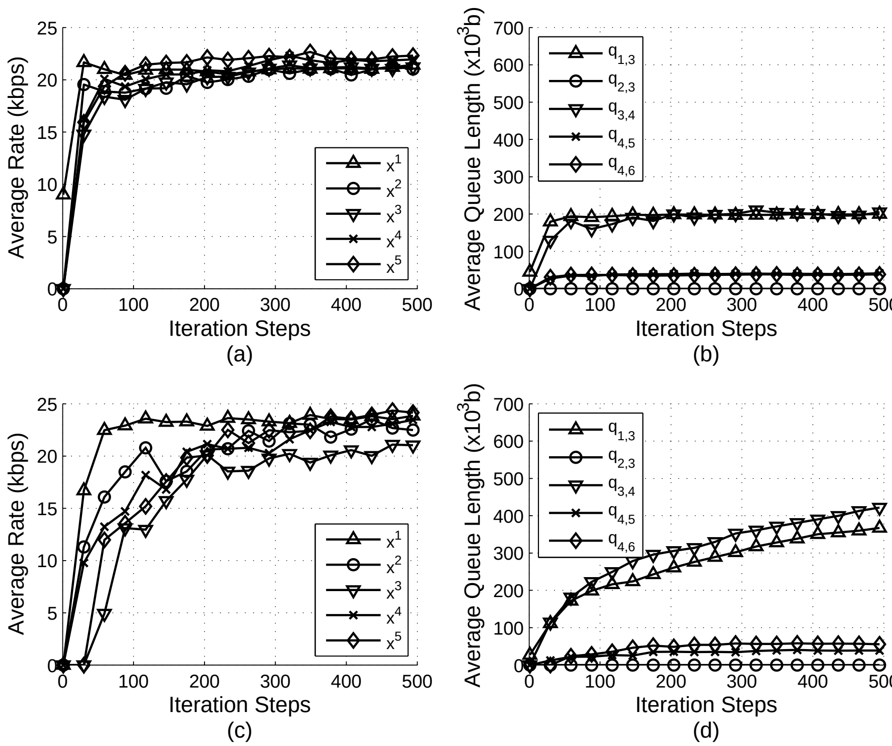

Fig. 2. Transmission rate and queue length of dumbbell topology: (a) The average transmission rate under Alg. 1 and per-link queueing. (b) The average queue length under Alg. 1 and per-link queueing. (c) The average transmission rate under MaxWeight and per-destination queueing. (d) The average queue length under MaxWeight and per-destination queueing.

optimal under per-destination queueing. To get the optimal throughput under Algorithm 1, we choose transmission rate as the utility function for each node. The routing table applied to Algorithm 1 is listed in Table I, while the MaxWeight algorithm uses the back-pressure routing.

Fig. 2 illustrates the time-averaged rates of data reaching each destination, and queue length vs. time under both queueing models. From Fig. 2(a), we can see the time-averaged data rate at each destination converges to about 22kb/s, which means the per-link queues in the dumbbell network is able to support such data rates. The time-averaged queue length also converges fast to finite value as shown in Fig. 2(b). This data rate can also be supported by the per-destination queue/MaxWeight (not shown in simulation results). However, with a transmission rate 23kb/s for each service, the network under per-destination queueing is unstable. It is seen from Fig. 2(c) that the time-average data rates at some destination nodes are obviously below 23kb/s. Accordingly, the queues are not stable in 2(d) because the rate of data entering the network is more than the rate of data reaching the destination. In the per-link queueing, node 3 is the only next hop node of node 1, thus the source rates of node 1 will converge to the same rate according to Algorithm 1 under the same utility function, which is also validated by Fig. 2(a).

Since it is very difficult to get an analytical results of the capacity region of a wireless network, we have to resort to simulation to roughly compare the throughput of a network under the per-link queueing and the per-destination queueing. It is seen that with reasonable routing algorithm applied, Algorithm 1 can achieve a throughput performance close to MaxWeight but using much less queues. There are $6 \times 6 = 36$ queues under the per-destination queueing whereas only 10 queues under the per-link queueing.

2) Supporting Multi-class Services: In this section, we evaluate whether Algorithm 1 can simultaneously support various QoS required by different services. Both elastic and inelastic services are simulated in a 4-node ad hoc network in chain topology, as shown in Fig. 1(b). There are four services in the network. $s1$ and $s3$ are BE flows such as email and file transfer protocol (FTP). $s2$ has a minimum transmission rate requirement of 350kb/s (in time-average sense under time-varying channels), which can provide the VBR flows such as video over IP. $s4$ is a CBR flow transmitting at 100kb/s, which represents 10 VoIP streams (each with 10kb/s).

Fig. 3 shows the time average transmission rates vs. iteration steps without/with QoS requirements under constant channels (CC) in the 4-node chain ad hoc network. In Fig. 3(a), we compare the convergent transmission rates of the ad hoc network under Algorithm 1 (per-link queueing) and the MaxWeight algorithm (per-destination queueing). The transmission rates under MaxWeight are denoted by $x1^1, x1^2, x1^3$. Since $s4$ transmits at a constant rate, it is not shown in the figure. It is shown that the throughput of Algorithm 1 is slightly smaller than the Maxweight algorithm which is throughput optimal and uses more queues. It is also noticed that the QoS constraints under Maxweight are not satisfied. In addition, similar simulation results can be observed in Fig. 4 under time-varying channels. If QoS requirement is not considered, the transmission rate for x^1 is much lower than $350kbps$, as shown in Fig. 3(a). In comparison, the QoS requirements are satisfied in Fig. 3(b) for all flows under Algorithm 1. Comparing the network utility under no QoS with that under QoS requirements, the former is larger than the latter, which is reasonable because the feasible region is smaller under QoS constraints. In this chain topology, $l_{1,2}$ and $l_{3,4}$ can be activated simultaneously. $s1$, $s2$ and $s4$ pass $l_{2,3}$, which has to be activated with no other active links. Thus, $l_{2,3}$ is the bottleneck and x^2 is expected to be smaller without QoS constraints, which also matches our simulation results. Similar results are observed in Figs. 4(a) and 4(b) under time-varying channels where $s2$ is the VBR flow. In summary, all these simulation results demonstrate that Algorithm 1 is capable of supporting traffic with multi-class QoS requirements. One possible disadvantage of Algorithm 1 is that the convergence of the transmission rates may be slower.

