Limit theorems for Markovian bandwidth-sharing networks with rate constraints

Citation for published version (APA):

DOI:
10.1287/opre.2014.1321

Document status and date:
Published: 01/01/2014

Document Version:
Publisher’s PDF, also known as Version of Record (includes final page, issue and volume numbers)

Please check the document version of this publication:

• A submitted manuscript is the version of the article upon submission and before peer-review. There can be important differences between the submitted version and the official published version of record. People interested in the research are advised to contact the author for the final version of the publication, or visit the DOI to the publisher’s website.
• The final author version and the galley proof are versions of the publication after peer review.
• The final published version features the final layout of the paper including the volume, issue and page numbers.

Link to publication

General rights
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

• Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
• You may not further distribute the material or use it for any profit-making activity or commercial gain
• You may freely distribute the URL identifying the publication in the public portal.

If the publication is distributed under the terms of Article 25fa of the Dutch Copyright Act, indicated by the “Taverne” license above, please follow below link for the End User Agreement:
www.tue.nl/taverne

Take down policy
If you believe that this document breaches copyright please contact us at:
openaccess@tue.nl
providing details and we will investigate your claim.

Download date: 17. Feb. 2019
Limit Theorems for Markovian Bandwidth-Sharing Networks with Rate Constraints

Josh Reed, Bert Zwart

To cite this article:

Full terms and conditions of use: http://pubsonline.informs.org/page/terms-and-conditions

This article may be used only for the purposes of research, teaching, and/or private study. Commercial use or systematic downloading (by robots or other automatic processes) is prohibited without explicit Publisher approval, unless otherwise noted. For more information, contact permissions@informs.org.

The Publisher does not warrant or guarantee the article’s accuracy, completeness, merchantability, fitness for a particular purpose, or non-infringement. Descriptions of, or references to, products or publications, or inclusion of an advertisement in this article, neither constitutes nor implies a guarantee, endorsement, or support of claims made of that product, publication, or service.

Copyright © 2014, INFORMS

Please scroll down for article—it is on subsequent pages

INFORMS is the largest professional society in the world for professionals in the fields of operations research, management science, and analytics. For more information on INFORMS, its publications, membership, or meetings visit http://www.informs.org
Limit Theorems for Markovian Bandwidth-Sharing Networks with Rate Constraints

Josh Reed
Stern School of Business, New York University, New York, New York 10012, jreed@stern.nyu.edu
Bert Zwart
Centrum Wiskunde & Informatica (CWI), Amsterdam, The Netherlands, Bert.Zwart@cwi.nl

Bandwidth-sharing networks provide a natural modeling framework for describing the dynamic flow-level interaction among elastic data transfers in computer and communication systems, and can be used to develop traffic pricing/charging mechanisms. At the same time, such models are exciting from an operations research perspective because their analysis requires techniques from stochastic modeling and optimization.

In this paper, we develop a framework to approximate bandwidth-sharing networks under the assumption that the number of users as well as the capacities of the system are large, and the assumption that the traffic that each user is allowed to submit is bounded above by some rate, which is standard in practice. We also assume that customers on each route in the network abandon according to exponential patience times. Under Markovian assumptions, we develop fluid and diffusion approximations, which are quite tractable: for most parameter combinations, the invariant distribution is multivariate normal, with mean and diffusion coefficients that can be computed in polynomial time as a function of the size of the network.

Subject classifications: networks; limit theorems; Markovian.
Area of review: Stochastic Models.
History: Received December 2010; revisions received March 2013, December 2013; accepted July 2014. Published online in Articles in Advance October 23, 2014.

1. Introduction

Bandwidth-sharing networks considered by Massoulié and Roberts (1999) and Roberts and Massoulié (1998) provide a natural extension for modeling the dynamic interaction among competing elastic flows that traverse several links along their source-destination paths in a network. They offer insight into the complex behavior of communication networks and have also recently been suggested as a tool in analyzing problems in road traffic (see, for instance, Kelly and Williams 2010). From an operations research (OR) perspective, bandwidth-sharing networks are exciting since their static behavior is governed by nonlinear optimization problems, while understanding their dynamics requires a separate set of OR tools, namely, stochastic models.

