Stability, Fairness, and Performance: A Flow-Level Study on Nonconvex and Time-Varying Rate Regions

Jiaping Liu, Alexandre Proutière, Yung Yi, *Member, IEEE*, Mung Chiang, *Senior Member, IEEE*, and H. Vincent Poor, *Fellow, IEEE*

***Abstract*—The flow-level stability and performance of data networks with utility-maximizing allocations are studied in this paper. Similarly to prior works on flow-level models, exogenous data arrivals with finite workloads are considered. However, to model many realistic situations, the rate region, which constrains the feasibility of resource allocation, may be either nonconvex or time-varying. When the rate region is fixed but nonconvex, sufficient and necessary conditions are characterized for stability for a class of α-fair allocation policies, which coincide when the set of allocated rate vectors have continuous contours. When the rate region is time-varying according to a Markovian stationary and ergodic process, the precise stability region is obtained. In both cases, the size of the stability region depends on the resource allocation policy, in particular, on the fairness parameter α in α-fair utility maximization. This is in sharp contrast with the substantial existing literature on stability under fixed and convex rate regions, in which the stability region coincides with the rate region for many utility-based resource allocation schemes, independent of the value of the fairness parameter. It is further shown that for networks which consist of flows from two different classes under α-fair allocations, there exists a tradeoff between the stability region and the fairness parameter α. Moreover, the impact of this fairness–stability tradeoff on the system performance, e.g., average throughput and mean flow response time, is studied, and numerical experiments that illustrate the new stability region and the performance versus fairness tradeoff are presented.**

***Index Terms*—Fairness, fluid limit, Markov process, network utility maximization, resource allocation, stability, tradeoff.**

I. Introduction

A. Motivation

FLOWS (or equivalently, end-to-end connections) in networks dynamically share resources (such as link capacities) according to various resource allocation schemes. These flows can be identified through their "classes," which define the set of network resources they require for the transfer of the corresponding packets. A popular family of schemes allocate resources to competing flows by distributively solving a network utility maximization problem [21]. The optimization objective in the form of utility functions can capture important notions such as traffic elasticity, user satisfaction, and fairness. The optimization constraint set captures the feasibility of allocations. In this paper, we focus on the case in which the resource variables are the feasible transmission rates, and the constraint set, referred to as "rate region," can be either the achievable set of rates such as the polytope formed by linear constraints on flows or the information-theoretic capacity region. For other types of resource variables, the analysis of this paper is also applicable if the constraint set can be transformed to an equivalent rate region. For instance, when the resource variables are feasible power levels in the application of power control, the transmission rates are written as functions of signal-to-noise ratios (SNRs), and the constraint set of power levels can be translated into a corresponding rate region, as shown in the example of Fig. 1.

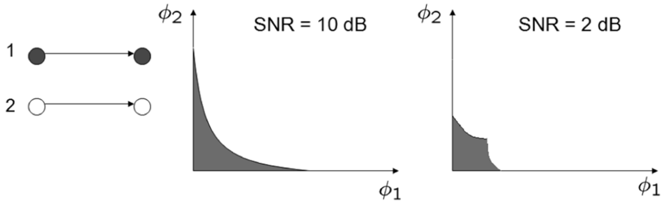


Fig. 1. A two-link interfering wireless network: rate regions when SNR $=10$ and 2 dB.

Extensive work on deterministic models of utility maximization has been conducted since the late 1990s, where flows constitute a static population and are assumed to have infinite backlogs. In practice, the numbers of flows are varying as flows are randomly generated by users and cease upon completion. This system can be viewed as a queueing network where the service rates depend on the solution to an optimization problem, which in turn depends on the number of active flows in each class, thus forming an interesting coupling between stochastic network evolution and distributed optimization algorithm.

A key performance requirement in data networks is that all flows are completed within a finite time, or equivalently, that the numbers of active flows do not grow unbounded. Mathematically, this corresponds to the ergodicity of the process representing the numbers of flows of various classes. This property is referred to as *flow-level stability*. One of the objectives in the design of resource sharing schemes is to provide flow-level

Manuscript received November 21, 2007; revised January 01, 2009. Current version published July 15, 2009. This work was supported in part by the NSF under Grants CNS-0417607, CNS-0625637, CCF-0448012, CCF-0635034, DARPA CBMANET, KRCF, and IR R&D program of MKE/ITTA [2009-F-045-01]. The material in this paper was presented in part at ACM SIGMETRICS, San Diego, CA, July 2007.

J. Liu, M. Chiang, and H. V. Poor are with the School of Electrical Engineering, Princeton University, Princeton, NJ, USA (e-mail: jiapingl@princeton.edu; chiangm@princeton.edu; poor@princeton.edu).

A. Proutière is with Microsoft Research, CB3 0FB Cambridge, U.K. (e-mail: alexandre.proutiere@microsoft.com).

Y. Yi is with the Department of Electrical Engineering and Computer Science, KAIST, Daejeon, Korea (e-mail: yiyung@ee.kaist.ac.kr).

Communicated by S. Ulukus, Associate Editor for Communication Networks.

Color versions of Figures 6 and 12 in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIT.2009.2023680

0018-9448/$25.00 © 2009 IEEE

stability whenever possible, or to maximize the (flow-level) stability region, defined as the set of vectors representing the traffic intensities of the various flow classes such that the network is stable at flow level. It is worth differentiating the notion of stability region studied here to Shannon theoretic notion of capacity region, which refers to the largest set of "achievable" rates for a fixed population of users with an infinite backlog of messages, and achievability is defined based on vanishing probability of decoding errors. The capacity region of a network, when characterized through achievability and converse theorems, can be used as one of the models for the constraint set of the utility maximization, i.e., the rate region.

As will be briefly reviewed later in this section, a series of papers in the literature have provided necessary and sufficient conditions for flow-level stability in various models. With a few recent exceptions, these models assume a fixed and convex rate region. In this paper, we investigate conditions for flow-level stability when the rate region is either nonconvex or time-varying.

Indeed, as explained in more detail in Section II-B and also in the survey in [11], in many applications we cannot assume convexity or time invariance of the rate region in Network Utility Maximization models when studying flow-level stability. For example, nonconvexity of the rate region naturally arises in wireless cellular and *ad hoc* networks [3], [19]. Fig. 1 shows a simple network consisting of two interfering links 1 and 2 whose capacities are shared by two classes of flows (class-i flows use link i only for $i = 1, 2$). The transmitters of both links are always active, but the transmission power may be adapted to the population of flows of the different classes. When interference at both receivers is treated as noise, the feasible transmission rates of links 1 and 2 are given by

$$\phi_1 \leq W \log_2 \left(1 + \frac{u_1 P}{N + u_2 P}\right)$$
$$\phi_2 \leq W \log_2 \left(1 + \frac{u_2 P}{N + u_1 P}\right)$$

where W denotes the bandwidth, P is the maximum power received at both receivers, N is the noise power, and $u_1, u_2 \in (0,1)$. The corresponding rate region is shown in Fig. 1 at SNR levels 2 and 10 dB, respectively $(\mathrm{SNR} = P/N)$.

On the other hand, time variation of rate region is common in practical systems due to mobility, link failures, route or topology changes, and priority structures in resource allocation. It turns out that new proof techniques are needed to prove stability conditions in these scenarios, and intriguing tradeoffs between fairness and stability are discovered.

B. Related Work

The first analysis of the flow-level stability focused on wired networks supporting data traffic only [5], [14]. For such networks, the rate region is a (convex) polytope formed by the intersection of a finite number of linear capacity constraints, and it has been shown that all α-fair allocations with $\alpha > 0$ provide flow-level stability if and only if the vector representing the average traffic intensities of flow classes lies in the rate region. In other words, the rate region in the utility maximization problem is also the stability region under flow-level stochastic dynamics. This result has been generalized by many papers, e.g., [32], [33], in particular, to the case of networks with arbitrary convex rate regions [6], to the case without assuming time scale separation [23], and recently to the case of general flow arrival processes and general flow size distributions [10], [12], [17], [22], [26], [33]. It has been shown in [6] that if the traffic intensity vector is outside of the rate region, then there is no allocation stabilizing the network at flow level. These results imply that for fixed, convex rate regions, α-fair allocations maximize the flow-level stability region. This is sometimes called the throughput-optimality property for the utility-maximization-based resource allocations.

The analysis of flow-level stability in the case of fixed but *nonconvex* rate region is generally very difficult, and has been investigated in very few existing works. In cases of networks with two flow classes only, the stability condition of a large class of allocations can be exactly characterized [7]. However, when the number of classes is greater than two, it has been has been found to be extremely difficult to derive an explicit and exact stability condition. This is mainly due to the fact that the stability condition depends on detailed statistical characteristics of the flow arrival processes, and flow departure processes, which are determined by the solutions of nonconvex optimization problems. Some papers provide bounds on the stability region for specific networks under particular allocations, see, e.g., [3], [25]. These papers study the stability of networks where the rate region reduces to a single point depending on the set of classes with active flows. Some other papers aim at providing exact stability conditions: in [9], [20], [29], a recursive (with respect to the number of flow classes) stability condition is given for a particular class of networks, including those studied in [3], [25]. Unfortunately, this kind of recursive formula often proves difficult to exploit: the stability condition of networks with S classes of flows depends on that of the network with $S - 1$ classes and also on more detailed characterizations such as the probability that a given class has no active flows. Usually, these characterizations cannot be efficiently computed.

The analysis of the flow-level stability of networks with time-varying rate regions has not been extensively studied so far. To the best of our knowledge, the only existing results provide the flow-level stability of wireless networks with user mobility under certain α-fair allocations [4], [8]. In [23], the authors show that one can obtain the largest possible stability region at flow level when applying opportunistic resource-sharing algorithms that explicitly take advantage of rate-region variations. In contrast, we investigate the case where one cannot apply such an approach, because rate regions may be slowly varying and implementing opportunistic resource sharing could result in serious fairness issues. A detailed discussion on the various time scales of the system dynamics is presented in Section II. Under a different system model, [15] investigated the stability of packet-level dynamics with a stochastic channel model and opportunistic scheduling.

As will be shown in Section V, there are interesting tradeoffs between fairness and flow-level stability when the rate region is nonconvex or time-varying. This tradeoff is different from that between fairness and efficiency investigated for a static population of flows with infinite backlogs (see, e.g., [27], [28] in wired networks, or [16], [24] for wireless networks, and [30] for a discussion on the absence of this tradeoff in a general topology). Here, we investigate the tradeoff between fairness and stochastic stability region, which quantifies the impact of fairness on the performance as perceived by users in a dynamic population of flows.

Moreover, although an exact characterization of performance metrics such as average throughput and mean response time (as defined in Section II-F) of such systems has shown to be complicated to obtain [5], a variety of numerical examples are provided for performance evaluation [5], [6], where the numerical experience suggests that the fairness parameter has a greater impact on performance in wireless networks than in wired networks. We also study toy network examples in both wired and wireless cases to observe the impact of the fairness–stability tradeoff on the system performance. Similarly, as illustrated in Section VI, it is observed that wireless networks tend to be more sensitive to the fairness parameter since the rate regions may have sharper variations over time compared to wired networks.

C. Overview

In this paper, we provide general stability conditions of α-fair allocations in networks with nonconvex or time-varying rate regions. The main results are the following.

(i) In networks with arbitrary numbers of classes and with fixed but nonconvex rate regions, we give sufficient and necessary conditions for flow-level stability of α-fair allocations, for all $\alpha > 0$. We also prove that these conditions coincide when the set of allocated rate vectors is continuous (in a sense that will be defined at the end of this section), leading to an explicit stability condition for such networks (Theorems 4, 5, and Corollary 1).

(ii) We extend our analysis to networks with time-varying rate regions, for which we also provide the stability conditions of α-fair allocations, for all $\alpha > 0$ (Theorems 7). The results and proof techniques in (i) and (ii) can be readily combined for the general case of any time-varying rate region which is either convex or nonconvex at any fixed time instant.

(iii) When the rate region is either nonconvex or time-varying, the stability condition is proven to depend on the chosen fairness parameter α. The exact degree of sensitivity with respect to α depends on the considered network, which can be significant (possibly changing the shape of stability region from concave to convex) or negligible. We provide examples for both situations. In two-class networks, we also prove that, as α increases, the flow-level stability region shrinks (Corollary 3). In other words, there is a tradeoff between fairness and flow-level performance. Fairness can be enhanced but at the expense of reduced network stability. This is in sharp contrast to the case of fixed and convex rate regions, where fairness has no impact on stability. This new phenomenon shows that the choice of the utility function is crucial to ensure a high user-level performance under nonconvex or time-varying rate regions.

The paper is organized as follows. Section II is devoted to describing the system model and presenting the assumptions. In Sections III and IV, we provide the stability conditions for nonconvex and time-varying rate regions, respectively. We discuss the tradeoff between fairness and stability in Section V. We illustrate our theoretical results with examples from both wired and wireless networks in Section VI, and conclude the paper in Section VII. Proofs that are essential to the flow of the paper are presented right after the theorems, and other proofs are collected in the Appendix.

Notation and Definitions: We summarize the major definitions and notations used throughout the paper.

- For all $\boldsymbol{A}$, $\boldsymbol{B}$ in $\mathbb{R}^S$, $\boldsymbol{A} \le \boldsymbol{B}$ (resp., $\boldsymbol{A} < \boldsymbol{B}$) means that $\boldsymbol{A}$ is component-wise less (resp., strictly less) than $\boldsymbol{B}$.
- A set $\mathcal{Y} \subset \mathbb{R}_+^S$ is *coordinate–convex* when the following is true: if $\boldsymbol{B} \in \mathcal{Y}$, then for all $\boldsymbol{A}$: $0 \le \boldsymbol{A} \le \boldsymbol{B}$, $\boldsymbol{A} \in \mathcal{Y}$.
- A set $\mathcal{Y} \subset \mathbb{R}_+^S$ is a *Pareto-type* set if for any couple $A, B \in \mathcal{Y}$, $A \le B$ implies that $A = B$.
- $\mathcal{Y}_\circ$ denotes the largest open subset of $\mathcal{Y}$.
- $c(\mathcal{Y})$ denotes the smallest closed set containing $\mathcal{Y}$.
- Define $\mathcal{U} = \left\{x \in \mathbb{R}_+^S : \sum_s x_s = 1\right\}$ and $\mathcal{D} : \mathbb{R}_+^S \mapsto \mathcal{U}$ the application giving the direction of vectors, i.e., $\mathcal{D}(v) = v/|v|$, where $|v| = \sum_s v_s$. We say that a Pareto-type set $\mathcal{Y}$ is continuous in direction u if the two following conditions are satisfied: (i) there exists $\epsilon > 0$ such that $\{v \in \mathcal{U} : |v - u| < \epsilon\} \subset \mathcal{D}(c(\mathcal{Y}))$; (ii) the application $\mathcal{D}^{-1} : \mathcal{U} \mapsto \mathcal{Y}$ is continuous at u. Condition (i) means that there are vectors in $\mathcal{Y}$ in all directions around u. Note that $\mathcal{D}^{-1}$ is well defined since $\mathcal{Y}$ is a Pareto-type set.
- A Pareto-type set $\mathcal{Y} \in \mathbb{R}_+^S$ is said to be continuous if $\forall u \in \mathcal{D}(c(\mathcal{Y}))$, $\mathcal{Y}$ is continuous in direction u.

II. System Model

A. Traffic Demand and Network State

We consider a data network where flows are randomly generated by users and cease upon completion. Flows are classified according to the set of resources required to transfer the corresponding packets. For example, in wired networks with fixed routing, the class of a flow is defined by the set of links that the flow traverses from the source to the destination. We have a finite set $\mathcal{S}$ of S classes of flows. Flows of class s are generated according to a Poisson process of intensity λ_s flows per second. The sizes of class-s flows are independent and identically distributed (i.i.d.) exponentially distributed with mean size $1/\mu_s$ bits. We define the traffic intensity/offered load of flows of class s by $\rho_s = \lambda_s/\mu_s$ bits per second. More general flow arrival processes are considered in [22], [26], [33], and more general file size distributions are considered in [12], [17], [26], respectively, but all for convex and fixed-rate region. In this paper, we instead investigate the stability region after the restrictive assumption of convexity and time invariance of the rate region is removed, while maintaining the Markovity assumption on traffic.

At time t, the network state is denoted by $\boldsymbol{N}(t) = (N_1(t), \ldots, N_S(t))$ where $N_s(t)$ is the number of active class-s flows. $\{\boldsymbol{N}(t)\}_{t=0}^{\infty}$ is a stochastic process governed by the random arrivals and departures of flows.

B. Rate Region

The rate region $\mathcal{R}$ of the network is defined as the set of achievable rate vectors $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_S)$ where ϕ_s is the total rate allocated to class-s flows. A rate vector $\boldsymbol{\phi}$ is said to be achievable if there exist resource allocation mechanisms that can realize this vector. We assume here that the rate region does not depend on the network state $\boldsymbol{N}$. For example, consider a wired network with two links of respective capacities C_1 and C_2. Two flow classes compete for the use of these resources, class-1 flows require the use of both links whereas class-2 flows require that of the second link only. The corresponding rate region is then $\mathcal{R} = \{\boldsymbol{\phi} : \phi_1 + \phi_2 \leq C_2, \phi_1 \leq C_1\}$.