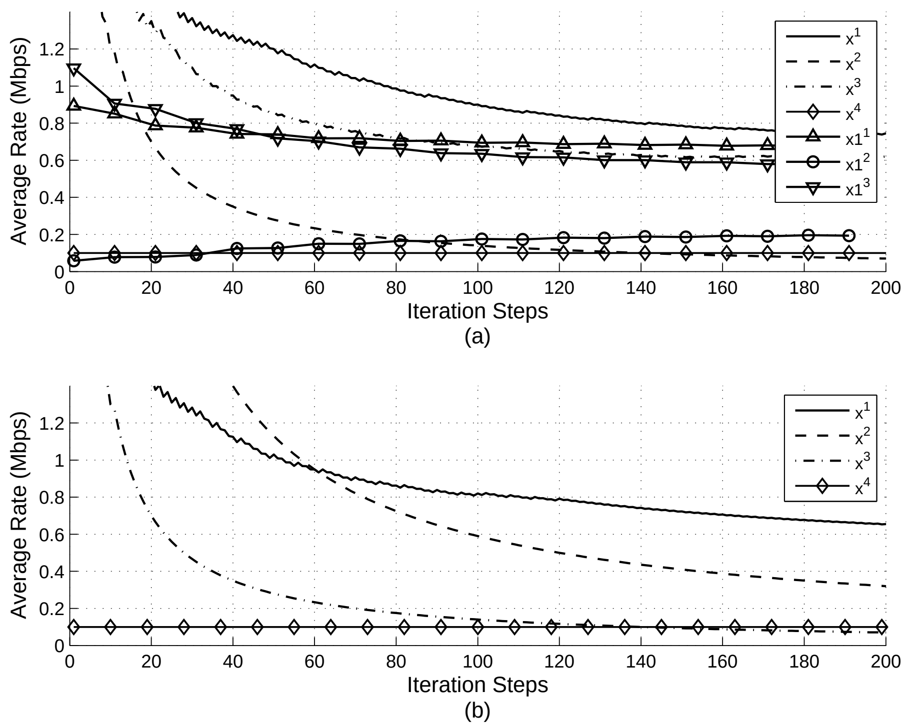

Fig. 3. Transmission rate for chain topology under constant channels: (a) The average transmission rate under Alg. 1 and MaxWeight without QoS requirements. (b) The average transmission rate under Alg. 1 with QoS requirements.

3) Robustness against Link Failure: In this section, we study the robustness of Algorithm 1 against link failure under time-varying channels. To emphasize that Algorithm 1 can support multi-path, a 4-node loop ad hoc network is simulated, as shown in Fig. 1(c). $s1$ is a video-over-IP flow with minimum transmission rate requirement of 350kb/s and $s2$ is BE flow. There is multi-path routing identified for $s1$ through $l_{1,2}$ and $l_{1,4}$ respectively.

Fig. 5 illustrates the multi-path transmission rates vs. iteration steps under loop topology shown in Fig. 1(c) under time-varying channels. In Figs. 5(a) and 5(b), the QoS requirement is satisfied for $s1$ by the sum of the transmission rates over two paths. The time-average queue length also converges. Since $s1$ and $s2$ have to share $l_{2,3}$, the transmission rate $x^1_{1,4}$ over $l_{1,4}$ is expected to be larger than $x^1_{1,2}$ over the other path, which matches the simulation results in Fig. 5(a). For comparison, Figs. 5(c) and 5(d) show the transmission rate and queue length in the case of the link failure of $l_{1,4}$. When encountering a link failure, it is shown in Fig. 5(d) that $q_{1,4}$ corresponding to $l_{1,4}$ increases quickly, which decreases the transmission rate $x^1_{1,4}$ quickly to zero. In this case, the QoS requirement is not satisfied and the

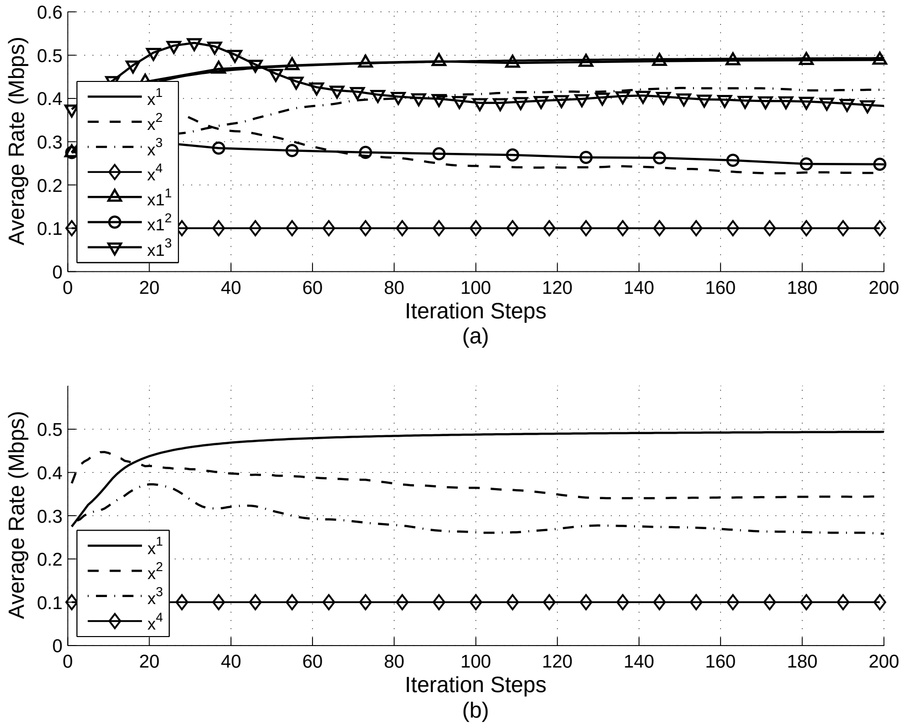

Fig. 4. Transmission rate for chain topology under time-varying channels: (a) The average transmission rate under Alg. 1 and MaxWeight without QoS requirements. (b) The average transmission rate under Alg. 1 with QoS requirements.

the transmission rate $x^1_{1,2}$ of the other path increases according to (22). Finally the QoS requirement is satisfied as shown in Fig. 5(c) and the network is stable. These simulation results show that Algorithm. 1 is robust against link failure because the multi-path transmission adaptively does load balancing to satisfy QoS requirements without changing the routing table.