Contemporary research has devoted a significant amount of effort to analyzing bandwidth-sharing networks and to deriving stability conditions for bandwidth-sharing networks. This question is still not settled, in general, and it is not the subject matter of the present paper although a variety of results may be found in De Veciana et al. (1999, 2001), Bonald and Massoulié (2001), Mo and Walrand (2000), Massoulié (2007), Gromoll and Williams (2008), and Chiang et al. (2006). Another significant issue, which is more central to the present paper, is concerned with second order phenomena, i.e., methods to evaluate the performance of bandwidth-sharing models. For the right combination of network topology and bandwidth-sharing policy, it is possible to show that the steady-state distribution of the network not only exists, but is of product form and is insensitive with respect to the flow size distribution. In some cases, it is even possible to derive necessary and sufficient conditions for steady-state distributions of this type to exist. This work is well summarized in Bonald et al. (2006).

In general, such nice structure on the network topology and bandwidth-sharing policy, as mentioned above, cannot be expected to hold and one has to resort to approximations. Fundamental papers on fluid limit approximations for bandwidth-sharing networks are Kelly and Williams (2004) and Gromoll and Williams (2009). Properties of overloaded bandwidth-sharing networks have subsequently been derived by Borst et al. (2014) and Egorova et al. (2007). A diffusion approximation for bandwidth-sharing networks was derived in Kang et al. (2009). Ye and Yao (2008, 2010) considered diffusion approximations of some bandwidth-sharing networks where the service discipline per class is FIFO rather than PS, which coincides with regular bandwidth-sharing networks in the case of exponential flow sizes. As we see it, the main message of these works is that the performance of bandwidth-sharing networks in heavy traffic can sometimes be described by a linear transformation of a vector of independent, exponential random variables. Although this line of research is exciting and still...
ongoing, this computationally tractable insight seems limited to specific network topologies and bandwidth-sharing mechanisms.

The current paper proposes a different perspective leading to another class of tractable approximations, namely, multivariate normal approximations. As we shall show, such approximations arise naturally from the observation that overall system capacity and individual user download speeds may be of different orders of magnitude. It is common in applications (see, for instance, Bonald and Proutière 2004) that network capacity is measured in gigabits or terabits per second, whereas individual user maximal download speeds are measured in megabits. In the present paper, we assume that individual user download speeds are bounded above by some maximum, whereas overall system capacity may be arbitrarily large. This stands apart from the above-mentioned works in which system capacities and user download speeds are assumed to be comparable. A consequence of our limit on the individual download speed is that a significant number of users are required to saturate a link. As a result, we consider a system with large arrival rates and system capacities and view the system on a fixed time scale, whereas many of the above-mentioned works focus on the large time properties of a network with fixed arrival rates and capacities.

Our framework can be seen as an extension of the many-server scaling found in the literature on call center approximations. We refer to Gans et al. (2003) for a survey on this literature and note that the results in this paper for the most simple case of a single node/class network reduce to the classical diffusion approximation of Halfin and Whitt (1981) for many-server queues. In fact, the model we consider is a highly nontrivial example of a Markovian service network, considered by Mandelbaum et al. (1998). Unfortunately, we could not directly fit our assumptions into theirs, so we verify the necessary details from scratch.

In the call center queueing literature, one often makes a distinction between several qualitatively different regimes: the quality driven (QD), efficiency driven (ED), and quality and efficiency driven regime (QED). We shall see in the present paper, in a multiclass multinode bandwidth-sharing network, it is not a priori clear in which regime a class will operate from the outset. This is actually determined endogenously (rather than exogenously, which is the case in simple call center models) through the dynamics of the bandwidth-sharing allocation algorithm. We provide a key optimization problem for a model with user impatience that determines whether in steady-state (on a fluid scale) the maximal service rate of a class of users will be met or not.

To the best of our knowledge, the first paper to consider diffusion approximations of bandwidth-sharing networks with rate constraints is Ayesta and Mandjes (2009). This work begins with existing explicit scheduling policies without individual capacity constraints, and then truncates the capacity constraints at the individual maxima. Our allocation policies take a more integrated approach, allowing users that operate below maximal capacity to take up bandwidth that is not used by other (rate-constrained) users, so that bandwidth allocations are Pareto optimal. Moreover, rather than formulating the fluid and diffusion approximations directly, we rigorously establish that these approximations arise in a large-capacity scaling. We also do not need to make additional assumptions that yield an explicit representation of the bandwidth-sharing allocation function. In fact, all that is necessary is a directional differentiability property that is established in complete generality in §3 of this paper, using results from the sensitivity analysis of nonlinear programs as developed in Bonnans and Shapiro (2000). This directional differentiability result (we actually give a necessary and sufficient condition for differentiability) is one of the main technical results of the paper. In particular, we hope that the general methodology we use (which does not seem to be well known in the applied probability community) will avoid the use of laborious bare hand calculations in the future. We believe that this connection between stochastic networks and continuous optimization will be interesting for other works as well.