As illustrated in the previous example, the rate regions of wired networks are often convex and coordinate-convex sets. This is also the case for some wireless systems, mainly when a centralized resource allocation is permitted and a time-sharing argument *convexifies* the rate region. See, e.g., [6] for many other examples of networks with convex rate regions. However, there are many situations where the rate region loses its convexity, for example, due to distributed resource allocation in wireless networks, or due to the fact that the achievable set of capacities is discrete. In cellular networks, the fact that the transmissions of the various base stations are not coordinated leads to nonconvex rate regions [3]. In particular, when the achievable transmission power levels of a base station form a countable set, the rate region is discrete [7]. In wireless local-area (LANs), mesh, or *ad hoc* networks, users or nodes randomly access the radio channel in a distributed manner, which again induces nonconvexity [19]. See [7] and Section VI of the present paper for the example on distributed medium-access control (MAC) scheduling. The first focus of this paper is to analyze the performance of networks with nonconvex rate region. In Section III, we do not make any assumption on the rate region except that it is a compact subset of $\mathbb{R}_+^S$.

The second focus of this paper is to study networks with time-varying capacities according to some exogenous processes (independent of the evolution of the network state). For example, in wired multiservice networks supporting low-priority data traffic and high-priority real-time traffic, the available capacity for data traffic is what is left by real-time traffic. The variations can also stem from link failures or from routing table changes. In wireless systems, fading as well as user mobility (in cellular networks) or node mobility (in *ad hoc* networks) also generate capacity variations. Here we denote by $\mathcal{R}(t)$ the rate region at time t. We assume that the set $\mathcal{I}$ of indices of possible states $\{\mathcal{R}_i\}$ for the process $\{\mathcal{R}(t)\}_{t=0}^{\infty}$ is finite, and that $\{\mathcal{R}(t)\}_{t=0}^{\infty}$ is a stationary and ergodic Markov process. We denote by $\boldsymbol{\pi}$ the stationary distribution of $\{\mathcal{R}(t)\}$, i.e., $\mathbb{P}\{\mathcal{R}(t) = \mathcal{R}_i\} = \pi_i$, $i \in \mathcal{I}$. By convention, each possible rate region $\mathcal{R}_i$ is a compact subset of $\mathbb{R}_+^S$.

C. Resource Allocation Algorithms

Resource allocation algorithms allocate network resources to different flow classes according to the current network state $\boldsymbol{N}(t)$ and the current rate region $\mathcal{R}(t)$. Since the seminal work of Kelly *et al.* [21], optimization approaches have been extensively used to model and design the way these algorithms share the network resources. Most existing resource allocations aim at maximizing a certain notion of *utility* of the network. The realized allocation is then the solution of the following optimization problem:

$$\begin{aligned} \text{maximize} \quad & \sum_s N_s(t) U_s \left(\frac{\phi_s}{N_s(t)} \right) \\ \text{subject to} \quad & \boldsymbol{\phi} \in \mathcal{R}(t) \end{aligned} \tag{1}$$

where the utility functions U_s are usually assumed to be concave and nondecreasing. Here we also assume that all flow classes share the same utility function, i.e., $U_s = U$ for all s.

A large class of resource allocations are obtained based on the utility functions $U^\alpha(\cdot) = (\cdot)^{1-\alpha}/(1-\alpha)$ for $\alpha > 0$, and $\log(\cdot)$, for $\alpha = 1$ [28]. The parameter α represents the degree of fairness of the allocation: when $\alpha = 0$, the total throughput of the network is maximized but the allocation may lead to user starvation and thus will not be considered in this paper; $\alpha = 1$ gives the Proportional Fair allocation; when $\alpha \to \infty$, it corresponds to the max-min fairness.

We denote the optimal solution of (1) at time t by $\boldsymbol{\phi}(\boldsymbol{N}(t))$ or $\boldsymbol{\phi}(\boldsymbol{N}(t), \mathcal{R}(t))$. For time-varying rate regions, this solution is denoted by $\boldsymbol{\phi}^{(i)}(\boldsymbol{N}(t))$ if $\mathcal{R}(t) = \mathcal{R}_i$. Since $\mathcal{R}(t)$ is compact, a solution of (1) exists. However, for the nonconvex rate region, the solution is not necessarily unique.

Note that we could replace $\boldsymbol{N}(t)$ in (1) by any vector $\boldsymbol{N}$ in $\mathbb{R}_+^S$ (as will be shown later, this is used to evaluate the system dynamics in the fluid limit regime). We then denote by $\boldsymbol{\phi}(\boldsymbol{N}, \mathcal{R}(t))$ or $\boldsymbol{\phi}(\boldsymbol{N})$ the solution of optimization problem. For any compact rate region $\mathcal{R}(t)$, the solution $\boldsymbol{\phi}(\boldsymbol{N})$ corresponding to any α-fair allocation is unique and has the following properties.

Property 1 (Continuity): For any fixed $\boldsymbol{N}^o \in \mathbb{R}_+^S$, let $\{\boldsymbol{N}^{(k)}\}_{k=1}^{\infty}$ be a sequence of states such that $\lim_{k\to\infty} \boldsymbol{N}^{(k)} = \boldsymbol{N}^o$, then $\boldsymbol{\phi}(\boldsymbol{N}^{(k)}) \to \boldsymbol{\phi}(\boldsymbol{N}^o)$ as $k \to \infty$.

Property 2 (Homogeneity): For any $\boldsymbol{N}$ and any scalar $a > 0$, $\boldsymbol{\phi}(a\boldsymbol{N}) = \boldsymbol{\phi}(\boldsymbol{N})$.

Property 3 (Pareto Efficiency): The set $\{\boldsymbol{\phi}(\boldsymbol{N}), \forall \boldsymbol{N} \in \mathbb{R}_+^S\}$ is a Pareto-type set.

The proof of Property 1 is provided in [33]; Property 2 can be easily conducted by the expression of α-fair utilities; Property 3 is due to the fact that any α-fair allocation with a compact rate region is Pareto efficient.

D. Time Scale Assumptions

The global system dynamics are induced by the flow arrivals/departures, the possible variations of the rate region, and the packet-level dynamics of the underlying resource allocation algorithms. The different time scales of these sources of system dynamics play an important role in the performance analysis, denoted as follows:

(i) $\mathbb{T}_1$: the time scale of the flow-level dynamics;
(ii) $\mathbb{T}_2$: the time scale of the rate region variations;
(iii) $\mathbb{T}_3$: the time scale of resource allocation algorithm's convergence.

We assume that the time scale of flow-level dynamics is much larger than that of resource allocation algorithms, i.e., $\mathbb{T}_1 \gg \mathbb{T}_3$. When the network state changes, the resource allocation algorithms are assumed to converge instantaneously to adapt the realized rate vector to this change. This assumption is often referred to as the *time-scale separation* assumption in the literature.

When the time scales of rate region variations and of the resource allocation algorithms are similar, i.e., $\mathbb{T}_2 \approx \mathbb{T}_3$, these algorithms can directly take advantage of the rate region variations. Such systems are said to be *opportunistic*. A typical example of such systems is channel-aware scheduling in cellular networks [2], [23], where fast-fading variations of the channels are exploited to get a greater throughput. When the rate region variations are not that fast, i.e., $\mathbb{T}_2 \gg \mathbb{T}_3$, being opportunistic proves more difficult and these variations can be exploited only at the expense of compromising the delay allowance of users. In this paper, the rate region variations are assumed to be relatively slow, as they can be generated by phenomena such as node mobility in wireless networks and link failures in wired networks.

To summarize, we assume that $\mathbb{T}_1, \mathbb{T}_2 \gg \mathbb{T}_3$, which means that the resource allocation algorithms instantaneously adapt the rate vector to either the numbers of active flows of various classes or the rate region variations. No assumption is made on the relative time scales $\mathbb{T}_1$ and $\mathbb{T}_2$.

E. Flow-Level Stability

One of the main focuses in this paper is to show necessary and sufficient conditions for flow-level stability: when will the durations of flows remain finite (almost surely)? Mathematically, stability means that the process $\{\boldsymbol{N}(t)\}_{t=0}^{\infty}$ is ergodic. With the assumptions in Section II-D, this process is Markovian and evolves as follows: for each class s

$$\begin{aligned}&N_s(t) \to N_s(t)+1, \text{ with rate } \lambda_s\\&N_s(t) \to N_s(t)-1, \text{ with rate } \mu_s\phi_s(\boldsymbol{N}(t), \mathcal{R}(t)).\end{aligned}$$

The flow-level stability is now equivalent to the positive recurrence of the Markov process $\boldsymbol{N}(t)$, which implies the *almost sure* finiteness of the number of active flows in the system, i.e., flows that are being served or remain in the queues. In the following, we characterize the set of traffic intensity vectors $\boldsymbol{\rho} = (\rho_1, \ldots, \rho_S)$ such that flow-level stability can be realized. This set is referred to as the *stability region*, which also depends on the considered resource allocation algorithm. We say a compact set Γ is the stability region under certain resource allocation, if $\forall\, \boldsymbol{\rho} \in \Gamma_o$ such that the system is stable, and if $\forall\, \boldsymbol{\rho} \notin \Gamma$, the system is unstable.

On a related but different notion, the maximum stability region is defined by the union of all possible stability regions under all possible resource allocations, i.e., for any traffic intensity vector outside this set, there exists no resource allocation algorithm that can stabilize the network at flow level. Note that such resource allocation may not be utility-based or implementable in a distributed fashion.

F. Performance Metrics

Besides the fundamental stability requirement, we are also interested in characterizing the performance of the system in our model. We introduce a series of performance metrics to evaluate the quality-of-service (QoS) level in different applications which will be studied in various numerical examples in Section VI.

Average flow throughput: the average throughput of class s flow is defined by

$$\gamma_s = \frac{\rho_s}{E[N_s]}. \tag{2}$$

Similarly, the average throughput of the system over all flow classes is considered as

$$\gamma = \frac{\sum\limits_{s\in\mathcal{S}} \rho_s}{\sum\limits_{s\in\mathcal{S}} E[N_s]}. \tag{3}$$

Mean flow response time: the mean flow response time of class s is considered as the average duration time of a class s flow in the network. By Little's law, it is given by

$$T_s = \frac{E[N_s]}{\lambda_s} = \frac{1}{\gamma_s\mu_s} \tag{4}$$

which is equivalent to the throughput metric.

Conditional mean response time: let x_s denote the file size of a class s flow, then we can define the conditional mean response time (conditioned on the file size) as the mean response time of this "tagged" flow, denoted by $T_s(x_s)$.

Standard deviation of response time: we also study the second-order statistical characteristics of flow response times, i.e., we will examine the standard deviation of flow response time $\sigma(T_s)$, and the conditional standard deviation, denoted by $\sigma(T_s(x_s))$.

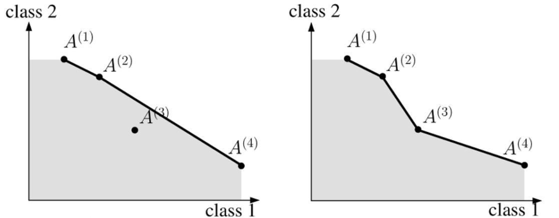
III. Stability With Arbitrary Fixed Rate Region

In this section, we investigate the flow-level stability of α-fair allocations for networks with arbitrary, but fixed rate region. We first recall the stability result for convex, coordinate–convex rate region $\mathcal{R}$, see, e.g., [6], [23].

Theorem 1 (Convex Rate Regions): For any convex, coordinate-convex rate region, the maximum stability region is the rate region, and is achieved by all α-fair allocations, provided that $\alpha > 0$.

The above theorem states that α-fair allocations are optimal with respect to (w.r.t.) the flow-level stability. In particular, they all have the same stability region. Hence, for convex, coordinate-convex rate region, fairness is not imposed at the expense of a reduction of the stability region. We now investigate the case where the rate region is not convex, in which case the stability region may strongly depend on the fairness parameter α. We begin by recalling the result providing the maximum stability region in case of arbitrary rate region [7]. We then study the case of discrete rate regions, i.e., rate regions composed by a finite number of rate vectors, and conclude this section with an analysis on the stability for arbitrary *continuous* rate regions. The maximum stability region is given in the following theorem [7].

Theorem 2 (Maximum Stability Region—Arbitrary Rate Region [1], [7]): For a network with an arbitrary rate region $\mathcal{R}$,



(a) $\alpha = 0.5$ (b) $\alpha = 1$

Fig. 2. Different chosen rate vectors, contours, and stability regions of a two-class network, for $\alpha = 0.5$ and $\alpha = 1$, with $A^{(1)} = (1,4)$, $A^{(2)} = (2,3.5)$, $A^{(3)} = (3,2)$, and $A^{(4)} = (6,1)$.

the maximum stability region is the smallest convex, coordinate-convex set containing $\mathcal{R}$.

In particular, it has been proven in [7] that the so-called Max-Projection (MP) allocation introduced in [1] achieves the maximum stability. For a given network state $\boldsymbol{N}$, the MP allocation allocates rates that form the solution of the following problem:

$$\max \sum_{s \in \mathcal{S}} N_s \phi_s, \quad \text{subject to } \boldsymbol{\phi} \in \mathcal{R}. \tag{5}$$

It is worth noting that this allocation is not utility-based, and there is no existing distributed implementation of this kind of allocation.

A. Discrete Rate Region

In the case of arbitrary discrete rate regions, the stability condition of α-fair allocations turns out to be sensitive to detailed traffic demand characteristics, such as the flow size distribution, see, e.g., [3], [25]. This explains why deriving an exact expression for the stability region proves quite challenging. However, for networks with two flow classes only, the stability region is known and given by Theorem 3 [7]. In the rest of this paper, we denote by $\mathcal{R}^\alpha$ the set of rate vectors actually chosen by the α-fair allocation, i.e., the set of vectors $A \in \mathcal{R}$ such that there exists a state $\boldsymbol{N}$ with $\boldsymbol{\phi}(\boldsymbol{N}) = A$ for this allocation. Also for notational convenience, we prove the results of this section for α-fair allocations with $\alpha \neq 1$. They can be similarly proved for Proportional fair allocation given by $\alpha = 1$.

Theorem 3 (Discrete Rate Region—Two Classes [7]): The stability region of an α-fair allocation, for $\alpha > 0$, is the smallest coordinate convex set containing the contour of $\mathcal{R}^\alpha$.

Here for a two-class network, the contour of $\mathcal{R}^\alpha \in \mathbb{R}_+^2$ is defined as the broken line joining the allocated rate vectors from left to right. In general, $\mathcal{R}^\alpha$ depends on the allocation considered, which in turn leads to the dependence of the stability region on α. In Fig. 2, we present an example of a two-class network with discrete rate region $\mathcal{R} = \{A^{(1)}, A^{(2)}, A^{(3)}, A^{(4)}\}$, and illustrate the dependence of the stability region on α. When $\alpha = 0.5$, $\mathcal{R}^\alpha = \{A^{(1)}, A^{(2)}, A^{(4)}\}$, and when $\alpha = 1$, $\mathcal{R}^\alpha = \mathcal{R}$. As a consequence, the Proportional fair allocation achieves a smaller stability region than the 0.5-fair allocation.

We generalize the result of Theorem 3 to the case of networks with an arbitrary number of flow classes. As explained above, deriving an exact expression for the stability region proves generally impossible. Hence, we separately derive sufficient and necessary conditions for stability. Later, we will show that the gap between the sufficient and necessary conditions vanishes as the set of rate vectors chosen by the considered allocation gets continuous.

Consider the α-fair allocation in a network with a fixed discrete rate region $\mathcal{R} = \{A^{(1)}, \ldots, A^{(K)}\}$. We use fluid limits [13] to investigate stability, see the Appendix for more details. We denote by $\boldsymbol{n}$ the network state in the fluid limit with a continuous state space. Since the rate region $\mathcal{R}$ is bounded, then following the same argument in [5], [33], the fluid limit $\boldsymbol{n}$ of the original dynamics $\boldsymbol{N}$ is deterministic and continuous function of time t, and it is differentiable almost everywhere except at the intersections of different cones

$$\frac{dn_s}{dt} = \begin{cases} \mu_s \left(\rho_s - A_s^{(l)}\right), & \text{if } n_s \neq 0 \\ \max\left(\mu_s \left(\rho_s - A_s^{(l)}\right), 0\right), & \text{if } n_s = 0 \end{cases} \tag{6}$$

for all $s \in \mathcal{S}$ and when the rate vector $A^{(l)}$ is allocated. The fluid limit is stable if $\boldsymbol{n}(t)$ reaches and stays at 0 within finite time. If starting from any initial point, the fluid limit reaches 0 in finite time, then the initial process $\{\boldsymbol{N}(t)\}_{t=0}^{\infty}$ is ergodic. The fluid limit is said to be unstable if $\|\boldsymbol{n}(t)\|$ grows at least linearly (after a finite time), and this instability implies the transience of the process $\{\boldsymbol{N}(t)\}_{t=0}^{\infty}$.