VII. Conclusions

In this paper, we study the joint congestion control and scheduling problem for multi-hop, multi-path per-link queueing wireless networks with QoS constraints. It is formulated as an NUM problem under both constant and time-varying channels, which is decomposed into a congestion control problem and a scheduling problem. Then, a distributed SM based controller is designed to iteratively solve the congestion control problem, which can provide multi-path rate adaption to satisfy QoS constraints. The scheduling problem is identified as a maximum weighted independent set problem and can also be solved in a distributed manner. We extend this algorithm to the case with time-varying channels and prove that the dynamic control law, i.e., Algorithm 1 is stable and the network utility converges

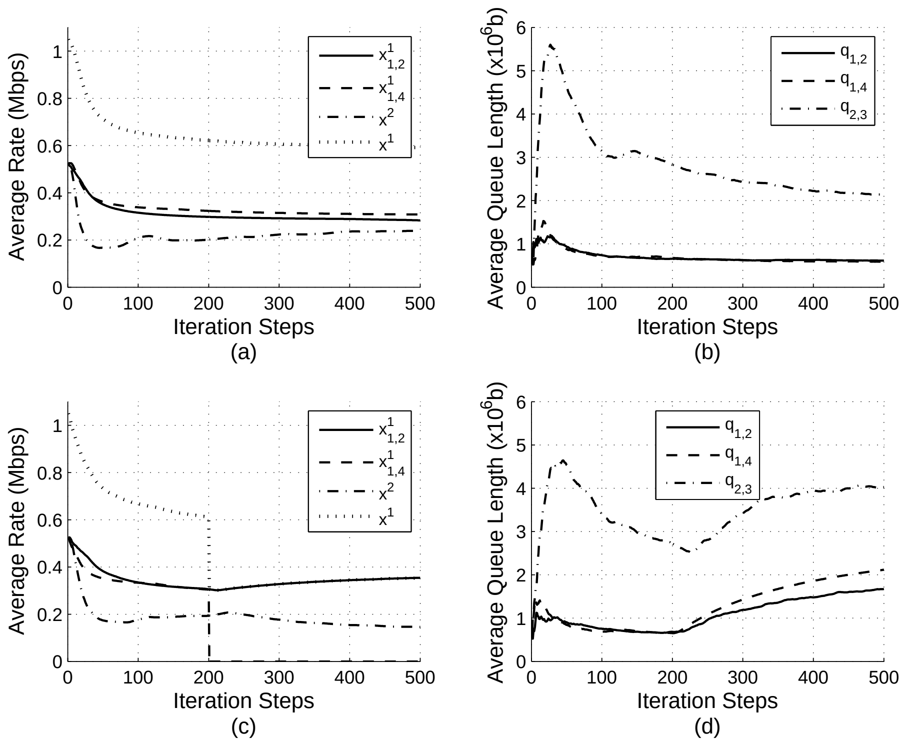

Fig. 5. Transmission rate for loop topology: (a) The transmission rate under multi-path with TVC and QoS requirements. (b) The queue length under multi-path with TVC and QoS requirements. (c) The transmission rate under link failure for $l_{1,4}$ with TVC and QoS requirements. (d) The queue length under link failure for $l_{1,4}$ with TVC and QoS requirements.

to a bounded neighborhood of the optimal. Simulation results show that Algorithm 1 is capable of providing heterogenous multi-class services with different QoS requirements. Because of the multi-path load balancing feature of this algorithm, it is robust against network anomalies such as link failure.

The application of the SM control theory in wireless networks is tightly related to the decomposition property of the protocol stack. Since the joint congestion control and scheduling is a cross-layer optimization, these two subproblems interact with each other through the Lagrangian prices. Therefore, the sources adapt the transmission rate to the price, which masks the link status such as time-varying channel capacities and dynamic scheduling. At the same time, the scheduling problem is independent of the source transmission rates. This property makes it possible to apply SM control theory. Furthermore, the convergence conditions of the SM controller are also relaxed because the decomposed sub-problems provide a way of "negative feedback" to each other in wireless networks.

APPENDIX A
PROOF OF THEOREM 1

Proof: We sketch the proof of Theorem 1 below, and interested readers can find the design law of the SM controller in [15] [16]. To prove the convergence and optimality of the algorithm, we construct the Lyapunov function as

$$W(X,\Lambda) \quad = \quad W_1(X) + W_2(\Lambda) \tag{28}$$

where

$$\begin{aligned} W_1(\Lambda) &\triangleq \frac{1}{2}\sum_{(i,j)\in\mathcal{L}} \left(\lambda_{i,j} - \hat{\lambda}_{i,j}\right)^2 = \frac{1}{2}\| \ \Lambda - \hat{\Lambda}\|_2^2 \\ W_2(X) &\triangleq \frac{1}{2}\sum_{s\in S, j=n^s} \left(x_{i,j}^s - \hat{x}_{i,j}^s\right)^2 = \frac{1}{2}\| \ X - \hat{X}\|_2^2; \end{aligned} \tag{29}$$

and ($[\hat{\lambda}_{i,j}]$, $[\hat{x}^s_{i,j}]$) are the optimal solutions to problem (7) (assumed to be feasible), thus they are constants. For brevity, we omit the "(t)" of the quantities below. Differentiating (28) with respect to t, we obtain:

$$\begin{aligned} &\dot{W}(X,\Lambda) \\ =&\sum_{s,j} \left(x_{i,j}^s - \hat{x}_{i,j}^s\right)\left[z_{i,j}^s\left(u_{i,j}^s + vb_{i,j}^s - \lambda_{i,j}\right)\right]^+_{x_{i,j}^s} + \sum_{i,j}\left(\lambda_{i,j} - \hat{\lambda}_{i,j}\right)\left[y_{i,j}\left(x_{i,j} + F_{i,j}^i - F_{i,j}^o\right)\right]^+_{\lambda_{i,j}} && (30) \\ \leq&\sum_{s,j} \left(x_{i,j}^s - \hat{x}_{i,j}^s\right) z_{i,j}^s\left(u_{i,j}^s + vb_{i,j}^s - \lambda_{i,j}\right) + \sum_{i,j}\left(\lambda_{i,j} - \hat{\lambda}_{i,j}\right) y_{i,j}\left(x_{i,j} + F_{i,j}^i - F_{i,j}^o\right), && (31) \end{aligned}$$