Define the subset $\mathcal{C}^{(j)}$ of the state space $\mathbb{R}_+^S$ (in the fluid limit) where the α-fair allocation allocates the rate vector $A^{(j)}$

$$\mathcal{C}^{(j)} = \left\{\boldsymbol{n} : \boldsymbol{\phi}(\boldsymbol{n}) = A^{(j)}\right\}. \tag{7}$$

Note that each $\mathcal{C}^{(j)}$ is a cone due to the Property 2 (i.e., homogeneity) of the α-fair allocation. Some cones may be empty, in which case the corresponding rate vector is never allocated by the α-fair allocation. The cones defined in (7) satisfy:

(i) $\bigcup_{1 \leq j \leq K} \mathcal{C}^{(j)} = \mathbb{R}_+^S$;
(ii) $\mathcal{C}_o^{(j)} \bigcap \mathcal{C}_o^{(j')} = \emptyset$, for all $j \neq j'$;
(iii) the rate vector $A^{(j)}$ is allocated, if $\boldsymbol{n} \in \mathcal{C}_o^{(j)}$;
(iv) if $\boldsymbol{n} \in \mathcal{C}^{(j)} \bigcap \mathcal{C}^{(j')}$, then either $A^{(j)}$ or $A^{(j')}$ is allocated.

Notice that property (iv) can be interpreted that $\mathcal{C}^{(i)} \bigcap \mathcal{C}^{(j)}$ is a subset of $\mathbb{R}_+^S$ with zero measure, thus the allocation when $\boldsymbol{n} \in \mathcal{C}^{(i)} \bigcap \mathcal{C}^{(j)}$ does not affect the evolution of $\boldsymbol{n}(t)$.

Now after defining the cone allocation, we will see in the following theorem that the stability condition depends on the comparison of traffic load and service rate (allocated rate vector).

Theorem 4 (Discrete Rate Region—Sufficient Stability Condition): For the discrete rate region $\mathcal{R} = \{A^{(1)}, A^{(2)}, \ldots, A^{(K)}\}$, if $\mathcal{R}^\alpha$ is the set of allocated rate vectors under α-fair allocation, then the stability region of the α-fair allocation, for $\alpha > 0$, contains Λ^α, the smallest coordinate convex set containing $\mathcal{R}^\alpha$.

Proof: Let $\boldsymbol{\rho} \in \Lambda^\alpha$; then there exists at least one point $A^{(l)}$ such that $\boldsymbol{\rho} < A^{(l)}$. Define

$$V_j(\boldsymbol{n}) = \sum_{s \in \mathcal{S}} n_s^\alpha \frac{(A_s^{(j)})^{1-\alpha} - (\rho_s)^{1-\alpha}}{1-\alpha}. \tag{8}$$

Then at any time t, if the fluid limit is at state $\boldsymbol{n}$ and the allocated rate vector $\boldsymbol{\phi}(\boldsymbol{n}) = A^{(j)}$, we have $V_j(\boldsymbol{n}) \geq 0$ since

$$\sum_{s \in \mathcal{S}} n_s^\alpha \frac{\left(A_s^{(j)}\right)^{1-\alpha}}{1-\alpha} \geq \sum_{s \in \mathcal{S}} n_s^\alpha \frac{\left(A_s^{(l)}\right)^{1-\alpha}}{1-\alpha} > \sum_{s \in \mathcal{S}} n_s^\alpha \frac{(\rho_s)^{1-\alpha}}{1-\alpha}.$$

Now we introduce the function

$$L(\boldsymbol{n}(t)) = \sum_{1\le j\le K} V_j(\boldsymbol{n}(t))\mathbf{1}_{\{\boldsymbol{n}(t)\in\mathcal{C}^{(j)}\}} \qquad (9)$$

which is continuous and differentiable almost everywhere (except at times where $\boldsymbol{n}(t)$ is at the intersections of cones $\mathcal{C}^{(j)}$). Moreover, $L(\boldsymbol{n}) \ge 0$ for all $\boldsymbol{n} \in \mathcal{R}_+^S$ and the equality holds if and only if $\boldsymbol{n} = 0$.

Assume that at time t, $L(\boldsymbol{n}(t)) > 0$ with $\boldsymbol{n}(t) \in \mathcal{C}^{(j)}$, and that it is differentiable almost everywhere. Then we have

$$\frac{dL}{dt} = \sum_{1\le j\le K} \mathbf{1}_{\{\boldsymbol{n}(t)\in\mathcal{C}^{(j)}\}} \sum_{s\in\mathcal{S}} \alpha n_s^{\alpha-1}\mu_s \left(\rho_s - A_s^{(j)}\right) \frac{\left(A_s^{(j)}\right)^{1-\alpha} - (\rho_s)^{1-\alpha}}{1-\alpha} < 0. \qquad (10)$$

Now we show that $L(\boldsymbol{n})$ decreases to 0 within finite time. We divide $\mathcal{S}$ as $\mathcal{S}_a(j)\bigcup\mathcal{S}_b(j)$ when $A^{(j)}$ is allocated and $\mathcal{S}_a(j) = \{s : \rho_s \ge A_s^{(j)}\}$, $\mathcal{S}_b(j) = \{s : \rho_s < A_s^{(j)}\}$. Since $\boldsymbol{\rho} \in \Lambda^\alpha$, then $\mathcal{S}_b(j)$ is nonempty for all j. Thus, we may rewrite $L(\boldsymbol{n}(t))$ as $L_a(\boldsymbol{n}(t)) + L_b(\boldsymbol{n}(t))$, where

$$L_a(\boldsymbol{n}) = \sum_{1\le j\le K} \mathbf{1}_{\{\boldsymbol{n}(t)\in\mathcal{C}^{(j)}\}} \sum_{s\in\mathcal{S}_a(j)} n_s^\alpha \frac{\left(A_s^{(j)}\right)^{1-\alpha} - (\rho_s)^{1-\alpha}}{1-\alpha} \qquad (11)$$

$$L_b(\boldsymbol{n}) = \sum_{1\le j\le K} \mathbf{1}_{\{\boldsymbol{n}(t)\in\mathcal{C}^{(j)}\}} \sum_{s\in\mathcal{S}_b(j)} n_s^\alpha \frac{\left(A_s^{(j)}\right)^{1-\alpha} - (\rho_s)^{1-\alpha}}{1-\alpha}. \qquad (12)$$

We have $L_a(\boldsymbol{n}(t)) \le 0$, so that $0 \le L(\boldsymbol{n}(t)) \le L_b(\boldsymbol{n}(t))$. Then it suffices to show that $L_b(\boldsymbol{n}(t))$ reaches 0 in a finite time. Notice that

$$L_b(\boldsymbol{n}(t)) = \sum_{1\le j\le K} \mathbf{1}_{\{\boldsymbol{n}(t)\in\mathcal{C}^{(j)}\}} \sum_{s\in\mathcal{S}_b(j)} b_s(j) n_s^\alpha$$

where

$$b_s(j) = \frac{\left(A_s^{(j)}\right)^{1-\alpha} - (\rho_s)^{1-\alpha}}{1-\alpha} > 0.$$

As in (10), for $\boldsymbol{n}(t) \in \mathcal{C}^{(j)}$, we write

$$\frac{dL_b(\boldsymbol{n}(t))}{dt} = -\sum_{s\in S_b(j)} a_s(j) n_s^{\alpha-1}(t) \qquad (13)$$

where

$$a_s(j) = -\alpha\mu_s\left(\rho_s - A_s^{(j)}\right) \frac{\left(A_s^{(j)}\right)^{1-\alpha} - (\rho_s)^{1-\alpha}}{1-\alpha} > 0.$$

For $\alpha > 1$ and $\boldsymbol{n} \in \mathcal{C}^{(j)}$, consider two norms on $\mathbb{R}^{S_b(j)}$

$$\|x\|^{(1)} = \left(\sum_{s\in\mathcal{S}_b(j)} b_s(j) x_s^{\frac{\alpha}{\alpha-1}}\right)^{\frac{\alpha-1}{\alpha}}$$

and

$$\|x\|^{(2)} = \sum_{s\in\mathcal{S}_b(j)} a_s(j)|x_s|.$$

Notice that $\|x\|^{(1)}$ and $\|x\|^{(2)}$ are equivalent on $\mathbb{R}^{S_b(j)}$, then there exists $\beta(j) > 0$ such that $\|x\|^{(2)} \ge \beta(j)\|x\|^{(1)}$. Letting $x_s = n_s^{\alpha-1}$ and $\beta = \min_{1\le j\le K}\beta(j)$, it gives $-\frac{dL_b}{dt}(t) \ge \beta L_b^{\frac{\alpha-1}{\alpha}}(t)$. Hence, we have

$$\frac{dL_b}{dt}(t) \le -\beta L_b^{\frac{\alpha-1}{\alpha}}(t), \quad \alpha > 1. \qquad (14)$$

For $0 < \alpha < 1$, notice that $0 < 1-\alpha < 1$, then by *Hölder's inequality* we have

$$\left(\sum_{s\in\mathcal{S}_b(j)} a_s(j) n_s^{\alpha-1}\right)^\alpha \cdot \left(\sum_{s\in\mathcal{S}_b(j)} b_s(j) n_s^\alpha\right)^{1-\alpha} \qquad (15)$$

$$= \left(\sum_{s\in\mathcal{S}_b(j)} a_s(j)\left(\frac{1}{n_s^{\alpha(1-\alpha)}}\right)^{1/\alpha}\right)^\alpha \cdot \left(\sum_{s\in\mathcal{S}_b(j)} b_s(j)\left(n_s^{\alpha(1-\alpha)}\right)^{1/(1-\alpha)}\right)^{1-\alpha} \qquad (16)$$

$$\ge \sum_{s\in\mathcal{S}_b(j)} a_s(j)^\alpha b_s(j)^{1-\alpha} \triangleq \gamma(j) \qquad (17)$$

which gives

$$\sum_{s\in\mathcal{S}_b(j)} a_s(j) n_s^{\alpha-1} \ge \gamma(j)^{\frac{\alpha-1}{\alpha}} \left(\sum_{s\in\mathcal{S}_b(j)} b_s(j) n_s^\alpha\right)^{\frac{\alpha-1}{\alpha}};$$

then with some abuse of notation, there exists $\beta(j) = \gamma(j)^{\frac{\alpha-1}{\alpha}} > 0$, and by defining β in the same way we have

$$\frac{dL_b(\boldsymbol{n}(t))}{dt}(t) \le -\beta L_b(\boldsymbol{n}(t))^{\frac{\alpha-1}{\alpha}}, \quad 0 < \alpha < 1. \qquad (18)$$

Then by (14) and (18) we have

$$L_b(\boldsymbol{n}(t)) \le \left(L_b(\boldsymbol{n}(0))^{1/\alpha} - \beta t/\alpha\right)^\alpha.$$

Thus, after finite time $T_1 = L_b(\boldsymbol{n}(0))^{1/\alpha}\alpha/\beta$, $L_b(\boldsymbol{n}(t)) = 0$. This implies that for all $t \ge T_1$, $L(\boldsymbol{n}(t)) = 0$ and thus $\boldsymbol{n}(t)$ must hit 0. □

We now provide a necessary stability condition for α-fair allocations. Note that Theorem 2 already provides a necessary stability condition, since the traffic load cannot exceed the maximum stability region. The following result gives tighter necessary conditions.

Theorem 5 (Discrete Rate Region—Necessary Stability Condition): For a discrete rate region $\mathcal{R}$, the system is unstable under the α-fair allocation, for $\alpha > 0$, if one of the following conditions holds.

(i) There exists $A^{(l)} \in \mathcal{R}^\alpha$ such that $\boldsymbol{\rho} - A^{(l)} > 0$.

(ii) There exists a class s such that $\rho_s > \max_{A^{(j)}\in\mathcal{R}^\alpha} A_s^{(j)}$.

The proof of Theorem 5 is provided in the Appendix.

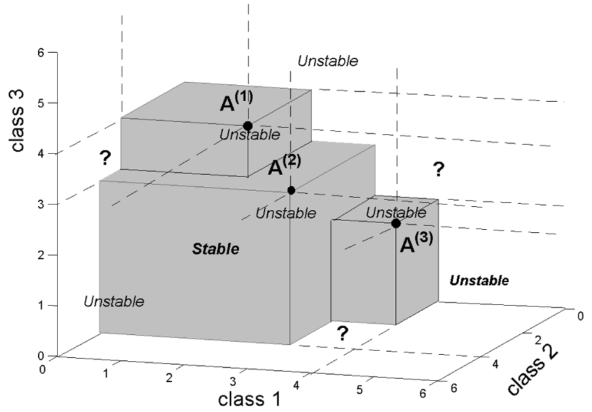


Fig. 3. Inner bound on the stability region (Theorem 4) of an α-fair allocation in a 3-flow class system—$\alpha = 0.5, \mathcal{R}^\alpha = \{(4,2,2),(3,4,3),(2,3,4)\}$.

B. Continuous Nonconvex Rate Region

When the number of rate vectors allocated by the α-fair allocation is small, there can be a significant gap between the sufficient and necessary conditions for stability regions derived in Theorems 4 and 5, as shown in Fig. 3. When $\mathcal{R}^\alpha$ has more points, the gap reduces, and ultimately tends to 0 when $\mathcal{R}^\alpha$ becomes continuous, which is an important special case often encountered in utility maximization models. The following result formalizes this observation and the proof is provided in the Appendix.

Corollary 1: (Continuous and Nonconvex Rate Regions): If the set $\mathcal{R}^\alpha$ for α-fair allocation is continuous, then the stability region of this allocation is the smallest coordinate-convex set containing $c(\mathcal{R}^\alpha)$.

In Section IV, we will consider time-varying rate regions, which can be either convex or nonconvex at any fixed time instant. In particular, in the case of time-varying convex rate regions, the set of allocated rate vectors is continuous by the convexity of each possible rate region. Then, a similar phenomenon as described in Corollary 1 occurs, which explains why we will be able to exactly characterize the stability region.

IV. Stability With Time-Varying Rate Region

We now investigate the stability region of various resource allocations in networks with time-varying rate regions. The network state is described by $(\boldsymbol{N}(t), \mathcal{R}(t))$ where we assume $\{\mathcal{R}(t)\}_{t=0}^{\infty}$ is a stationary and ergodic process as described in Section II-B, i.e., $\mathbb{P}\{\mathcal{R}(t) = \mathcal{R}_i\} = \pi_i$, $i \in \mathcal{I}$ while each $\mathcal{R}_i$ is finite and compact.

We first describe the evolution of the network in the fluid limit for any type of allocation. Consider an allocation which allocates the rate vector $\boldsymbol{\phi}^{(i)}(\boldsymbol{N})$ at state $\boldsymbol{N}$ when the rate region is $\mathcal{R}_i$, and satisfies Properties 1–3 defined in Section II-C. The evolution of the system fluid limit is given by

$$\frac{dn_s}{dt} = \lambda_s - \mu_s \sum_{i\in\mathcal{I}} \pi_i \phi_s^{(i)}(\boldsymbol{n}), \quad \forall\, s \in \mathcal{S}. \tag{19}$$

The proof of the above statement is presented in the Appendix.

The proof techniques applied to obtain sufficient and necessary conditions for stability in the case of time-varying rate regions are similar to those used in the previous section. We will characterize the maximum stability region, and then derive the stability region of α-fair allocations.

A. Maximum Stability Region

The following theorem is the analog of Theorem 2 for networks with fixed and arbitrary rate region. In that case, it turns out that the MP allocation also achieves maximum stability. Recall that the MP allocation solves (5) with rate region $\mathcal{R}(t)$ at any time t, based on which we have the following.

Theorem 6 (Maximum Stability Region—Time-Varying Rate Region): Consider a network with time-varying rate region $\mathcal{R}(t)$, if we define[1]

$$\overline{\mathcal{R}} = \sum_{i\in\mathcal{I}} \pi_i \mathcal{R}_i, \tag{20}$$

then the maximum stability region is the smallest convex, coordinate convex set containing $\overline{\mathcal{R}}$ and it can be achieved by the MP allocation.

Proof: Let $\boldsymbol{\phi}^{(i),M}(\boldsymbol{n})$ denote the allocated rate vector under the MP policy when $\mathcal{R}(t) = \mathcal{R}_i$, and also let $\bar{\boldsymbol{\phi}}^M = \sum_{i\in\mathcal{I}} \pi_i \phi_s^{(i),M}$.

As stated in [1, Theorem 2], the stability region of the MP allocation is just the convex hull of the rate region, or the set of rate vectors allocated by the MP policy. Then we show that, in the case of time-varying rate regions, the average allocated rate vector $\bar{\boldsymbol{\phi}}^M$ in turn solves the MP allocation in case the rate region is $\overline{\mathcal{R}}$. In other words, it implies that in the fluid limit model in the case of time-varying rate regions, the service rate of the fluid dynamics is provided by an MP policy over rate region $\overline{\mathcal{R}}$. Thus, the stability region of such dynamics follows from Theorem 2 is the convex hull of $\overline{\mathcal{R}}$ which is also the maximum stability region.