where the last inequality follows that (30) and (31) are equal if the projection in (13) and (17) are inactive, and if the projection is active, the projection in (13) and (17) are zero, while the expression in (31) are positive due to $x_{i,j} + F_{i,j}^i - F_{i,j}^o < 0$ and $u_{i,j}^s + vb_{i,j}^s - \lambda_{i,j} < 0$. Here, "$\sum_{i,j}$" and "$\sum_{s,j}$" is a shorthand notations for $\sum_{(i,j)\in\mathcal{L}}$ and $\sum_{s\in S, i=b^s, j\in\mathcal{N}_i}$.

Since $[\hat{x}_{i,j}^s]$ is the optimal solution to the NUM problem, then the fact comes that $\hat{\lambda}_{i,j} = \hat{u}_{i,j}^s$, where

$\hat{u}^s_{i,j} = \frac{\partial U^s(x^s)}{\partial x^s_{i,j}}\Big|_{\hat{x}^s_{i,j}}$. Substituting it into (31) yields

$$\begin{aligned}
&\dot{W}(X,\Lambda) \\
\leq\ & \sum_{s,j}\left(x^s_{i,j}-\hat{x}^s_{i,j}\right)z^s_{i,j}\left(u^s_{i,j}-\hat{u}^s_{i,j}-\lambda_{i,j}+\hat{\lambda}_{i,j}+vb^s_{i,j}\right)+\sum_{i,j}y_{i,j}\left(\lambda_{i,j}-\hat{\lambda}_{i,j}\right)\left(x_{i,j}-\hat{x}_{i,j}\right)+ \\
& \sum_{i,j}y_{i,j}\left(\lambda_{i,j}-\hat{\lambda}_{i,j}\right)\left(\hat{x}_{i,j}+F^i_{i,j}-F^o_{i,j}\right) \qquad (32) \\
=\ & \sum_{s,j}z^s_{i,j}\left(x^s_{i,j}-\hat{x}^s_{i,j}\right)\left(u^s_{i,j}-\hat{u}^s_{i,j}\right)+\sum_{s,j}z^s_{i,j}\left(x^s_{i,j}-\hat{x}^s_{i,j}\right)\left(\hat{\lambda}_{i,j}-\lambda_{i,j}\right)+ \\
& \sum_{s,j}z^s_{i,j}\left(x^s_{i,j}-\hat{x}^s_{i,j}\right)vb^s_{i,j}+\sum_{i,j}y_{i,j}\left(\lambda_{i,j}-\hat{\lambda}_{i,j}\right)\left(x_{i,j}-\hat{x}_{i,j}\right)+ \\
& \sum_{i,j}y_{i,j}\left(\lambda_{i,j}-\hat{\lambda}_{i,j}\right)\left(\hat{x}_{i,j}+F^i_{i,j}-F^o_{i,j}\right) \qquad (33) \\
=\ & \sum_{s,j}\left(x^s_{i,j}-\hat{x}^s_{i,j}\right)z^s_{i,j}\left(u^s_{i,j}-\hat{u}^s_{i,j}\right)+\sum_{s,j}z^s_{i,j}\left(x^s_{i,j}-\hat{x}^s_{i,j}\right)vb^s_{i,j}+ \\
& \sum_{i,j}y_{i,j}\left(\lambda_{i,j}-\hat{\lambda}_{i,j}\right)\left(\hat{x}_{i,j}+F^i_{i,j}-F^o_{i,j}\right) \qquad (34)
\end{aligned}$$

The second term and the forth term in (33) cancel each other if $z^s_{i,j} = y_{i,j}$ $(\forall s \in S, b^s = i, n^s = j)$ because flows sharing the same source node i and the next hop node j relate to the same $\lambda_{i,j}$, then

$$\sum_{s,j}z^s_{i,j}\left(x^s_{i,j}-\hat{x}^s_{i,j}\right)(\hat{\lambda}_{i,j}-\lambda_{i,j}) = \sum_{i,j}y_{i,j}\left(\hat{\lambda}_{i,j}-\lambda_{i,j}\right)\left(x_{i,j}-\hat{x}_{i,j}\right)$$

Because U^s is concave w.r.t. x^s, and $u^s_{i,j} = u^s$ $(\forall j \in \mathcal{N}_i)$, where $u^s \triangleq \frac{dU^s}{dx^s}\Big|_{x^s}$. Then we have:

$$\begin{aligned}
& \sum_{s,j}z^s_{i,j}\left(x^s_{i,j}-\hat{x}^s_{i,j}\right)\left(u^s_{i,j}-\hat{u}^s_{i,j}\right) \\
=\ & z^s_{i,j}\sum_{s}\left(\sum_{j\in\mathcal{N}_i}x^s_{i,j}-\sum_{j\in\mathcal{N}_i}\hat{x}^s_{i,j}\right)\left(u^s_{i,j}-\hat{u}^s_{i,j}\right) \qquad (35) \\
=\ & z^s_{i,j}\sum_{s}\left(x^s-\hat{x}^s\right)\left(u^s-\hat{u}^s\right) \qquad (36) \\
\leq\ & 0
\end{aligned}$$

where $\hat{u}^s \triangleq \frac{dU^s}{dx^s}\Big|_{\hat{x}^s}$ and (35) follows that $u^s_{i,j} - \hat{u}^s_{i,j}$ depends on $x^s_{i,j}$ in terms of x^s for a source s; and (36) follows that U^s is a concave function of x^s.