Now we show the optimality of $\bar{\boldsymbol{\phi}}^M$ at state $\boldsymbol{n}$. Since for any $\boldsymbol{x} \in \overline{\mathcal{R}}$, $\boldsymbol{x}$ can be represented as $\sum_{i\in\mathcal{I}} \pi_i \boldsymbol{x}^{(i)}$, with $\boldsymbol{x}^{(i)} \in \mathcal{R}_i$, for all $i \in \mathcal{I}$, we thus have

$$\begin{aligned}\sum_{s\in\mathcal{S}} n_s \sum_{i\in\mathcal{I}} \pi_i \phi_s^{(i),M}(\boldsymbol{n}) &= \sum_{i\in\mathcal{I}} \pi_i \sum_{s\in\mathcal{S}} n_s \phi_s^{(i),M}(\boldsymbol{n}) \\ &\geq \sum_{i\in\mathcal{I}} \pi_i \sum_{s\in\mathcal{S}} n_s x_s^{(i)} \\ &= \sum_{s\in\mathcal{S}} n_s \sum_{i\in\mathcal{I}} \pi_i x_s^{(i)}.\end{aligned}$$

This completes the proof. □

Note that even with different time-scale assumptions, a similar result on the maximum stability region in the case of time-varying rate regions was provided in [23] for the specific case where $\mathcal{R}_i$ is a convex polytope, where a certain channel-aware scheduling policy is adopted to obtain the maximum stability region in the context.

B. Stability Region of α-Fair Allocations

We now turn to the characterization of the stability region of α-fair allocations. Observe that by (19), the possible service

[1] The addition of sets is defined as follows: $\mathcal{R}_1 + \mathcal{R}_2 = \{x_1 + x_2 : x_1 \in \mathcal{R}_1, x_2 \in \mathcal{R}_2\}$.

rate for an α-fair allocation in the fluid limit is the average of the allocated rate vectors in the various rate regions. It is then natural to define the *average* set of rate vectors allocated by the α-fair allocation in the fluid limit as

$$\overline{\mathcal{R}^\alpha} = \left\{ \boldsymbol{\phi} : \exists \boldsymbol{n} \in \mathbb{R}_+^S, \boldsymbol{\phi} = \sum_i \pi_i \times \boldsymbol{\phi}^{(i)}(\boldsymbol{n}) \right\}. \tag{21}$$

This is the set of all possible service rate vectors in the fluid limit. We further define the average rate region in the fluid limit for the α-fair allocation as the smallest coordinate-convex set containing $\overline{\mathcal{R}^\alpha}$, i.e.,

$$\overline{\Lambda^\alpha} = \{\boldsymbol{y} : \exists \boldsymbol{x} \in \overline{\mathcal{R}^\alpha} \text{ s.t. } 0 \le \boldsymbol{y} \le \boldsymbol{x}\}. \tag{22}$$

Now if we first consider the case where each $\mathcal{R}_i$ is finite and discrete, based on the concepts of $\overline{\mathcal{R}^\alpha}$ and $\overline{\Lambda^\alpha}$ and following similar procedures to those in Section III, we are able to see that the stability conditions in the case of time-varying rate regions can be a direct analog to Theorems 4 and 5. Assume that at time t, the rate region $\mathcal{R}(t) = \mathcal{R}_i$. We write $\mathcal{R}_i = \{A^{(i,1)}, \dots, A^{(i,K_i)}\}$, and then the set of average allocated rate vector in the fluid limit can be represented as

$$\overline{\mathcal{R}^\alpha} = \left\{ \sum_{i\in\mathcal{I}} \pi_i A^{(i,l_i)} : \exists \boldsymbol{n} \in \mathbb{R}_+^S, \boldsymbol{\phi}^{(i)}(\boldsymbol{n}) = A^{(i,l_i)},\ \forall\, i \in \mathcal{I} \right\}.$$

Since the α-allocation is a cone policy, the rate vector $A^{(i,l_i)}$ is allocated if and only if the state $\boldsymbol{n}(t)$ belongs to the cone $\mathcal{C}^{(i,l_i)}$. We introduce a set of cones $\mathcal{C}^{(l_1,\dots,l_I)}$, which corresponds to the states where the rate vector in the fluid limit is $\sum_{i\in\mathcal{I}} \pi_i A^{(i,l_i)}$. These cones are defined by the intersections of cones

$$\mathcal{C}^{(l_1,\dots,l_I)} = \bigcap_{i\in I} \mathcal{C}^{(i,l_i)}.$$

Theorem 7 (Discrete and Time-Varying Rate Region): For a discrete and time-varying rate region $\mathcal{R}(t)$ and $\alpha > 0$, the stability conditions are given as follows:

(i) if $\boldsymbol{\rho} \in \overline{\Lambda_o^\alpha}$, then the system is stable;

(ii) if $\boldsymbol{\rho}$ dominates a vector in $\overline{\mathcal{R}^\alpha}$, i.e., $\exists \boldsymbol{A}' \in \overline{\mathcal{R}^\alpha}$ s.t. $\boldsymbol{A}' < \boldsymbol{\rho}$, or for some class s, $\rho_s > A_s$ for all $\boldsymbol{A} \in \overline{\mathcal{R}^\alpha}$, then the system is unstable.

Proof: (i) Now if $\boldsymbol{\rho} \in \overline{\Lambda_o^\alpha}$, then there exists $\sum_{i\in\mathcal{I}} \pi_i A^{(i,l_i)} \in \overline{\mathcal{R}^\alpha}$ such that $\boldsymbol{\rho} < \sum_{i\in\mathcal{I}} \pi_i A^{(i,l_i)}$. Thus, then there must exit a decomposition $\boldsymbol{\rho} = \sum_{i\in\mathcal{I}} \pi_i \boldsymbol{\rho}^{(i)}$ such that $\boldsymbol{\rho}^{(i)} < A^{(i,l_i)}$. For each rate region $\mathcal{R}_i$, we define

$$V_j^{(i)}(\boldsymbol{n}) = \frac{\sum_{s\in\mathcal{S}} n_s(t)^\alpha \left(\left(A_s^{(i,j)} \right)^{1-\alpha} - \left(\rho_s^{(i)} \right)^{1-\alpha} \right)}{(1-\alpha)}$$

and notice that $V_j^{(i)}(\boldsymbol{n}) > 0$ whenever $A^{(i,j)}$ is allocated. Then we consider the following function:

$$L(\boldsymbol{n}(t)) = \sum_{i\in\mathcal{I}} \sum_{1\le j\le K_i} V_j^{(i)}(\boldsymbol{n}) \mathbf{1}_{\{\boldsymbol{n}(t)\in\mathcal{C}^{(j)}\}} \tag{23}$$

which is strictly positive and equals 0 if and only if $\boldsymbol{n}(t) = 0$. Moreover, we observe that (23) yields a form of linear combination of (9), thus by applying the arguments in Theorem 4 to (23), we can see that $L(\boldsymbol{n}(t))$ vanishes within finite time.

(ii) When there exists $\sum_{i\in\mathcal{I}} \pi_i A^{(i,l_i)} \in \overline{\mathcal{R}^\alpha}$ such that $\boldsymbol{\rho} > \sum_{i\in\mathcal{I}} \pi_i A^{(i,l_i)}$, we can decompose $\boldsymbol{\rho} = \sum_{i\in\mathcal{I}} \pi_i \boldsymbol{\rho}^{(i)}$ such that $\boldsymbol{\rho}^{(i)} > A^{(i,l_i)}$. Thus, the same arguments in Theorem 5 based on the Lyapunov function defined in (23) completes the proof. □

Similarly, the stability conditions provided in the preceding theorem for discrete rate regions may have a gap between the sufficient and necessary conditions, which can also be reduced in the case where $\overline{\mathcal{R}^\alpha}$ is continuous by discrete approximation as shown in the proof of Corollary 1.

Corollary 2 (Continuous and Time-Varying Rate Region): When $\overline{\mathcal{R}^\alpha}$ is continuous, the stability region of the α-fair allocation is $\overline{\Lambda^\alpha}$.

In particular, Corollary 2 characterizes the stability region of the case where each $\mathcal{R}_i$ is convex. (Recall that the set of allocated vectors under α-fair allocation of a convex rate region is exactly its Pareto boundary.) In fact, the case that every $\mathcal{R}_i$ is convex caters many practical scenarios. In Section V, we will focus on the case of time-varying but convex rate regions and study its sensitivity to α.

V. Fairness–Stability Tradeoff

In this section, we discuss the tradeoff between fairness and flow-level stability and study the sensitivity of the stability region of α-fair allocations to the fairness parameter α. When the rate region is fixed convex and coordinate-convex, we know from Theorem 1 that the stability region is insensitive to α. This property is lost for networks with nonconvex or time-varying rate regions. In this case, quantifying the sensitivity of the stability region w.r.t. α proves to be quite challenging, and we restrict the analysis to the case of networks with two classes only.

A. Sensitivity in the Case of a Nonconvex Rate Region

A preliminary sensitivity analysis in the case of a nonconvex rate region has been performed in [7]. It has been proved that there exist two fairness parameters $\alpha_{\min}$ and $\alpha_{\max}$ with $\alpha_{\min} < \alpha_{\max}$ such that the stability regions of $\alpha_{\min}$-fair and $\alpha_{\max}$-fair allocations are the minimum and the maximum, respectively. This result indicates that the stability region tends to be larger for smaller values of α. In particular max-min fairness always leads to the smallest stability region, whereas the allocation maximizing the network throughput leads to the greatest region.

B. Sensitivity in Case of Time-Varying Rate Region

We now investigate the sensitivity of the stability region to the choice of the resource allocation for time-varying rate regions, and focus on the case where each rate region is convex. We provide two preliminary results indicating that for time-varying rate region, the stability region of α-fair allocations is reduced when α grows.

We consider time-varying rate regions satisfying the *scaling rule*, i.e., $\mathcal{R}_i = a^{i,j} \times \mathcal{R}_j$ for all $i, j \in \mathcal{I}$ and some $a^{i,j} \in \mathbb{R}_+^S$, where the product is defined as the component-wise scaling by factor $a_s^{i,j}$ for class s coordinate. The scaling rule indicates that the basic shape of rate regions does not change. This assumption is valid in many practical systems since the type of network and resource allocation scheme (e.g., time or power sharing in wireless networks) determines the basic shape of the rate region. On assuming that there are M possible rate regions, we sort them as $\mathcal{R}_1, \mathcal{R}_2, \ldots, \mathcal{R}_M$ such that $\mathcal{R}_i = a^{(i)} \times \mathcal{R}_1$, $i = 2, \ldots, M$, and

$$1 < \frac{a_2^{(2)}}{a_1^{(2)}} < \frac{a_2^{(3)}}{a_1^{(3)}} < \cdots < \frac{a_2^{(M)}}{a_1^{(M)}}. \tag{24}$$

Then we have the following two corollaries.

Corollary 3 (Convex and Time-Varying Rate Region): For $\alpha \geq 1$, the stability region $\overline{\Lambda_\alpha}$ is decreasing as α increases. In addition, if we let $\partial\mathcal{R}_i$ denote the Pareto boundary of $\mathcal{R}_i$, and $\partial\mathcal{R}_i$ is a line segment, the monotonicity of $\overline{\Lambda_\alpha}$ holds for all $\alpha > 0$.

Corollary 4: The maximum stability region $\overline{\mathcal{R}}$ is achieved as $\alpha \rightarrow 0$, i.e., $\overline{\Lambda_\alpha} \rightarrow \bar{\mathcal{R}}$.

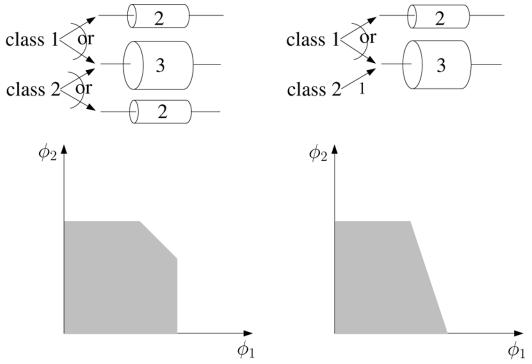
In fact, we will see in Section VI that in some cases, there exists $0 < \alpha_0 < 1$ such that when $\alpha < \alpha_0$, the maximum stability is achieved, i.e., $\overline{\mathcal{R}_\alpha} = \overline{\mathcal{R}}$.

We conclude this subsection by presenting possible cases where even with time-varying rate region, the stability region of α-fair allocations is insensitive to α. This can be the case when all flow classes suffer from the same capacity variations. This special case for insensitivity is true for a system with an arbitrary number of classes, if for all $i \in \mathcal{I}$, there exists a constant c_i such that $\mathcal{R}_i = c_i\mathcal{R}_1$, the stability region is given as $\overline{\Lambda_\alpha} = \sum_{i\in\mathcal{I}} \pi_i c_i \mathcal{R}_i$. This is because when solving (1) with different rate regions $\mathcal{R}_i$ where $\mathcal{R}_i = c_i\mathcal{R}_1$, by scaling the decision variable $\boldsymbol{\phi}^{(j,\alpha)}$ with the same constant c_i, we have $\boldsymbol{\phi}^{(i,\alpha)} = c_i\boldsymbol{\phi}^{(1,\alpha)}$. An example of such systems is the downlink of a cell in a wireless network where the power of the base station allocated to data traffic may vary because of the presence of high-priority traffic such as voice.

As discussed in this section, when the rate region is nonconvex or time-varying, the stability region of a resource allocation scheme depends on the chosen fairness parameter α. Fairness can be imposed only at the expense of reducing the stability region. In a number of practical networks where this fairness–stability tradeoff exists, it becomes crucial to choose a fairness objective that achieves the right balance between fairness and performance.

VI. Examples

In this section, we present some numerical experiments to illustrate the analytical results derived in the previous sections on various types of data networks: wired networks and wireless networks with centralized or distributed resource allocation, whose particular applications lead to nonconvex or time-varying rate regions. The sensitivity of the stability region of α-fair allocations to the fairness parameter α depends on the considered network. We also observe that the sensitivity is usually much higher for wireless networks than for wired networks due to the sharp variation of rate regions. We further study the impact of the stability–fairness tradeoff on the system performance in this section, especially in the case of time-varying rate regions. Various performance metrics defined in Section II-F are examined in the numerical examples.



(a) $\mathcal{R}_1$ (b) $\mathcal{R}_2$

Fig. 4. A wired network with link failures: multipath routing without flow splitting.

A. Wired Networks With Link Failures

In this subsection, we investigate time-varying rate regions in wired networks due to link failures. The different sets of time-varying broken links generate various link failure states, which in turn defines time-varying rate regions. We study two different cases depending on the underlying routing and flow management mechanism: (i) multipath routing without flow splitting, and (ii) multipath routing with flow splitting.

A wired network is represented as a set of L links and K routes where each route k is defined as a subset r_k of the set of links $\{1, \ldots, L\}$. Let $C = \{C_1, \ldots, C_L\}$ be the capacity vector with $C_l > 0$. We refer to the routing matrix as the $K \times L$-dimensional matrix R whose k, l-entry is equal to 1 if $l \in r_k$, and 0 otherwise. The routing matrix varies with link failure states, and we denote by R_i the routing matrix in link failure state i.

1) Multipath Routing Without Flow Splitting: We now assume that each class is assigned a set of routes for each link failure state i, but at any instant of time, each class s can choose only a single route in the subset of routes $m_s(i)$. We let $\mathcal{M}_i$ be the set of $S \times K$ stochastic matrices such that on each row s, the s, k-entries are equal to 0 for all k except in the set $m_s(i)$. Each matrix $M \in \mathcal{M}$ corresponds to a particular route choice. We let $\boldsymbol{\phi}$ denote a row vector; then the rate region in the link failure state i is the convex hull of the capacity sets associated with the routing matrices $M \in \mathcal{M}_i$, given by

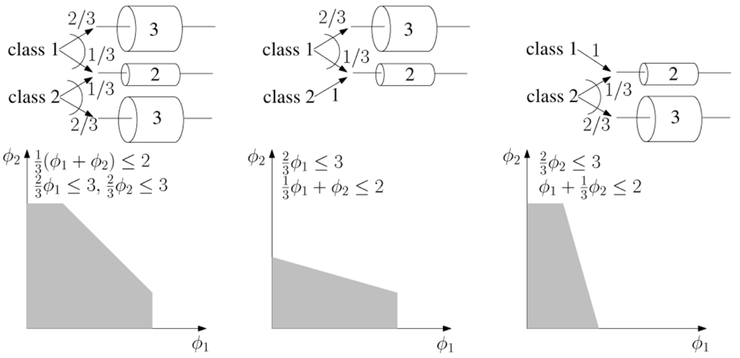
$$\mathcal{R}_i = \text{convex hull of } \{\boldsymbol{\phi} : \exists M \in \mathcal{M}_i, \boldsymbol{\phi} M R_i \leq C\}.$$

Consider the example of Fig. 4, where $C = (2, 3, 2)$ and

$$R_1 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \quad R_2 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix},$$

$$\mathcal{M}_1 = \left\{ \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix}, \begin{pmatrix} 0 & 1 & 0 \\ 0 & 1 & 0 \end{pmatrix}, \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \right\},$$

$$\mathcal{M}_2 = \left\{ \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \begin{pmatrix} 0 & 1 \\ 0 & 1 \end{pmatrix} \right\}.$$


class 1
class 2
2/3
1/3
1/3
2/3
3
2
3
ϕ_2
$\frac{1}{3}(\phi_1+\phi_2)\le 2$
$\frac{2}{3}\phi_1 \le 3,\ \frac{2}{3}\phi_2 \le 3$
ϕ_1
class 1
class 2
2/3
1/3
1
3
2
ϕ_2
$\frac{2}{3}\phi_1 \le 3$
$\frac{1}{3}\phi_1 + \phi_2 \le 2$
ϕ_1
class 1
class 2
1
1/3
2/3
2
3
ϕ_2
$\frac{2}{3}\phi_2 \le 3$
$\phi_1 + \frac{1}{3}\phi_2 \le 2$
ϕ_1

(a) $\mathcal{R}_1$ (b) $\mathcal{R}_2$ (c) $\mathcal{R}_3$

Fig. 5. A wired network with link failures: multipath routing with flow splitting.

Fig. 6(a) shows the stability regions for different values of α when $\pi_1 = \pi_2 = 1/2$. We observe that the stability region decreases as α increases. However the sensitivity to α is rather limited, and when $\alpha < 0.5$, the maximum stability region is achieved.

2) Multipath Routing With Flow Splitting: Suppose now that for link failure state i, each class s can use all routes in the set $m_s(i)$ at the same time. Abusing the notation somewhat, we again let $\mathcal{M}$ be the set of $S \times K$ stochastic matrices such that on each row s, the (s,k)-entries are equal to 0 for all k except those in the set $m_s(i)$. Each matrix $M \in \mathcal{M}_i$ corresponds to a particular traffic splitting scheme at the failure state i. Then, the rate region is given by

$$\mathcal{R}_i = \{\boldsymbol{\phi} : \exists M \in \mathcal{M}_i, \boldsymbol{\phi} M R_i \le C\}.$$

Consider the example in Fig. 5 with three links, where $C = (3,2,3)$, and

$$R_1 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \quad R_2 = R_3 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix},$$
$$\mathcal{M}_1 = \left\{ \begin{pmatrix} \frac{2}{3} & \frac{1}{3} & 0 \\ 0 & \frac{1}{3} & \frac{2}{3} \end{pmatrix} \right\}, \mathcal{M}_2 = \left\{ \begin{pmatrix} \frac{2}{3} & \frac{1}{3} & 0 \\ 0 & 1 & 0 \end{pmatrix} \right\},$$
$$\mathcal{M}_3 = \left\{ \begin{pmatrix} 0 & 1 & 0 \\ 0 & \frac{1}{3} & \frac{2}{3} \end{pmatrix} \right\}.$$

Fig. 6(b) shows the change of stability regions for different values of α, where we assume that $\pi_1 = \pi_2 = \pi_3 = 1/3$. As illustrated, the sensitivity to α is more substantial compared to the case without flow splitting. When $\alpha < 0.2$, the maximum stability region is achieved.

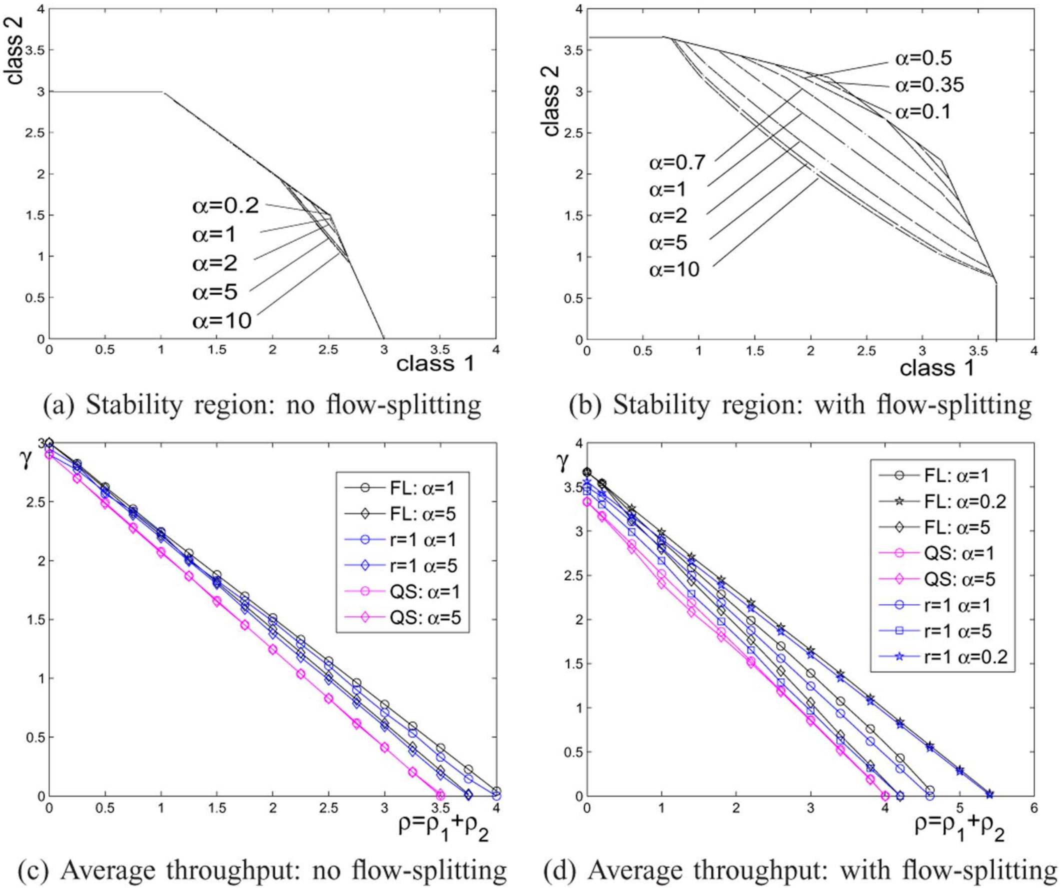
3) Throughput Performance in Different Regimes: We see that in Fig. 6(a) and (b) the stability region generally shrinks as the fairness parameter α increases. However, it is hard to quantitatively characterize the change on the stability region. In this case, one way is to observe the average system throughput $\gamma(\boldsymbol{\rho})$ with different fairness parameter α while scaling the vector of offered load by a positive parameter η, i.e., $\boldsymbol{\rho} = \eta \boldsymbol{u}$ where $\boldsymbol{u}$ is chosen as a particular unit vector. Thus, the threshold value of the scaling parameter η when $\gamma(\eta \boldsymbol{u})$ decreases to zero implies the load vector $\boldsymbol{\rho} = \eta \boldsymbol{u}$ exceeds the stability region.

We also examine the problem of different time scales of rate region variation. It has been shown in [4] that for a processor-sharing type network, where the service rate of each flow is modulated by an independent stochastic process, the throughput varies monotonically with the speed of rate variations in the sense of stochastic ordering. For data networks with utility-based rate allocation, in this paper, we consider a family of systems parameterized by the speed of rate region variation, denoted by r, i.e., $\mathcal{R}^{(r)}(t) = \mathcal{R}(rt)$. When $r \to \infty$ (resp., $r \to 0$), it represents the limiting regime where the rate region process evolves on an infinitely fast (resp., slow), which is termed as the *fluid* (resp., *quasi-stationary*) regime.

Now we let the average system throughput be labeled by the three factors, the fairness parameter α, the load vector $\boldsymbol{\rho}$ (or the scaling factor η), and speed of rate region variation r, denoted by $\gamma(\alpha, \boldsymbol{\rho}, r)$. Fig. 6(c) and (d) shows $\gamma(\alpha, \boldsymbol{\rho}, r)$ of the example of wired network with multipath routing. We choose the load vector with $\rho_1 = \rho_2$, and let $\alpha = 0.2, 1, 5$, $r = 0, 1, \infty$. We observe that for any fixed $\boldsymbol{\rho}$ and r, as α increases we have $\gamma(0.2, \boldsymbol{\rho}, r) < \gamma(1, \boldsymbol{\rho}, r) < \gamma(5, \boldsymbol{\rho}, r)$, and the threshold $\boldsymbol{\rho}$ when γ decreases to 0 also yields the monotonicity with respect to α. For different regimes of rate region variations, we observe that in the fluid regime the system has the best throughput performance, and when the speed parameter r decreases, the average throughput generally decreases, i.e., $\gamma(\alpha, \boldsymbol{\rho}, 0) < \gamma(\alpha, \boldsymbol{\rho}, 1) < \gamma(\alpha, \boldsymbol{\rho}, \infty)$ for any fixed α and $\boldsymbol{\rho}$. This monotonicity in the simulation result coincides with the result in [4] under a more complicated network model where the service rate of each flow is state-dependent. A stronger result would be to prove the monotonicity of system performance with respect to the speed of rate region variations in fact exists, which is a subject of our future work.

B. Wireless Networks With Interference

1) Wireless Cellular Networks With Random Interference: We consider the downlink of a cell covered by base station (BS) 1. BS 1 serves two classes of flows generated by some users with fixed positions, as shown in Fig. 7. The rate regions of the system are time-varying due to variations of the interference generated by BSs 2 and 3. To simplify the analysis, we assume that BSs 2 and 3 cannot be active at the same time, and when active they transmit at full and fixed power. We also assume the downlink resources are shared at the BS 1 according to a time-division multiple-access (TDMA) scheme. A symmetric case and an asymmetric case with respect to the variation of



(a) Stability region: no flow-splitting (b) Stability region: with flow-splitting

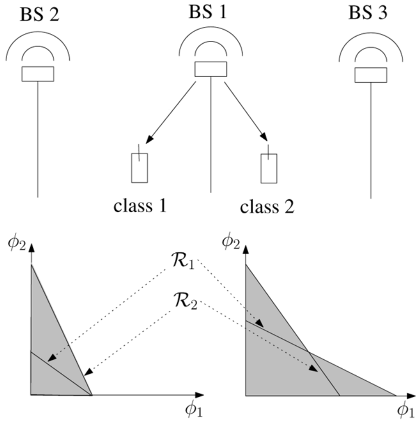
(c) Average throughput: no flow-splitting (d) Average throughput: with flow-splitting

Fig. 6. Stability regions and average flow throughput: multipath routing.

the rate region are considered, respectively. For the symmetric case, when BS 2 (resp., 3) is on and BS 3 (resp., 2) is off, the noise plus interference at the position of class-1 (resp., class-2) users is 4 and that for class-2 (resp., class-1) users is 1. For the asymmetric case, when BS 2 is on and BS 3 is off, the noise plus interference level at the position of class-1 users is 4 and that for class-2 users is 1; on the other hand, when BS 3 is on and BS 2 is off, the noise plus interference level at the position of class-1 user is 1 and that for class-2 users is 2. The corresponding rate regions are presented in Fig. 7.

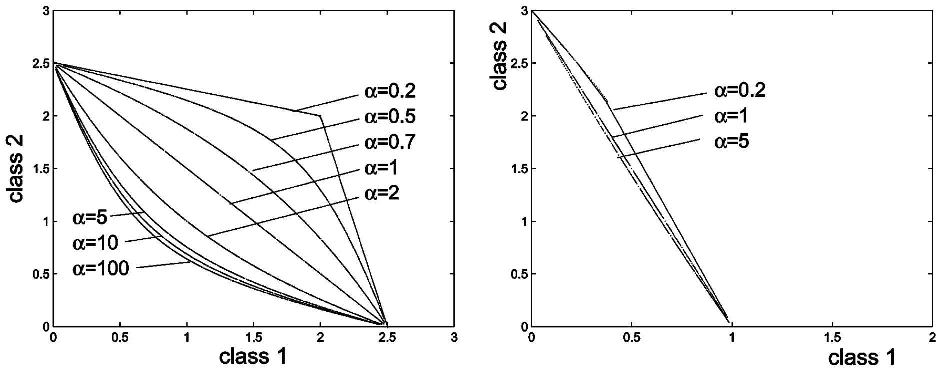
When BS 1 allocates its full power to users of class 1 (or 2), the corresponding flows are served at different rates depending on the activities of BS 2 and 3. Now the stability region for different α is shown in Fig. 8. The maximum stability region is achieved when $\alpha < 0.2$ for both cases from numerical experiments. We also observe that the sensitivity of α is even more significant in this type of networks due to the sharp variations of rate regions in wireless network.

We further examine a different set of performance metrics in the wireless cellular network model. Figs. 9(a)–(d) and 10(a)–(h) show the mean response times T (or conditioned on file size $T(x)$) and the second-order statistics of class 1 and class 2 flows under different fairness parameter $\alpha = 0.2, 1, 5$ in the fluid regime (different speeds of rate region variations are considered in Section VI-A1). In the symmetric case, obviously all the performance metrics of the two classes are equal. We observe that as α decreases, T_1, $T_1(X_1)$, $\sigma(T_1)$ and $\sigma(T_1(X_1))$ all increase monotonically, which yields the same monotonicity of performance with respect to α as in Section VI-A1. However, in the asymmetric case as shown in Fig. 10(a)–(h), it shows that when α is small, the performance experienced by class 1 and class 2 data flows react differently to the change of α. Fig. 10(a) and (b) shows that when T_2 decreases as α decreases, class 1 is penalized as T_1 tends to increase, which can be interpreted as the result of “unfairness” in the rate allocation when α is small. This “unfairness”



(a) Asymmetric (b) Symmetric

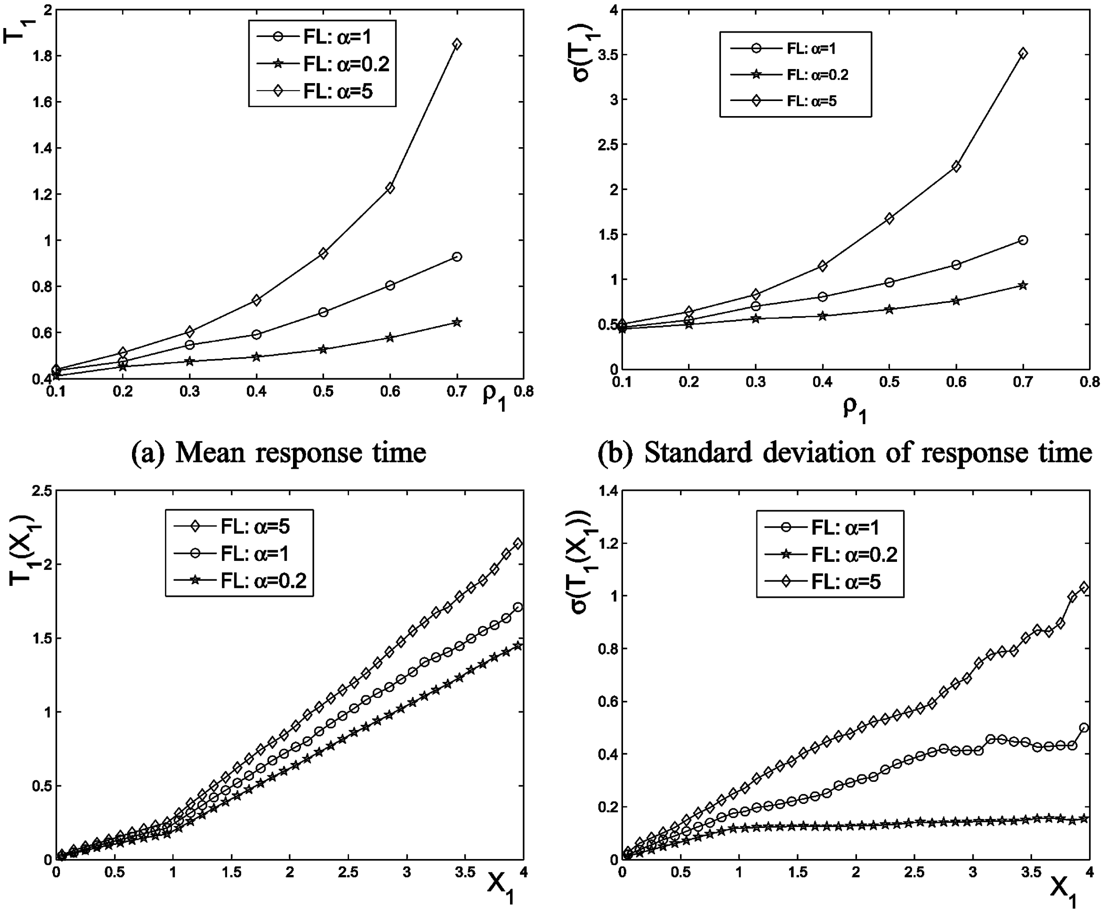
Fig. 7. Wireless cellular network with random interference.



(a) Stability region: symmetric case

(b) Stability region: asymmetric case

Fig. 8. Stability regions for wireless cellular networks with random interference.