From constraint (7c), we have $\hat{x}_{i,j} \leq \hat{F}^o_{i,j} - \hat{F}^i_{i,j}$. Because $F^i_{i,j}$, $F^o_{i,j}$ and $\lambda_{i,j}$ are the optimal solution for (10) at time t and $\hat{F}^i_{i,j}$ and $\hat{F}^o_{i,j}$ are feasible solutions to the outflow and inflow, thus

$$\sum_{i,j}\lambda_{i,j}(\hat{x}_{i,j}+F^i_{i,j}-F^o_{i,j}) \leq \sum_{i,j}\lambda_{i,j}(\hat{F}^o_{i,j}-\hat{F}^i_{i,j}+F^i_{i,j}-F^o_{i,j}) \leq 0. \qquad (37)$$

From the KarushKuhnKucker (KKT) conditions, $\lambda_{i,j}\left(\hat{x}_{i,j}+\hat{F}^{i}_{i,j}-\hat{F}^{o}_{i,j}\right)=0$ holds and

$$\begin{aligned}
&\sum_{i,j}\left(-\hat{\lambda}_{i,j}\right)\left(\hat{x}_{i,j}+F^{i}_{i,j}-F^{o}_{i,j}\right)\\
=\ &\sum_{i,j}\left(-\hat{\lambda}_{i,j}\right)\left(\hat{F}^{o}_{i,j}-\hat{F}^{i}_{i,j}-F^{o}_{i,j}+F^{i}_{i,j}\right)\leq 0
\end{aligned} \tag{38}$$

because the quantities with hat are the optimal solution to (10). Thus, $\sum_{i,j} y_{i,j}\left(\lambda_{i,j}-\hat{\lambda}_{i,j}\right)\left(\hat{x}_{i,j}+F^{i}_{i,j}-F^{o}_{i,j}\right)\leq 0$. We also have $\sum_{s,j}(x^{s}_{i,j}-\hat{x}^{s}_{i,j})vb^{s}_{i,j}\leq 0$ according to the definition of $b^{s}_{i,j}$.

Therefore, $\dot{W}\leq 0$, which implies that the joint congestion control and scheduling scheme is optimal and convergent according to LaSalle's Invariance Principle [2]. ■

APPENDIX B
PROOF OF THEOREM 2 AND 3

A. Proof of Theorem 2

Proof: The Lyapunov function is the same as defined in Appendix A. Here, for simplicity, it is assumed that $y_{i,j}(X,t)=z^{s}_{i,j}(t)=1$ $(\forall s\in\mathcal{F},(i,j)\in\mathcal{L},\forall X,t)$. Consider the expectation of the one-step Lyapunov drift

$$\begin{aligned}
&\mathbb{E}\left\{W(t+1)-W(t)|X(t),\Lambda(t)\right\}\\
=\ &\mathbb{E}\{W_1(t+1)-W_1(t)|X(t),\Lambda(t)\}+\mathbb{E}\{W_2(t+1)-W_2(t)|X(t),\Lambda(t)\}\\
=\ &\mathbb{E}\{W_1(X(t+1))-W_1(X(t))|X(t),\Lambda(t)\}+\mathbb{E}\{W_2(\Lambda(t+1))-W_2(\Lambda(t))|X(t),\Lambda(t)\}
\end{aligned} \tag{39}$$

By substituting (21) into (39), the first term of (39) is

$$\begin{aligned}
&\mathbb{E}\{W_1(t+1)-W_1(t)|X(t),\Lambda(t)\}\\
=\ &\mathbb{E}\left\{W_1\left(\left\lfloor[\Lambda(t)-\Delta\mathbf{d}]^{+}\right\rfloor\right)-W_1(\Lambda(t))|X(t),\Lambda(t)\right\}\\
=\ &\mathbb{E}\left\{W_1\left([\Lambda(t)-\Delta\mathbf{d}]^{+}-\epsilon_1\right)\middle|\,X(t),\Lambda(t)\right\}-\mathbb{E}\left\{W_1(\Lambda(t))|\,X(t),\Lambda(t)\right\}\\
\leq\ &\mathbb{E}\left\{W_1\left(\Lambda(t)-\Delta\mathbf{d}\right)-W_1(\Lambda(t))|\,X(t),\Lambda(t)\right\}-\\
&\mathbb{E}\left\{\epsilon_1\left[(\Lambda(t)-\Delta\mathbf{d})^{+}-\hat{\Lambda}\right]-\frac{1}{2}\|\epsilon_1\|_2^2\middle|\,X(t),\Lambda(t)\right\}\\
\leq\ &\mathbb{E}\left\{W_1\left(\Lambda(t)-\Delta\mathbf{d}\right)-W_1(\Lambda(t))|\,X(t),\Lambda(t)\right\}+\epsilon_1\hat{\Lambda}+\frac{1}{2}\|\epsilon_1\|_2^2,
\end{aligned} \tag{40}$$

where $\mathbf{d}=[d_{i,j}(t)]$ with $d_{i,j}(t)=F^{o}_{i,j}(t)-F^{i}_{i,j}(t)-x_{i,j}(t)$ as defined in (12) and $\epsilon_1=[\Lambda(t)-\Delta\mathbf{d}]^{+}-\left\lfloor[\Lambda(t)-\Delta\mathbf{d}]^{+}\right\rfloor$; the first inequality comes from expanding the term $W_1\left([\Lambda(t)-\Delta\mathbf{d}]^{+}-\epsilon_1\right)$ and relax the projection in it. Note that $\mathbf{0} \le \epsilon_1 < \Delta^2\mathbf{1}$, where $\mathbf{0}$ and $\mathbf{1}$ are zero and unit matrix respectively. Thus, $\epsilon_1\hat{\Lambda} + \frac{1}{2}\|\epsilon_1\|_2^2 \le \Delta^2\|\hat{\Lambda}\|_2^2 + \frac{1}{2}\Delta^4\|\mathbf{1}\|_2^2 \triangleq \Delta^2 Z_1$. Substitute it into (40), and we have

$$\begin{aligned}
&\mathbb{E}\{W_1(t+1) - W_1(t)|X(t),\Lambda(t)\}\\
\le\ &\mathbb{E}\left\{ W_1\left(\Lambda(t) - \Delta\mathbf{d}\right) - W_1(\Lambda(t))\right| X(t),\Lambda(t)\} + \Delta^2 Z_1\\
=\ &-\Delta\mathbb{E}\left\{\sum_{i,j} d_{\lambda_{i,j}}\left(\lambda_{i,j} - \hat{\lambda_{i,j}}\right)\right| X(t),\Lambda(t)\left.\vphantom{\sum_{i,j}}\right\} + \frac{1}{2}\Delta^2\mathbb{E}\left\{\|d\|_2^2\right| X(t),\Lambda(t)\} + \Delta^2 Z_1
\end{aligned} \tag{41}$$

where the second equality follows from expanding the term $W_1\left(\Lambda(t) - \Delta\mathbf{d}\right) - W_1(\Lambda(t))$.