(a) Mean response time

(b) Standard deviation of response time

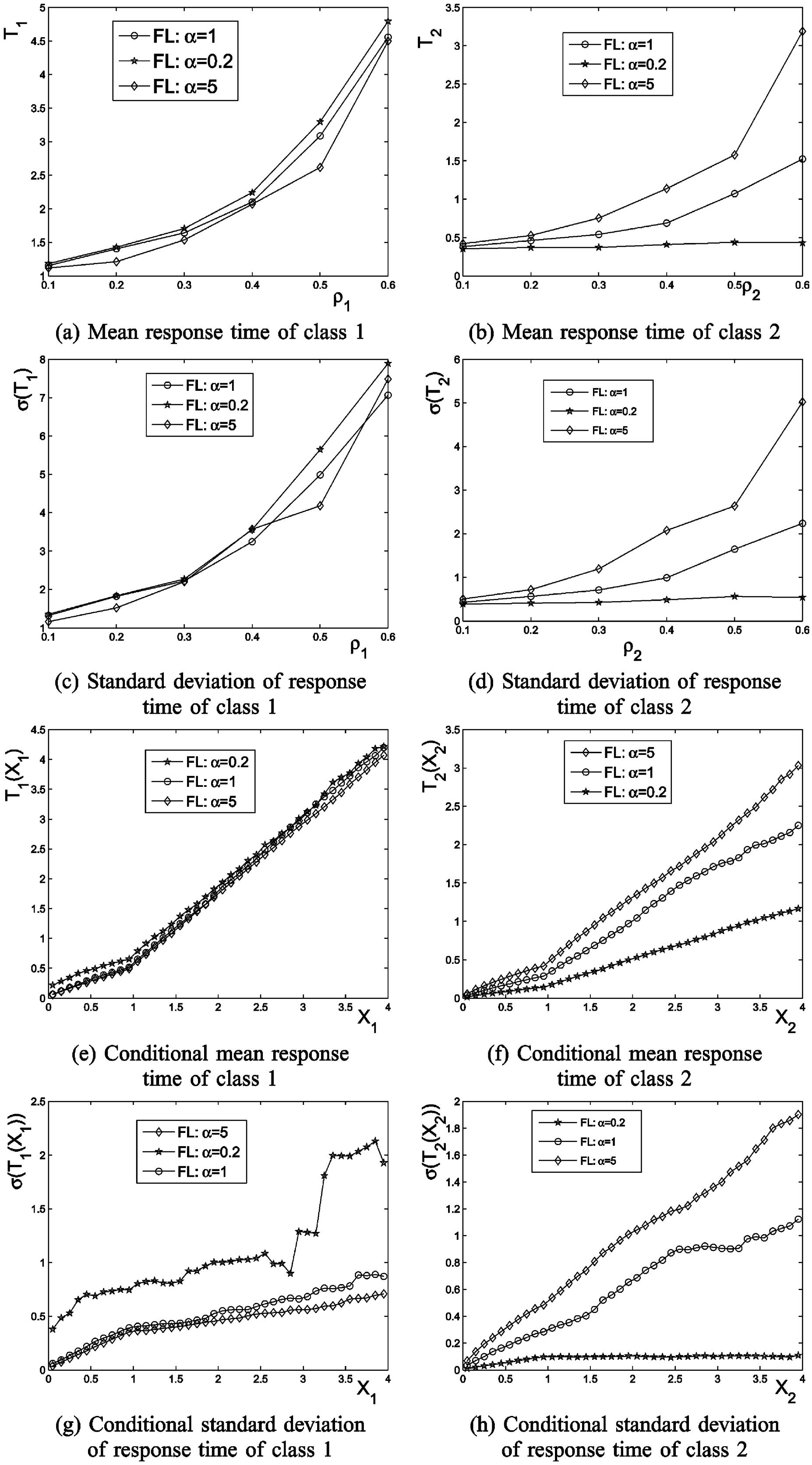
(c) Conditional mean response time

(d) Conditional standard deviation of response time

Fig. 9. Performance of response time of wireless cellular network with symmetric stability region.

also can be observed in the comparison of $\sigma(T_1)$ and $\sigma(T_2)$, $T_1(X_1)$ and $T_2(X_2)$ and $\sigma(T_1(X_1))$ and $\sigma(T_2(X_2))$ as shown in Fig. 10(c)–(h). In other words, it exemplifies the tradeoff between fairness and performance: higher average performance can possibly be achieved with smaller α but some classes of data flows are unfairly treated in the rate allocation and thus their performance is penalized.

2) Wireless Ad Hoc Networks With User Mobility: Consider a wireless *ad hoc* network where the flow class is determined by its source–destination pairs. We assume that any user node cannot transmit and receive at the same time, and thus the network adopts M contention-free coordination schemes to avoid interference. Meanwhile, each flow class chooses its routing according to the current network topology which varies constantly



(a) Mean response time of class 1

(b) Mean response time of class 2

(c) Standard deviation of response time of class 1

(d) Standard deviation of response time of class 2

(e) Conditional mean response time of class 1

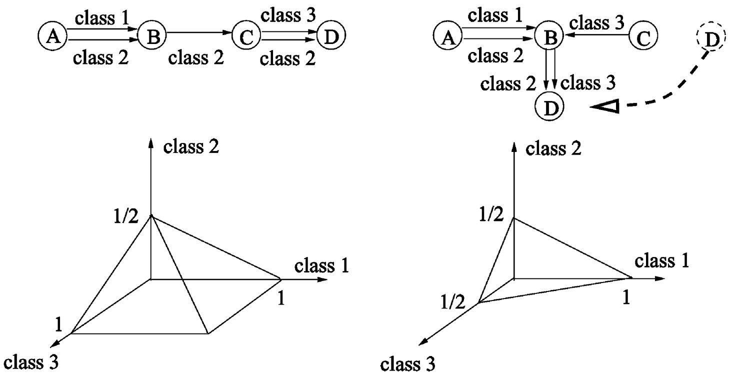
(f) Conditional mean response time of class 2

(g) Conditional standard deviation of response time of class 1

(h) Conditional standard deviation of response time of class 2

Fig. 10. Performance of response time of wireless cellular network with asymmetric stability region.

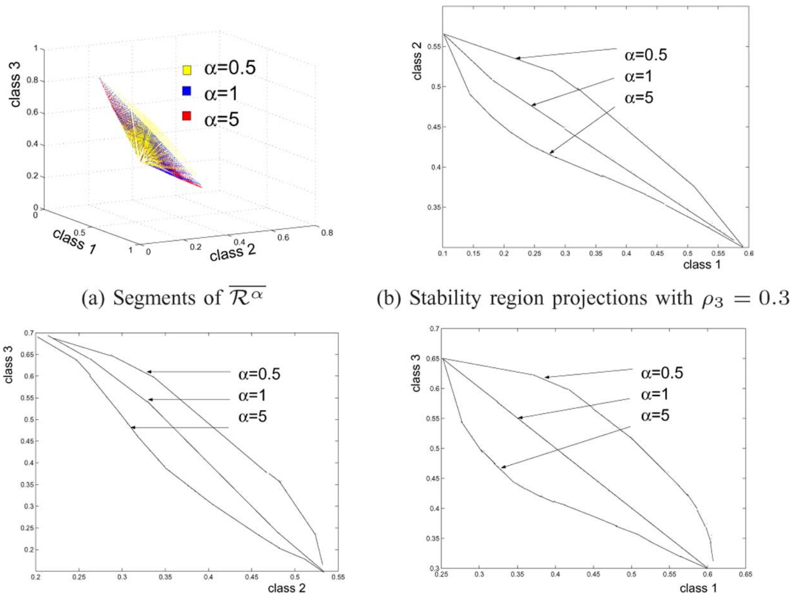
due to user mobility. If at time t, the network has a set $\mathcal{S}$ of flow classes and a set $\mathcal{L}$ of logical links, then let $A(t)$ denote the $S \times L$ routing matrix as in the wired network and $C(t)$ denote the $M \times L$ capacity matrix whose m, lth entry is the capacity



(a). Node topology type 1

(b). Node topology type 2

Fig. 11. A 3-class wireless *ad hoc* network. (a) Rate region $\mathcal{R}_1 = \{\boldsymbol{\phi} \in \mathbb{R}^3_+ : 2\phi_2 + \phi_1 \le 1, 2\phi_2 + \phi_3 \le 1\}$. (b) Rate region $\mathcal{R}_2 = \{\boldsymbol{\phi} \in \mathbb{R}^3_+ : \phi_1 + 2\phi_2 + 2\phi_3 \le 1\}$.



(a) Segments of $\overline{\mathcal{R}^\alpha}$

(b) Stability region projections with $\rho_3 = 0.3$

(c) Stability region projections with $\rho_1 = 0.3$

(d) Stability region projections with $\rho_2 = 0.3$

Fig. 12. Stability regions for different values of α for 3-class wireless *ad hoc* network with user node mobility.

of link l in transmission scheme m, then the considered *ad hoc* network has a time-varying rate region defined as

$$\mathcal{R}(t) = \{\boldsymbol{\phi} : \exists \tau \in \mathcal{T}, \boldsymbol{\phi} A(t) \le \tau C(t)\}. \tag{25}$$

Now we consider a 3-class network as shown in Fig. 11 with two different network topologies due to the mobility of user nodes, which leads to two possible rate regions described by

$$A_1 = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 1 & 1 \\ 0 & 0 & 1 \end{pmatrix}, \quad A_2 \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & 1 \end{pmatrix}$$

$$C_1 = \begin{pmatrix} 1 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix}, \quad C_2 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

The corresponding rate regions $\mathcal{R}_1, \mathcal{R}_2$ are shown in Fig. 11. For the sake of illustration, we compare the stability regions with different α-allocations by showing segments of the stability region boundaries and the projection of the stability region while fixing one coordinate of the traffic intensity $\boldsymbol{\rho}$ in Fig. 12. It shows clearly that in this 3-class example, the tradeoff between fairness α and stability also exists as we have rigorously proven for 2-class networks in Section V.

3) Wireless Random-Access Networks: We conclude this section by an example of a network with a nonconvex but fixed rate region. For this example, the results of Section III allow us to exactly characterize the stability region of some α-fair allocation.

The model is similar to that considered in [18], [31], and may represent typical wireless LANs (WLANs) or multihop wireless networks with random access Aloha-type protocols. The network is a collection $\mathcal{L}$ of L wireless links. We consider L flow classes, in which flows of class l use link l only. The links interact through interference. We assume that a transmission on link l can be successful only if none of the neighboring links is transmitting, otherwise there is a collision. Denote by $\mathcal{L}_l$ the set of links interfering link l. Time is slotted and packet transmissions last exactly one slot. At the beginning of each slot, links try to access the channel in a distributed manner, each link l transmits with probability p_l. The rate of link l, and then of class-l flows, is given by

$$\phi_l = p_l \prod_{k\in\mathcal{L}_l}(1-p_k). \tag{26}$$

The rate region of this system is in general nonconvex and given by

$$\mathcal{R} = \left\{ \boldsymbol{\phi} : \forall\, l, \phi_l \le p_l \prod_{k\in\mathcal{L}_l}(1-p_k), p_k \in (0,1) \right\}. \tag{27}$$

It is shown that for Proportional fair allocation with a nonconvex rate region (27), the transmission probabilities at state $\boldsymbol{N}$ are

$$p_l = \frac{N_l}{N_l + \sum\limits_{k:l\in\mathcal{L}_k} N_k}, \tag{28}$$

and by varying the network state $\boldsymbol{N}$ that the conditions of Subsection III-B are satisfied, namely, that the set of the allocated rate vectors is continuous and actually equals the entire rate region in this case. Therefore, we can conclude that the stability region of the Proportional fair allocation is equal to the largest open subset of $\mathcal{R}$.

VII. Conclusion and Future Work

In practical networks, the rate region that constrains resource allocation may not follow the standard assumptions of convexity and time invariance. We have shown that characterizing the stability region becomes more challenging for either nonconvex or time-varying rate regions, and its size and shape become dependent on the fairness parameter α in the utility objective function of resource allocation.

For networks with arbitrary numbers of classes and with fixed and nonconvex rate regions, we have given sufficient and necessary conditions for flow-level stability of α-fair allocations, for all $\alpha > 0$. We then have extended the analysis to networks with time-varying convex rate regions, for which we have provided the stability condition of α-fair allocations, for all $\alpha > 0$. We have also studied the sensitivity of the stability region of α-fair allocations to the fairness parameter α, and have demonstrated an intriguing tradeoff between fairness and flow-level stability, and numerical examples have shown the impact of the tradeoff on system performance. This fairness–stability tradeoff raises further questions on how to choose a resource allocation policy. Interesting future research directions are quantifying the tradeoff, characterizing performance metrics, and extending the results to general arrivals/service processes.

Appendix

A. Fluid Limits

Throughout the paper, we used fluid limit techniques to investigate stability. Here we justify the differential equations governing the evolution of the system in the fluid limit. In general, the evolution of $\boldsymbol{N}(t)$ is characterized by: for each class s

$$N_s(t) = N_s(0) + B_s(t) - D_s(t) \tag{29}$$

where B_s and D_s represent the arrival and the departure processes, respectively. Considering the network in the fluid limit consists of studying a sequence of systems where the initial value $\boldsymbol{N}(0)$ grows large. More precisely, we consider an increasing sequence of numbers M_k tending to ∞ as $k \to \infty$, and such that

$$\lim_{k\to\infty} \frac{\|\boldsymbol{N}^{(k)}(0)\|}{M_k} = 1.$$

Now the recurrence of the process $\boldsymbol{N}(t)$ can be determined through the evolution of the fluid limit $n(t)$ (when it exists) of the sequence of processes $\boldsymbol{N}(M_k t)/M_k$. Most often the fluid limit is a deterministic process but not always. In the case of a fixed discrete rate region (see Section III), the fluid limit is easy to derive. We now derive the fluid limit in the case of time-varying rate region.

The departure process D_s can be seen as a Poisson process of stochastic intensity $\mu_s \int_0^t \phi_s(\boldsymbol{N}(\tau), \mathcal{R}(\tau))d\tau$.

Define $\Xi_s(t) = \int_0^t \phi_s\,(\boldsymbol{N}(\tau), \mathcal{R}(\tau))d\tau$, and

$$\begin{aligned} N_s^{(k)}(t) &= \frac{N_s(M_k t)}{M_k} \\ B_s^{(k)}(t) &= \frac{B_s(M_k t)}{M_k} \\ D_s^{(k)}(t) &= \frac{D_s(M_k t)}{M_k} \\ \Xi_s^{(k)}(t) &= \frac{\Xi_s(M_k t)}{M_k}. \end{aligned}$$

Proposition 1: The sequence $\boldsymbol{N}^{(k)}(t)$ converges unifor-mmly on compact sets (u.o.c.) when $k \to \infty$ to the fluid limit $\boldsymbol{n}(t)$ whose evolution is driven by

$$\frac{dn_s}{dt} = \lambda_s - \mu_s \sum_{i\in\mathcal{I}} \pi_i \phi_s^{(i)}(\boldsymbol{n}), \quad \forall\, s \in \mathcal{S}. \tag{30}$$

Proof: By the functional law of large numbers, $B_s^{(k)}(t) \to \lambda_s t$ u.o.c. when $k \to \infty$. Since $\mathcal{R}_i$ is a bounded set for each $i \in \mathcal{I}$, then $D_s^{(k)}(t)$, $\Xi_s^{(k)}(t)$ satisfies the Lipschitz condition for every k, which guarantees the existence of the limit $n_s(t)$ as $k \to \infty$ (see [33]). Then we only need to show that

$$\frac{d\Xi_s^{(k)}(t)}{dt} \to \sum_{i\in\mathcal{I}} \pi_i \phi_s^{(i)}(\boldsymbol{n}(t))$$

as $k \to \infty$. Note that

$$\begin{aligned} &\left| \frac{1}{\Delta t}(\Xi_s^{(k)}(t+\Delta t) - \Xi_s^{(k)}(t)) - \sum_{i\in\mathcal{I}} \pi_i \phi_s^{(i)}(\boldsymbol{n}(t)) \right| \\ &= \left| \frac{1}{\Delta t} \int_t^{t+\Delta t} \phi_s(\boldsymbol{N}^{(k)}(u), \mathcal{R}(M_k u))du \right. \\ &\qquad \left. - \sum_{i\in\mathcal{I}} \pi_i \phi_s^{(i)}(\boldsymbol{n}(t)) \right| \\ &\le \left| \frac{1}{\Delta t} \left(\int_t^{t+\Delta t} \phi_s(\boldsymbol{N}^{(k)}(u), \mathcal{R}(M_k u))du \right. \right. \\ &\qquad \left. \left. - \int_t^{t+\Delta t} \phi_s(\boldsymbol{n}(u), \mathcal{R}(M_k u))du \right) \right| \end{aligned} \tag{31}$$

$$+\left|\frac{1}{\Delta t}\int_t^{t+\Delta t}\phi_s(\boldsymbol{n}(u),\mathcal{R}(M_k u))du\right.$$
$$\left.-\frac{1}{\Delta t}\int_t^{t+\Delta t}\phi_s(\boldsymbol{n}(t),\mathcal{R}(M_k u))du\right| \quad (32)$$
$$+\left|\frac{1}{\Delta t}\int_t^{t+\Delta t}\phi_s(\boldsymbol{n}(t),\mathcal{R}(M_k u))du-\sum_{i\in\mathcal{I}}\pi_i\phi_s^{(i)}(\boldsymbol{n})\right|. \quad (33)$$

The first equality in the preceding expression is due to the homogeneity of $\boldsymbol{\phi}(\boldsymbol{N})$. Also, when $k\to\infty$, (31) and (32) $\to 0$ a.s. by the continuity of $\boldsymbol{\phi}(\boldsymbol{N})$ (uniformly w.r.t. to the various possible values of the rate region, recalling that the rate region process can take a finite number of values only). For (33), we observe that

$$\frac{1}{\Delta t}\int_t^{t+\Delta t}\phi_s(\boldsymbol{n}(t),\mathcal{R}(M_k u))du$$
$$=\frac{1}{M_k\Delta t}\int_{M_k t}^{M_k(t+\Delta t)}\phi_s(\boldsymbol{n}(t),\mathcal{R}(u))du \quad (34)$$

and $(\phi_s(\boldsymbol{n}(t),\mathcal{R}(u)))_u$ is a stationary and ergodic stochastic process for every fixed t. Notice that for such a stochastic process, denoted by c, by the strong law of large numbers we have

$$\lim_{t\to\infty}\frac{1}{t}\int_0^t c(u)du=\lim_{t\to\infty}\lim_{m\to\infty}\frac{1}{m}\sum_{j=0}^{m-1}c\left(\frac{tj}{m}\right) \quad (35)$$
$$=\mathbb{E}c(t),\ \text{a.s.} \quad (36)$$

Then (33) $\to 0$ a.s. as $k\to\infty$. We conclude the proof letting $\Delta t\to 0$. □

B. Proof of Theorem 5

Assume (i) holds. We consider the same function $L(\boldsymbol{n})$ defined in (9), but given the assumption on $\boldsymbol{\rho}$, the sign of $L(\boldsymbol{n}(t))$ cannot be guaranteed. We write $L(\boldsymbol{n}(t))=L_a(\boldsymbol{n}(t))+L_b(\boldsymbol{n}(t))$ as defined in (11) and (12), and $L_a(\boldsymbol{n})\le 0$, $L_b(\boldsymbol{n})\ge 0$. Here for notational purpose, we write $\mathcal{S}_b(t)$ and $\mathcal{S}_a(t)$ to denote the division of index set $\mathcal{S}$ at time t. Note that in this case $\mathcal{S}_a(t)$ is nonempty for all $t>0$. Notice that

$$|L(\boldsymbol{n}(t))|=|L_b(\boldsymbol{n}(t))-|L_a(\boldsymbol{n}(t))||.$$

Now if we assume that $\mathcal{S}_b(t)$ is empty after a finite time, i.e., $\mathcal{S}_a=\mathcal{S}$, then n_s grows linearly to infinity for all $s\in\mathcal{S}$.