Denote the objective function in (9) as $V(X(t),\Lambda(t)) = \sum_{s\in S} U^s(x^s) - \sum_{(i,j)\in\mathcal{L}} \lambda_{i,j}x_{i,j}$. The partial derivative w.r.t. $x_{i,j}^s$ is $V_{i,j}^s = u_{i,j}^s - \lambda_{i,j}$ and $\mathbf{g} = [V_{i,j}^s]$, where $u_{i,j}^s$ is the partial derivative w.r.t. $x_{i,j}^s$ of $U^s(x^s)$. Substituting (22) into the second term of (39) and following the same manipulations as above, we have

$$\begin{aligned}
&\mathbb{E}\{W_2(t+1) - W_2(t)|X(t),\Lambda(t)\}\\
=&\mathbb{E}\left\{ W_2\left(\left\lfloor[X(t) + \Delta\left(\mathbf{g} + vB(t)\right)]^+\right\rfloor\right)\right| X(t),\Lambda(t)\} - \mathbb{E}\left\{W_2(X(t))|X(t),\Lambda(t)\right\}\\
=&\mathbb{E}\left\{ W_2\left([X(t) + \Delta\left(\mathbf{g} + vB(t)\right)]^+ - \epsilon_2\right)\right| X(t),\Lambda(t)\} - \mathbb{E}\left\{W_2(X(t))|X(t),\Lambda(t)\right\}\\
\le&\mathbb{E}\left\{ W_2\left(X(t) + \Delta\left(\mathbf{g} + vB(t)\right)\right)\right| X(t),\Lambda(t)\} - \mathbb{E}\left\{ W_2(X(t)) - \frac{1}{2}\|\epsilon_2\|_2^2\right| X(t),\Lambda(t)\left.\vphantom{\frac12}\right\} -\\
&\mathbb{E}\left\{\epsilon_2\left[\left(X(t) + \Delta\left(\mathbf{g} + vB(t)\right)\right)^+ - \hat{X}\right]\right| X(t),\Lambda(t)\left.\vphantom{\hat X}\right\}\\
\le&\mathbb{E}\left\{ W_2\left(X(t) + \Delta\left(\mathbf{g} + vB(t)\right)^+\right)\right| X(t),\Lambda(t)\} - \mathbb{E}\left\{ W_2(X(t))\right| X(t),\Lambda(t)\} + \epsilon_2\hat{\Lambda} + \frac{1}{2}\|\epsilon_2\|_2^2,
\end{aligned} \tag{42}$$

where $B(t) = [b_{i,j}^s(t)]$ and $\epsilon_2 = [X(t) + \Delta\ g]^+ - \left\lfloor[X(t) + \Delta\ g]^+\right\rfloor$. Further simplify (42) by $\epsilon_2\hat{X} + \frac{1}{2}\|\epsilon_2\|_2^2 \le \Delta^2\|\hat{X}\|_2^2 + \frac{1}{2}\Delta^4\|\mathbf{1}\|_2^2 \triangleq \Delta^2 Z_2$ and we get

$$\begin{aligned}
&\mathbb{E}\{W_2(t+1) - W_2(t)|X(t),\Lambda(t)\}\\
\le\ &\mathbb{E}\left\{ W_2\left(X(t) + \Delta\ \left(\mathbf{g} + vB(t)\right)^+\right)\right| X(t),\Lambda(t)\} - \mathbb{E}\left\{ W_2(X(t))\right| X(t),\Lambda(t)\} + \Delta^2 Z_2\\
=\ &\Delta\mathbb{E}\left\{\sum_{s,j} g_{x_{i,j}^s}\left(x_{i,j}^s - \hat{x}_{i,j}^s\right)\right| X(t),\Lambda(t)\left.\vphantom{\sum_{s,j}}\right\} + \Delta\sum_{s,j}\mathbb{E}\left\{ vb_{i,j}^s\left(x_{i,j}^s - \hat{x}_{i,j}^s\right)\right| X(t),\Lambda(t)\}\\
&+\frac{1}{2}\Delta^2\mathbb{E}\left\{\|(\mathbf{g} + vB(t))^+\|_2^2\right| X,\Lambda\} + \Delta^2 Z_2\\
\le\ &\Delta\mathbb{E}\left\{\sum_{s,j} g_{x_{i,j}^s}\left(x_{i,j}^s - \hat{x}_{i,j}^s\right)\right| X(t),\Lambda(t)\left.\vphantom{\sum_{s,j}}\right\} + \frac{1}{2}\Delta^2\mathbb{E}\left\{\|(\mathbf{g} + vB(t))^+\|_2^2\right| X,\Lambda\} + \Delta^2 Z_2
\end{aligned} \tag{43}$$

where the second equality follows the definition of $b^s_{i,j}(t)$. Substitute (41) and (43) into (39), then

$$\begin{aligned} & \mathbb{E}\left\{W(t+1)-W(t)|X(t),\Lambda(t)\right\} \\ \leq\ & -\Delta\mathbb{E}\left\{\sum_{i,j} d(\lambda_{i,j}-\hat{\lambda}_{i,j})\middle| X(t),\Lambda(t)\right\} + \Delta\mathbb{E}\left\{\sum_{s,j} g_{x^s_{i,j}}\left(x^s_{i,j}-\hat{x}^s_{i,j}\right)\middle| X(t),\Lambda(t)\right\} \\ & +\Delta^2(Z_1+Z_2+D_0^2+V_0^2), \end{aligned} \tag{44}$$

where the last inequality follows that the conditional expectation of $\|d\|_2^2$ and $\|(\mathbf{g}+vB(t))^+\|_2^2$ are uniformly bounded by constants $D_0>0$ and $V_0>0$. For $d_{\lambda_{i,j}}(t)$, the per-link flow variables $[F^o_{i,j}(t)]$, $[F^i_{i,j}(t)]$ and the transmission rate $[x_{i,j}(t)]$ are bounded. For $\|(\mathbf{g}+vB(t))^+\|_2^2$, if $\mathbf{g}+vB(t)\geq 0$, $\|(\mathbf{g}+vB(t))^+\|_2^2 \leq \|u^s_{i,j}(t)\|_2^2+\|vB(t)\|_2^2$ is bounded, otherwise $\|(\mathbf{g}+vB(t))^+\|_2^2=0$.

Follow similar manipulation procedure as in the proof of Theorem 1 and take conditional expectation, (44) can be simplified to

$$\begin{aligned} & \mathbb{E}\left\{W(t+1)-W(t)|X(t),\Lambda(t)\right\} \\ \leq\ & \Delta\sum_s \mathbb{E}\left\{(u^s-\hat{u}^s)\left(x^s(t)-\hat{x}^s\right)\right\} + \Delta^2(Z_1+Z_2+D_0^2+V_0^2), \end{aligned} \tag{45}$$

where $\hat{u}^s$ is the partial derivative of $U^s(x^s)$ w.r.t. $x^s_{i,j}$ and evaluated at $\hat{X}$.

Define $\mathcal{A}$, such that

$$\begin{aligned} \mathcal{A}\ \triangleq\ & \left\{X \middle| \mathbb{E}\left[\sum_{i,j}(u^s-\hat{u}^s)\left(\hat{x}^s-x^s(t)\right)\right]\right. \\ \leq\ & \left.\Delta(V_0^2+D_0^2+Z_1+Z_2)\right\}, \end{aligned}$$

which is not empty and has finite number of elements with properly chosen Δ.

Therefore, we have:

$$\begin{aligned} & \mathbb{E}\left\{W(t+1)-W(t)|X(t),\Lambda(t)\right\} \\ \leq\ & \frac{1}{2}\Delta^2(V_0^2+D_0^2+Z_1+Z_2)I_{X\in\mathcal{A}} - \frac{1}{2}\Delta^2(V_0^2+D_0^2+Z_1+Z_2)I_{X\in\mathcal{A}^c} \end{aligned} \tag{46}$$

where I is the indicator function. Thus, by Theorem 3.1 in [8], which is an extension of Foster's criterion, the Markov chain is stable. ■

APPENDIX C
PROOF OF THEOREM 3

Proof: Taking expectation of (46) w.r.t. (X, Λ), we have

$$\begin{aligned}&\mathbb{E}\{W(t+1) - W(t)\} \\ \leq\ & \frac{1}{2}\Delta^2(V_0^2 + D_0^2 + Z_1 + Z_2) + \Delta \sum_s \mathbb{E}\left\{(u^s - \hat{u}^s)(x^s(t) - \hat{x}^s)\right\}\end{aligned} \tag{47}$$

Summing (47) over $\tau = 0, \ldots, t-1$, it is obtained that

$$-\frac{1}{t}\sum_{\tau=0}^{t-1}\left(\sum_s \mathbb{E}\left\{(u^s - \hat{u}^s)(x^s(t) - \hat{x}^s)\right\}\right) \leq \frac{\mathbb{E}\{W(X(0), \Lambda(0))\} + \frac{t}{2}\Delta^2(V_0^2 + D_0^2 + Z_1 + Z_2)}{t\Delta} \tag{48}$$

If $x^s(t) \neq \hat{x}^s$, (48) can be simplified as

$$\begin{aligned}&\frac{1}{t}\sum_{\tau=0}^{t-1}\left(\sum_s \mathbb{E}\left\{\left|\frac{u_{i,j}^s - \hat{u}_{i,j}^s}{x_{i,j}(t) - \hat{x}_{i,j}}\right|(x_{i,j}(t) - \hat{x}_{i,j})^2\right\}\right) \\ \leq & \frac{\mathbb{E}\{W(X(0), \Lambda(0))\} + \frac{t}{2}\Delta^2(V_0^2 + D_0^2 + Z_1 + Z_2)}{t\Delta}.\end{aligned} \tag{49}$$

Since $U^s(x^s)$ is a strictly concave function of x^s and the transmission rate $x^s \in [0, x_m^s]$, $\left|\frac{u_{i,j}^s - \hat{u}_{i,j}^s}{x_{i,j}(t) - \hat{x}_{i,j}}\right|$ is non-negative and bounded. Then $\left|\frac{u_{i,j}^s - \hat{u}_{i,j}^s}{x_{i,j}(t) - \hat{x}_{i,j}}\right| \geq K_1$ holds, where $K_1 > 0$ is the largest lower bound of $\left|\frac{u_{i,j}^s - \hat{u}_{i,j}^s}{x_{i,j}(t) - \hat{x}_{i,j}}\right|$. Letting t going to infinity in (49) produces

$$\sum_s \mathbb{E}\left\{(x^s(\infty) - \hat{x}^s)^2\right\} \leq \frac{\Delta(V_0^2 + D_0^2 + Z_1 + Z_2)}{2K_1}. \tag{50}$$

Then it is obtained that

$$\begin{aligned}&\left[\sum_s \mathbb{E}\left\{U^s\left(x^s(\infty)\right)\right\} - \sum_s \hat{U}^s\right]^2 \\ \leq & \sum_s \left[\mathbb{E}\left\{U^s\left(x^s(\infty)\right)\right\} - \hat{U}^s\right]^2 \\ \leq & \sum_s \mathbb{E}\left\{\left(U^s\left(x^s(\infty)\right) - \hat{U}^s\right)^2\right\} \\ \leq & \sum_s \mathbb{E}\left\{\left|u^s\left(x^s(\infty)\right)\right|^2 (x^s(\infty) - \hat{x}^s)^2\right\} \\ \leq & \frac{K_2|\mathcal{F}|\Delta(V_0^2 + D_0^2 + Z_1 + Z_2)}{2K_1},\end{aligned} \tag{51}$$

where $K_2 = \max_s \left|u^s\left(x^s(\infty)\right)\right|^2$ is a constant; and $|\mathcal{F}|$ is the cardinality of $\mathcal{F}$. ■

REFERENCES

[1] F. Kelly, A. Maulloo, and D. Tan, "Rate control for communication networks: shadow prices, proportional fairness and stability," *Journal of the Operational Research society*, vol. 49, pp. 237–252, Mar. 1998.