If not, i.e., if there exists at least one time sequence $\{t_k\}_{k=1}^{\infty}$ such that $\lim_{k\to\infty}t_k=\infty$ and $\mathcal{S}_b(t_k)$ is nonempty for every t_k, by the proof of Theorem 4, $L_b(\boldsymbol{n}(t))=0$ after a finite time T for any fixed initial state, which also implies $n_s(t_k)=0$ for $s\in\mathcal{S}_b(t_k)$. However, on the other hand, we also have $\frac{dL_a}{dt}<0$ which gives that $|L_a(\boldsymbol{n}(t))|$ is monotonically increasing. In that case, if $|L_a(\boldsymbol{n}(t))|\to\infty$ as $t\to\infty$, then it naturally gives $\|\boldsymbol{n}(t)\|\to\infty$. If $|L_a(\boldsymbol{n}(t))|$ is upper-bounded instead, then we must have $|dL_a/dt|\to 0$ as $t\to\infty$. For $\alpha>1$, this is not possible since n_s linearly increases for all $s\in\mathcal{S}_a$, and thus we must have $|L_a(\boldsymbol{n}(t))|\to\infty$. For $\alpha<1$, $|dL_a/dt|\to 0$ again leads to $\|\boldsymbol{n}(t)\|\to\infty$. Hence, we have exhausted all possibilitiess and therefore conclude that the fluid limit is unstable under Assumption (i).

Now assume (ii) holds, then at any state $\boldsymbol{n}(t)$ and with any allocated rate vector $A^{(j)}$; $\delta_s^{(j)}>0$, i.e., the drift of class s is strictly positive and n_s is always increasing. Thus $\|\boldsymbol{n}\|$ linearly grows to infinity, and the network is unstable.

C. Proof of Corollary 1

Proof: When the set of allocated vectors $\mathcal{R}^\alpha$ is continuous, we approximate it by a sequence of discrete rate regions with finite number of rate vectors.

Let $\mathcal{R}^{(k)}$ be a discrete subset of $\mathcal{R}^\alpha$ such that $\mathcal{R}^{(k)}\uparrow\mathcal{R}^\alpha$ as $k\to\infty$. Correspondingly, we let $\boldsymbol{\phi}^{(k)}$ denote the allocated vector associated with $\mathcal{R}^{(k)}$ and $\boldsymbol{\phi}$ associated with $\mathcal{R}^\alpha$. Notice that by the homogeneity property, the allocated vector at state $\boldsymbol{n}$ can be viewed as a function of the unified state vector $\boldsymbol{n}/\|\boldsymbol{n}\|$, denoted by $\boldsymbol{\phi}^{(k)}(\boldsymbol{n}/\|\boldsymbol{n}\|)$. Thus, $\boldsymbol{\phi}^{(k)}(\boldsymbol{n}/\|\boldsymbol{n}\|)\to\boldsymbol{\phi}(\boldsymbol{n}/\|\boldsymbol{n}\|)$ uniformly on the compact set $[0,1]^S$. Hence, by (6), the trajectory of $\boldsymbol{n}^{(k)}$ also converges as $k\to\infty$. Now for the same system with discrete rate region $\mathcal{R}^{(k)}$, we let $\Lambda_{\text{suf}}^{(k)}$ denote the sufficient stability region defined by Theorem 4, which is the smallest coordinate-convex set containing $\mathcal{R}^{(k)}$, and we let $\Lambda_{\text{nec}}^{(k)}$ denote the necessary stability region defined by Theorem 5 as the complement of the unstable region. Thus, if $\Lambda^{(k)}$ denotes the exact stability region of the system with rate region $\mathcal{R}^{(k)}$, we must have

$$\Lambda_{\text{suf}}^{(k)}\subseteq\Lambda^{(k)}\subseteq\Lambda_{\text{nec}}^{(k)}. \quad (37)$$

If we let Λ^α denote the smallest coordinate-convex set containing $c(\mathcal{R}^\alpha)$, then by letting $k\to\infty$, $\Lambda_{\text{suf}}^{(k)}\uparrow\Lambda^\alpha$ and $\Lambda_{\text{nec}}^{(k)}\downarrow\Lambda^\alpha$, since $\mathcal{R}^{(k)}\uparrow\mathcal{R}^\alpha$ and $\mathcal{R}^\alpha$ is continuous. Hence, when $k\to\infty$, the gap between the sufficient and necessary stability conditions vanishes, and the exact stability region is given as Λ^α. □

D. Proofs of Corollaries 3 and 4

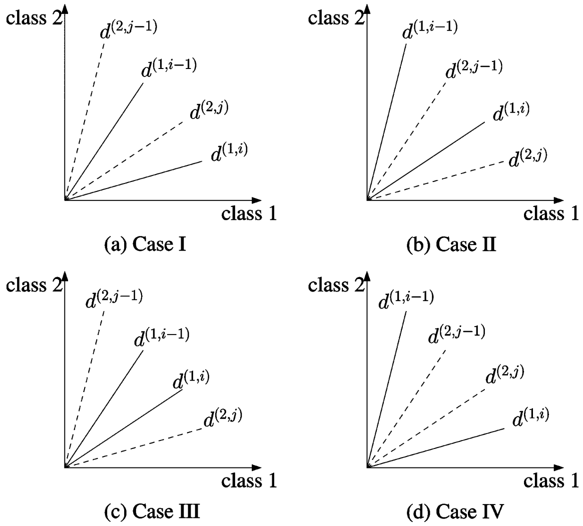
Proof of Corollary 3: We let $\boldsymbol{\phi}^{(i,\alpha)}(\boldsymbol{n})$ denote the α-fair allocated rate vector at state $\boldsymbol{n}$ when $\mathcal{R}(t)=\mathcal{R}_i$. First note that we can without loss of generality assume that every rate region is associated with equal probability (just scaling the rate regions). Moreover, another scaling argument allows to consider that the vectors of $\overline{\mathcal{R}_\alpha}$ are $\boldsymbol{\phi}^{(1,\alpha)}(\boldsymbol{n})+\boldsymbol{\phi}^{(2,\alpha)}(\boldsymbol{n})+\cdots+\boldsymbol{\phi}^{(M,\alpha)}(\boldsymbol{n})$. We start by discrete approximation of the rate regions. Consider a sequence of systems where the kth system has time-varying, finite, and discrete rate regions such that $\mathcal{R}_i^{(k)}$ is a subset of $\partial\mathcal{R}_i$, for all $i\in\mathcal{I}$. In particular, each discrete $\mathcal{R}_i^{(k)}=\{A^{(i,1)},\ldots,A^{(i,k)}\}$ is a Pareto-type set. The considered sequence is such that $\mathcal{R}_i^{(k)}\uparrow\partial\mathcal{R}_i$ as $k\to\infty$. Now let $\mathcal{R}_1^{(k)}=\{A^{(1,1)},\ldots,A^{(1,k)}\}$ satisfy the following.

(i) The allocated rate vectors are bordered from the left-top to the right-bottom in the sense that for all $1\le j\le k$, $A_1^{(1,j)}<A_1^{(1,j+1)}$ and $A_2^{(1,j)}>A_2^{(1,j)}$.

(ii) The allocated rate vectors are equally spaced by h_k w.r.t. class 1 coordinate, i.e., $A_1^{(1,i)}-A_1^{(1,i-1)}=h_k$.

(iii) $h_k\to 0$, as $k\to\infty$.

Thus, the rate vectors in $\mathcal{R}_i^{(k)}=a_i\times\mathcal{R}_1^{(k)}$ also satisfy *(i)*, *(ii)*, and *(iii)* but with equal space $a_1^{(i)}h_k$ on the class 1 axis. Let $\Lambda_\alpha^{(k)}$ denote the stability region associated with rate regions $\mathcal{R}_1^{(k)},\ldots,\mathcal{R}_M^{(k)}$, then $\Lambda_\alpha^{(k)}\uparrow\overline{\Lambda_\alpha}$ as $k\to\infty$. Now we proceed with the proof by induction. First, we consider $M=2$, and let $\mathcal{R}_2=a\times\mathcal{R}_1$ where $a_2>a_1$. By the monotone cone policy described in Section III-A, let $\mathcal{C}^{(1,i)}$ denote the set of



(a) Case I (b) Case II

(c) Case III (d) Case IV

Fig. 13. Cones and boundaries as $A^{(1,i)} + A^{(2,j)}$ is allocated.

states $\boldsymbol{n}$ when $A^{(1,i)} \in \mathcal{R}_1^{(k)}$ is allocated, then $\mathcal{C}^{(1,i)} \bigcap \mathcal{C}^{(1,i+1)}$ is a line containing $(0,0)$. Let $d^{(1,i)}$ denote the tangent of the angle between the line $\mathcal{C}^{(1,i)} \bigcap \mathcal{C}^{(1,i+1)}$ and the class–2 axis, then by [7], for $1 \leq i \leq k-1$

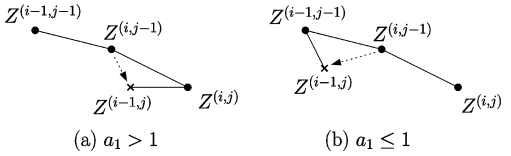
$$d^{(1,1)} \leq d^{(1,2)} \leq \cdots d^{(1,k-1)}.$$

Similar properties hold for $\mathcal{C}^{(2,i)}$, and we have

$$d^{(2,i)} = \left(\frac{a_2}{a_1}\right)^{(1-\alpha)/\alpha} d^{(1,i)}. \tag{38}$$

The rate vector $Z^{(i,j)} \triangleq A^{(1,i)} + A^{(2,j)}$ is allocated, if there exists some state $\boldsymbol{n}$, for which $A^{(1,i)}$ is allocated when the rate region is $\mathcal{R}_1^{(k)}$ and $A^{(2,j)}$ when it is $\mathcal{R}_2^{(k)}$. When $Z^{(i,j)}$ is actually allocated for some states, then $\mathcal{C}^{(1,i)} \bigcap \mathcal{C}^{(2,j)} \neq \emptyset$, which implies $d^{(2,j-1)} < d^{(1,i)}$ and $d^{(2,j)} > d^{(1,i-1)}$. We must have $j \geq i$ for $\alpha \geq 1$, and $j < i$ for $0 < \alpha < 1$.

Now we will slightly increase α and see how the set of allocated vectors changes. We concentrate the analysis on four consecutive vectors in the contour of the set of allocated vectors, i.e., $Z^{(i-1,j-1)}$, $Z^{(i,j-1)}$, $Z^{(i-1,j)}$, $Z^{(i,j)}$. There are four possible cases as shown in Fig. 13: Case I. $Z^{(i,j)}$, $Z^{(i-1,j)}$ are allocated; Case II. $Z^{(i,j)}$, $Z^{(i,j-1)}$; Case III. $Z^{(i,j)}$, $Z^{(i-1,j)}$; and Case IV. $Z^{(i,j)}$, $Z^{(i,j-1)}$. In this case, if α increases, the relative positions of $\mathcal{C}^{(1,i)}$ and $\mathcal{C}^{(2,i)}$ will change, which can be interpreted by a counterclockwise rotation of the boundary lines for $\mathcal{R}_2^{(k)}$ in Fig. 13. The possible transitions are: IV → I, II → IV, II → III, II → I, III → I, in which only IV, II → I, II → III causes change in allocated rate vectors as $Z^{(i,j)}, Z^{(i,j-1)} \rightarrow Z^{(i,j)}, Z^{(i-1,j)}$. We also observe that during the time just before and after the transition happens $Z^{(i-1,j-1)}$ is also allocated. Fig. 14 shows that after the transition, the new allocated rate vector $Z^{(i-1,j)}$ is always below the line segment connecting $Z^{(i-1,j)}$ and $Z^{(i,j)}$, or $Z^{(i-1,j)}$ and $Z^{(i-1,j-1)}$. Here we let $\angle(Z^{(i-1,j)} - Z^{(i,j)})$ denote the tangent of the angle between the line segment $Z^{(i-1,j)} - Z^{(i,j)}$ and class 1 axis: if $a_1 > 1$, $Z_1^{(i,j)} > Z_1^{(i-1,j)} > Z_1^{(i,j-1)}$, as shown in Fig. 14(a), then



(a) $a_1 > 1$ (b) $a_1 \leq 1$

Fig. 14. The transition of allocated rate vectors when α increases.

$$\begin{aligned}\angle(Z^{(i,j-1)} - Z^{(i,j)}) &= \frac{a_2\left(A_2^{(1,j-1)} - A_2^{(1,j)}\right)}{a_1 h_k} & (39)\\ &> \frac{A_2^{(1,i-1)} - A_2^{(1,i)}}{h_k} & (40)\\ &= \angle(Z^{(i-1,j)} - Z^{(i,j)}) & (41)\end{aligned}$$

as $A_2^{(1,j-1)} - A^{(1,j)} > A_2^{(1,i-1)} - A^{(1,i)}$ by the convexity of $\mathcal{R}_1$ when $j \geq i$ for $\alpha \geq 1$. Thus, $Z^{(i-1,j)}$ is below the line segment connecting $Z^{(i,j)}$, $Z^{(i,j-1)}$. Similarly, if $a_1 \leq 1$, $Z_1^{(i-1,j)} \geq Z_1^{(i,j-1)} > Z_1^{(i-1,j-1)}$, as shown in Fig. 14(b), then

$$\begin{aligned}\angle(Z^{(i-1,j-1)} - Z^{(i-1,j)}) &= \frac{a_2\left(A_2^{(1,j-1)} - A_2^{(1,j)}\right)}{a_1 h_k} & (42)\\ &> \frac{A_2^{(1,i-1)} - A_2^{(1,i)}}{h_k} & (43)\\ &= \angle(Z^{(i-1,j-1)} - Z^{(i,j-1)})\end{aligned}$$

which implies that $Z^{(i-1,j)}$ is below the line segment connecting $Z^{(i-1,j-1)}$ and $Z^{(i,j-1)}$. $Z^{(i-1,j)}$ is below the line segment connecting $Z^{(i-1,j-1)}$and $Z^{(i,j-1)}$. Hence, as α increases, the contour of the new allocated vectors is always below the previous one. Letting $k \rightarrow \infty$, we conclude that the stability region $\overline{\Lambda_\alpha}$ is decreasing as α increases if $\alpha \geq 1$.

Note that if $\partial\mathcal{R}_1$ and $\partial\mathcal{R}_2$ are hyperplanes (i.e., line segments in $\mathbb{R}_+^2$), which means we always have $A_2^{(1,j-1)} - A_2^{(1,j)} = A_2^{(1,i-1)} - A_2^{(1,i)}$ for any i, j, then the above arguments hold for all $\alpha > 0$. Hence, in this case $\overline{\Lambda_\alpha}$ decreases as α increases for all $\alpha > 0$.