[2] S. Shakkottai and R. Srikant, *Network optimization and control*. Boston: Now Publishers, 2008.

[3] X. Lin and N. Shroff, "Joint rate control and scheduling in multihop wireless networks," in *Proc. 43rd IEEE Conf. on Decision and Control*, Paradise Island, 2004, pp. 1484–1489.

[4] Y. Yi and M. Chiang, "Stochastic network utility maximisation-a tribute to Kelly's paper published in this journal a decade ago," *European Trans. Telecommunications*, vol. 19, pp. 421–442, Jun. 2008.

[5] S. Low and D. Lapsley, "Optimization flow control I: basic algorithm and convergence," *IEEE/ACM Trans. Netw.*, vol. 7, pp. 861–874, Dec. 1999.

[6] P. Huang, X. Lin, and C. Wang, "A low-complexity congestion control and scheduling algorithm for multihop wireless networks with order-optimal per-flow delay," in *Proc. 32nd IEEE International Conf. on Computer Communications*, Shanghai, 2011, pp. 2588–2596.

[7] H. Xiong, R. Li, A. Eryilmaz, and E. Ekici, "Delay-aware cross-layer design for network utility maximization in multi-hop networks," *IEEE J. Sel. Areas Commun.*, vol. 29, pp. 951–959, May 2011.

[8] 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, pp. 1936–1948, Dec. 1992.

[9] M. Neely, E. Modiano, and C. Rohrs, "Dynamic power allocation and routing for time-varying wireless networks," *IEEE J. Sel. Areas Commun.*, vol. 23, pp. 89–103, Jan. 2005.

[10] M. Chiang, "Balancing transport and physical layers in wireless multihop networks: Jointly optimal congestion control and power control," *IEEE J. Sel. Areas Commun.*, vol. 23, pp. 104–116, Jan. 2005.

[11] Z. Ding and D. Wu, "Sliding Mode Based Joint Congestion Control and Scheduling in Multi-hop Ad Hoc Networks with Multi-class Services," in *Proc. Global Communications Conf.*, Miami, 2010, pp. 1–5.

[12] L. Bui, R. Srikant, and A. Stolyar, "Novel architectures and algorithms for delay reduction in back-pressure scheduling and routing," in *Proc. 25th IEEE International Conf. on Computer Communications*, Rio de Janeiro, 2009, pp. 2936–2940.

[13] M. Neely, "Energy optimal control for time-varying wireless networks," *IEEE Trans. Inf. Theory*, vol. 52, pp. 2915–2934, Jul. 2006.

[14] M. Chiang, S. Low, A. Calderbank, and J. Doyle, "Layering as optimization decomposition: A mathematical theory of network architectures," *Proc. IEEE*, vol. 95, p. 255, Jan. 2007.

[15] V. Utkin, *Sliding modes in control and optimization*. Berlin: Springer-Verlag, 1992.

[16] C. Lagoa, H. Che, and B. Movsichoff, "Adaptive control algorithms for decentralized optimal traffic engineering in the internet," *IEEE/ACM Trans. Netw.*, vol. 12, pp. 415–428, Jun. 2004.

[17] C. Lagoa and H. Che, "Decentralized optimal traffic engineering in the internet," *ACM SIGCOMM Computer Communication Review*, vol. 30, pp. 39–47, Oct. 2000.

[18] B. Movsichoff, C. Lagoa, and H. Che, "Decentralized optimal traffic engineering in connectionless networks," *IEEE J. Sel. Areas Commun.*, vol. 23, pp. 293–303, Feb. 2005.

[19] M. Neely, "Universal scheduling for networks with arbitrary traffic, channels, and mobility," in *Proc. 49th IEEE Conf. Decision and Control*, Atlanta, 2010, pp. 1822–1829.

[20] L. Chen, S. Low, M. Chiang, and J. Doyle, "Cross-layer congestion control, routing and scheduling design in ad hoc wireless networks," in *Proc. 25th IEEE International Conf. on Computer Communications*, Barcelona, 2006, pp. 1–13.

[21] Y. Yi and S. Shakkottai, "Hop-by-hop congestion control over a wireless multi-hop network," *IEEE/ACM Trans. Netw.*, vol. 15, pp. 133–144, Feb. 2007.

[22] L. Georgiadis, M. Neely, M. Neely, and L. Tassiulas, *Resource allocation and cross layer control in wireless networks*. Boston: Now Pubulisher, 2006.

[23] D. Palomar and M. Chiang, "Alternative distributed algorithms for network utility maximization: Framework and applications," *IEEE Trans. Autom. Control*, vol. 52, pp. 2254–2269, Dec. 2007.

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

[25] Z. Ding and D. Wu, "Capacity Region and Dynamic Control of Per-Next-Hop-Queueing Wireless Networks under Time Varying Channels," http://www.mcn.ece.ufl.edu/public/Zongrui/files/CpctyRgn.pdf.

[26] S. Boyd and L. Vandenberghe, *Convex optimization*. Cambridge: Cambridge university press, 2004.

[27] S. Korovin and V. Utkin, "Using sliding modes in static optimization and nonlinear programming," *Automatica*, vol. 10, pp. 525–532, May 1974.

[28] C. Barnhart, E. Johnson, G. Nemhauser, M. Savelsbergh, and P. Vance, "Branch-and-price: Column generation for solving huge integer programs," *Operations Research*, vol. 46, pp. 316–329, 1998.

[29] X. Lin and N. Shroff, "The impact of imperfect scheduling on cross-layer rate control in wireless networks," in *Proc. 24th IEEE International Conf. on Computer Communications*, 2005, pp. 1804–1814.

[30] X. Wu and R. Srikant, "Regulated maximal matching: A distributed scheduling algorithm for multi-hop wireless networks with node-exclusive spectrum sharing," in *Proc. 44th IEEE Conference on Decision and Control*, Seville, 2005, pp. 5342–5347.

[31] R. Srikant, *The mathematics of Internet congestion control*. Berlin: Birkhauser, 2004.

[32] D. Wu, Y. Hou, and Y. Zhang, "Scalable video coding and transport over broad-band wireless networks," *Proc. IEEE*, vol. 89, pp. 6–20, Jan. 2001.