Now if we let $M = 3$, then the allocated rate vector can be viewed as $Z^{(i_1,i_2,i_3)} = Z^{(i_1,i_2)} + A^{(3,i_3)}$. Notice that $\mathcal{C}^{(i_1,i_2,i_3)} = \mathcal{C}^{(i_1,i_2)} \bigcap \mathcal{C}^{(3,i_3)}$, and the boundaries of $\mathcal{C}^{(i_1,i_2)}$ correspond to the boundary of either $\mathcal{C}^{(1,i)}$ or $\mathcal{C}^{(2,i)}$. By (24) and (38), if $Z^{(i_1,i_2,i_3)}$ is allocated, then $i_3 \geq i_2 \geq i_1$ for $\alpha \geq 1$ and $i_3 < i_2 < i_1$ for $0 < \alpha < 1$. Thus, it is reduced to an $M = 2$ problem with cones $\mathcal{C}^{(i_1,i_2)}$ and $\mathcal{C}^{(3,i_3)}$, respectively. Hence, when $M = 3$, the α–$\overline{\Lambda_\alpha}$ tradeoff in Corollary 3 is true. By induction, this argument can be developed for arbitrary M rate regions. □

Proof of Corollary 4: Following the proof of Corollary 3, we consider the scaled rate regions such that each rate region is chosen with equal probability. Also, by induction, we first consider the case $M = 2$, where $\mathcal{R}_2 = a \times \mathcal{R}_1$ with $a_2 > a_1$, and consider the discrete approximation of rate regions $\mathcal{R}_1^{(k)}$, $\mathcal{R}_2^{(2)}$. By (38), when $\alpha \rightarrow 0$, $d^{(2,i)}/d^{(1,i)} \rightarrow \infty$, then all the

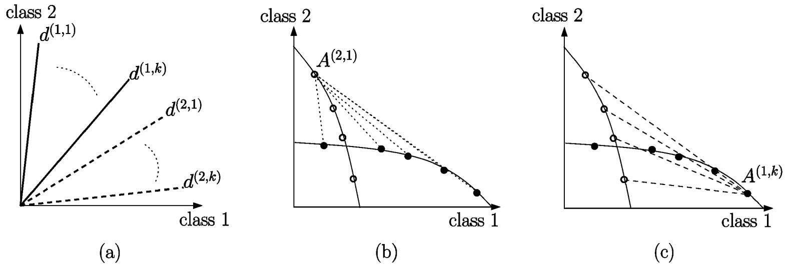


Fig. 15. When $\alpha \to 0$, (a) the cones and boundaries; (b)(c) the allocated rate vectors in $\mathcal{R}_1^{(k)}$ and $\mathcal{R}_2^{(k)}$.

boundaries of cones associated with $\mathcal{R}_1^{(k)}$ are above those associated with $\mathcal{R}_2^{(k)}$. Fig. 15(a) shows the relative positions of $\mathcal{C}^{(1,i)}$ and $\mathcal{C}^{(2,i)}$ in this case. Then the allocated rate vectors are $A^{(1,k)} + A^{(2,i)}$ and $A^{(1,i)} + A^{(2,1)}$ for all $1 \leq i \leq k$, as shown in Fig. 15(b) and (c). Notice that when $a_2 > a_1$, among all the possible combinations of rate vectors of the two rate regions, the vectors $A^{(1,k)} + A^{(2,i)}$ have maximum class-2 coordinates, and the vectors $A^{(1,i)} + A^{(2,1)}$ have maximum class-1 coordinates, for all $1 \leq i \leq k$, due to the convexity of $\mathcal{R}_1$, $\mathcal{R}_2$. This also implies that when $h_k \to 0$, the allocated vectors of $\mathcal{R}_1^{(k)}$, $\mathcal{R}_2^{(2)}$ when $\alpha \to 0$ are on the boundary of $\overline{R} = \mathcal{R}_1 + \mathcal{R}_2$. This finally indicates that the maximum stability region $\overline{R}$ can be achieved as $\alpha \to 0$.

For $M = 3$, $Z^{(i_1,i_2,i_3)} = Z^{(i_1,i_2)} + A^{(3,i_3)}$. By (24), when $\alpha \to 0$, all the boundaries of cones associated with $\mathcal{R}_1^{(k)}$ and $\mathcal{R}_2^{(k)}$ are above those associated with $\mathcal{R}_3^{(k)}$. The allocated vectors are $Z^{(i,j)} + A^{(3,1)}$ and $Z^{(k,k)} + A^{(3,i)}$, where $Z^{(i,j)}$ denotes any possible "matched" allocated vectors when $M = 2$. Hence, the allocated vectors for $M = 3$ are on the boundary of $\overline{R} = \mathcal{R}_1 + \mathcal{R}_2 + \mathcal{R}_3$ by the convexity of $\mathcal{R}_1$, $\mathcal{R}_2$, and $\mathcal{R}_3$. By induction, this argument can be developed for $M = 4, 5, \ldots$. Therefore, when $\alpha \to 0$, the maximum rate region $\overline{\mathcal{R}}$ is achieved. □

References

[1] M. Armony and N. Bambos, "Queueing dynamics and maximal throughput scheduling in switched processing systems," *Queueing Syst.: Theory and Applic.*, vol. 44, no. 3, pp. 209–252, 2003.

[2] P. Bender, P. Black, M. Grob, R. Padovani, N. Sindhushayana, and A. Viterbi, "CDMA/HDR: A bandwidth-efficient high-speed wireless data service for nomadic users," *IEEE Commun. Mag.*, vol. 38, no. 4, pp. 70–77, 2000.

[3] T. Bonald, S. Borst, N. Hegde, and A. Proutière, "Wireless data performance in multicell scenarios," in *Proc. ACM Sigmetrics/Performance*, New York, 2004, pp. 378–380.

[4] T. Bonald, S. Borst, and A. Proutière, "How mobility impacts the flow-level performance of wireless data networks," in *Proc. IEEE Infocom*, Hong Kong, Jun. 2004, pp. 1872–1881.

[5] T. Bonald and L. Massoulié, "Impact of fairness on internet performance," in *Proc. ACM Sigmetrics/Performance*, Cambridge, MA, Jun. 2001, pp. 82–91.

[6] T. Bonald, L. Massoulié, A. Proutière, and J. Virtamo, "A queueing analysis of max-min fairness, proportional fairness and balanced fairness," *Queueing Syst.: Theory and Applic.*, vol. 53, no. 1–2, pp. 65–84, 2006.

[7] T. Bonald and A. Proutière, "Flow-level stability of utility-based allocations for non-convex rate regions," in *Proc. 40th Conf. Information Sciences and Systems*, Princeton, NJ, Mar. 2006, pp. 327–332.

[8] S. Borst, N. Hegde, and A. Proutière, "Capacity of wireless networks with intra- and inter-cell mobility," in *Proc. IEEE Infocom*, Barcelona, Spain, Mar. 2006, pp. 1–12.

[9] S. Borst, M. Jonckheere, and L. Leskela, "Stability of parallel queueing systems with coupled rates," *Discr. Event Dyn. Syst.*, vol. 18, no. 4, pp. 447–472, 2008.

[10] M. Bramson, "Stability of networks for max-min fair routing," presented at the 13th INFORMS Applied Probability Conf., Ottawa, ON, Canada, 2005.

[11] M. Chiang, S. H. Low, A. R. Calderbank, and J. C. Doyle, "Layering as optimization decomposition: A mathematical theory of network architectures," *Proc. IEEE*, vol. 95, no. 1, pp. 255–312, Jan. 2007.

[12] M. Chiang, D. Shah, and A. Tang, "Stochastic stability of network utility maximization: General file size distribution," in *Proc. 44th Allerton Conf. Communication, Control and Computing*, Monticello, IL, Sep. 2006.

[13] J. G. Dai, "On positive harris recurrence of multiclass queueing networks: A unified approach via fluid limit models," *Ann. Appl. Probab.*, vol. 5, pp. 49–77, 1995.

[14] G. de Veciana, T. Lee, and T. Konstantopoulos, "Stability and performance analysis of networks supporting elastic services," *IEEE/ACM Trans. Netw.*, vol. 1, pp. 2–14, Feb. 2001.

[15] A. Eryilmaz and R. Srikant, "Joint congestion control, routing, and mac for stability and fairness in wireless networks," *IEEE J. Sel. Areas Commun.*, vol. 24, no. 8, pp. 1514–1524, Aug. 2007.

[16] V. Gambiroza, B. Sadeghi, and E. W. Knightly, "End-to-end performance and fairness in multihop wireless backhaul networks," in *Proc. ACM Mobicom*, Philadelphia, PA, 2004, pp. 287–301.

[17] H. C. Gromoll and R. Williams, Fluid Limit of a Network With Fair Bandwidth Sharing and General Document Size Distribution 2006, preprint.

[18] P. Gupta and A. Stolyar, "Optimal throughput allocation in general random-access networks," in *Proc. 38th Conf. Information Sciences and Systems*, Princeton, NJ, Mar. 2006, pp. 1254–1259.

[19] N. Hegde and A. Proutière, "Packet and flow level performance of wireless multihop networks," in *Proc. IEEE Globecom*, San Francisco, CA, Nov./Dec. 2006, pp. 1–5.

[20] M. Jonckheere and S. Borst, "Stability of multi-class queueing systems with state-dependent service rates," in *Proc. IEEE Value Tools Conf.*, Pisa, Italy, Oct. 2006.

[21] F. Kelly, A. Maulloo, and D. Tan, "Rate control in communication networks: Shadow prices, proportional fairness and stability," *J. Oper. Res. Soc.*, vol. 49, pp. 237–252, 1998.

[22] A. Lakshmikantha, C. L. Beck, and R. Srikant, "Connection level stability analysis of the internet using the sum of squares (SOS) techniques," in *Proc. 38th Conf. Information Sciences and Systems*, Princeton, NJ, 2004.

[23] X. Lin, N. B. Shroff, and R. Srikant, "On the connection-level stability of congestion-controlled communication networks," *IEEE Trans. Inf. Theory*, vol. 54, no. 5, pp. 2317–2338, May 2008.

[24] H. Luo, S. Lu, and V. Bharghavan, "A new model for packet scheduling in multihop wireless networks," in *Proc. IEEE Mobicom*, Boston, MA, 2000, pp. 76–86.

[25] W. Luo and A. Ephremides, "Stability of n interacting queues in random-access systems," *IEEE Trans. Inf. Theory*, vol. 45, no. 5, pp. 1579–1587, Jul. 1999.

[26] L. Massoulié, "Structural properties of proportional fairness: Stability and insensitivity," 2006, submitted for publication.

[27] L. Massoulié and J. Roberts, "Bandwidth sharing: Objectives and algorithms," *IEEE/ACM Trans. Netw.*, vol. 10, no. 3, pp. 320–328, Jun. 2002.

[28] J. Mo and J. Walrand, "Fair end-to-end window-based congestion control," *IEEE/ACM Trans. Netw.*, vol. 8, no. 5, pp. 556–567, Oct. 2000.

[29] W. Szpankowski, "Stability conditions for some multi-queue distributed systems: Buffered random access systems," *Ann. Appl. Probab.*, vol. 26, pp. 498–515, 1994.

[30] A. Tang, J. Wang, and S. Low, "Counter-intuitive throughput behaviors in networks under end-to-end control," *IEEE/ACM Trans. Netw.*, vol. 14, no. 2, pp. 355–368, Apr. 2006.

[31] X. Wang and K. Kar, "Cross-layer rate control for end-to-end proportional fairness in wireless networks with random access," in *Proc. MobiHoc*, Urbana-Champaign, IL, May 2005, pp. 157–168.

[32] H. Ye, "Stability of data networks under optimization-based bandwidth aladdress," *IEEE Trans. Aut. Contr.*, vol. 48, no. 7, pp. 1238–1242, Jul. 2003.

[33] H. Ye, J. Ou, and X. Yuan, "Stability of data networks: Stationary and bursty models," *Oper. Res.*, vol. 53, pp. 107–125, 2005.

Jiaping Liu received the bachelor's degree from Shanghai Jiao Tong University, Shanghai, China, in 2004, the M.A. degree from Princeton University, Princeton, NJ, in 2006, and the Ph.D. degree also from Princeton University in 2009, all in electrical engineering. Her Ph.D. dissertation focused on stochastic modeling and analysis of data networks, and its applications in wireless ad hoc networks, in particular the random access algorithm design of medium access control protocols.

She was a Research Assistant at Bell Laboratories in summer 2007 and a Visiting Scholar in the Department of Electrical Engineering at Stanford University, Stanford, CA, in fall 2007.

The Ph.D. work of Dr. Liu was supported in part by the Gordon Wu Fellowship of Princeton University.

Alexandre Proutière received the Ph.D. degree in mathematics from Ecole Polytechnique (Palaiseau, France) in 2003, graduated in mathematics from Ecole Normale Superieure (Paris), and qualified as an engineer at Ecole Nationale Superieure des Telecommunications (Paris).

He is a Researcher in the Systems and Networking group at Microsoft Research, Cambridge, U.K. His research interests are in the design and the performance evaluation of computer networks, with a specific interest in resource allocation and control in wireless systems. Before joining Microsoft in June 2007, he was a Senior Expert Researcher at France Telecom R&D and an Assistant Professor in the Computer Science Department of Ecole Normale Superieure (Paris, France).

Dr. Proutière (with Thomas Bonald), is the recipient of the best paper award at ACM Sigmetrics /Performance 2004.

Yung Yi (S'04–M'06) received the B.S. and M.S. degrees from the School of Computer Science and Engineering, Seoul National University, Seoul, Korea, in 1997 and 1999, respectively, and the Ph.D. degree from the Department of Electrical and Computer Engineering, University of Texas at Austin, in 2006.

From 2006 to 2008, he was a Postdoctoral Research Associate in the Department of Electrical Engineering, Princeton University, Princeton, NJ. He is now an Assistant Professor with the School of Electrical Engineering and Computer Science, Korea Advanced Institute of Science and Technology (KAIST), Daejeon, Korea. His current research interests include design and analysis of computer networking and communication systems, especially congestion control, scheduling, and interference management, with applications in wireless *ad hoc* networks and broadband access networks, and future Internet evolution.

Mung Chiang (S'00–M'03–SM'08) received the B.S. (Honors) degree in electrical engineering and mathematics, and the M.S. and Ph.D. degrees in electrical engineering from Stanford University, Stanford, CA, in 1999, 2000, and 2003, respectively.

He was an Assistant Professor at Princeton University, Princeton, NJ, from 2003 to 2008. He is now an Associate Professor of Electrical Engineering, and an Affiliated Faculty of Applied and Computational Mathematics and of Computer Science, at Princeton University. His research areas include optimization, distributed control, and stochastic analysis of communication networks, with applications to the Internet, wireless networks, broadband access networks, and content distribution.

Prof. Chiang received the Presidential Early Career Award for Scientists and Engineers 2008 from the White House, the Young Investigator Award 2007 from ONR, TR35 Young Innovator Award 2007 from Technology Review, Young Researcher Award Runner-up 2004–2007 from Mathematical Programming Society, CAREER Award 2005 from NSF, as well as Frontiers of Engineering Symposium participant 2008 from NAE and SEAS Teaching Commendation 2007 from Princeton University. He was a Princeton University Howard B. Wentz Junior Faculty and a Hertz Foundation Fellow. His paper awards include ISI citation Fast Breaking Paper in Computer Science, IEEE INFOCOM Best Paper Finalist, and IEEE GLOBECOM Best Student Paper. His guest and associate editorial services include IEEE/ACM TRANSACTIONS ON NETWORKING, IEEE TRANSACTIONS ON INFORMATION THEORY, IEEE JOURNAL ON SELECTED AREAS IN COMMUNICATIONS, IEEE TRANSACTIONS ON COMMUNICATIONS, IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS, and *Journal Optimization and Engineering*, and he cochaired the 38th Conference on Information Sciences and Systems.

H. Vincent Poor (S'72–M'77–SM'82–F'87) received the Ph.D. degree in electrical engineering and computer science from Princeton University, Princeton, NJ, in 1977.

From 1977 until 1990, he was on the faculty of the University of Illinois at Urbana-Champaign. Since 1990, he has been on the faculty at Princeton University, where he is the Dean of Engineering and Applied Science, and the Michael Henry Strater University Professor of Electrical Engineering. His research interests are in the areas of stochastic analysis, statistical signal processing and their applications in wireless networks, and related fields. Among his publications in these areas are the recent books *MIMO Wireless Communications* (Cambridge University Press, 2007), coauthored with Ezio Biglieri *et al.* and *Quickest Detection* (Cambridge University Press, 2009), coauthored with Olympia Hadjiliadis.

Dr. Poor is a member of the National Academy of Engineering, a Fellow of the American Academy of Arts and Sciences, and a former Guggenheim Fellow. He is also a Fellow of the Institute of Mathematical Statistics, the Optical Society of America, and other organizations. In 1990, he served as President of the IEEE Information Theory Society, and in 2004–2007 as the Editor-in-Chief of these TRANSACTIONS. He is the recipient of the 2005 IEEE Education Medal. Recent recognition of his work includes the 2007 IEEE Marconi Prize Paper Award, the 2007 Technical Achievement Award of the IEEE Signal Processing Society, and the 2008 Aaron D. Wyner Distinguished Service Award of the IEEE Information Theory Society.