The limit theorems obtained in this paper are fluid and diffusion limits under Markovian assumptions on the interarrival times, information sizes, and patience times of users. The resulting steady-state diffusion approximations often yield a multivariate normal law, where the means and covariances can be computed by, respectively, a concave programming problem with a polyhedral capacity set, and a set of linear equations. This results in a computational procedure that has complexity, which is polynomial in the size of the network, and is, in principle, valid for any network topology and a large class of utility-based bandwidth allocation mechanisms.

During the preparation of the final version of this paper, our main fluid limit result (Theorem 4.1) was extended to general distributions in Remerova et al. (2014). The proof of the fluid limit result of Remerova et al. (2014) requires the machinery of measure-valued processes, while the proofs of the fluid and diffusion limit results in the present paper work in a finite-dimensional setting. To avoid unnecessary overlap with Remerova et al. (2014), the discussion at several points of the main body of this paper has been shortened and we simply reference Remerova et al. (2014).

The remainder of this paper is organized as follows. A model description is provided in §2. Section 3 contains a detailed sensitivity analysis of the bandwidth allocation function. Fluid and diffusion approximations are presented in §4. Section 5 focuses on invariant points for the fluid model. In particular, we focus on a model with user impatience, for which we establish uniqueness of an invariant point, and we provide sufficient conditions for differentiability of the bandwidth allocation function at this invariant point, leading to a multivariate normal law for the diffusion approximation. In §6, we illustrate our results with some examples. Proofs can be found throughout the paper; some of the more technical proofs of §3 are deferred to §7.
while the proofs of §4 may be found in an accompanying electronic companion (available as supplemental material at http://dx.doi.org/10.1287/opre.2014.1321).

1.1. Notation

All random variables and processes in this paper are assumed to be defined on a common probability space \((\Omega, \mathcal{F}, P)\). All vectors \(x \in \mathbb{R}^d\) are assumed to be column vectors and we denote by \(\|x\|\) the Euclidean norm of \(x\). We denote by \(x^t\) the transpose of \(x\). We let \(D([0, \infty), \mathbb{R}^d)\) denote the space of right continuous with left limits functions defined on \([0, \infty)\) and taking values in \(\mathbb{R}^d\) for \(d \geq 1\). Unless otherwise stated, all stochastic processes in this paper are assumed to be measurable maps from \((\Omega, \mathcal{F})\) to \((D([0, \infty), \mathbb{R}^d), d_{\text{weak}})\), where the choice of dimension \(d \geq 1\) will be clear from context and \(d_{\text{weak}}\) is the metric defined in (16.4) of Billingsley (1999). We recall from Theorem 16.3 of Billingsley (1999) that the space \((D([0, \infty), \mathbb{R}^d), d_{\text{weak}})\) is separable and complete. Moreover, we recall that if \(x \in D([0, \infty), \mathbb{R}^d)\) is a continuous function, then \(x^n \to x\) in \((D([0, \infty), \mathbb{R}^d), d_{\text{weak}})\) if and only if \(\sup_{0 \leq t < T} \|x^n(t) - x(t)\| \to 0\) as \(n \to \infty\) for each \(T \geq 0\). We denote by \(D^2([0, \infty), \mathbb{R}^d)\) the product space \(D([0, \infty), \mathbb{R}^d) \times D([0, \infty), \mathbb{R}^d)\), which we endow with the maximum metric. We denote by \(\varepsilon = (t, t \geq 0)\) the identity function.

2. The Model

In this section, we provide the details of the bandwidth-sharing model that we consider for the remainder of the paper. We begin by describing what is referred to as the network topology. A bandwidth-sharing network consists of \(J \geq 1\) resources and \(I \geq 1\) routes. Each resource is given an index \(j \in \{1, 2, \ldots, J\}\) and a route \(i = 1, \ldots, I\) is a nonempty subset of \(\{1, 2, \ldots, J\}\). We then define the \(J \times I\) incidence matrix \(A\) to be such that \(A_{ji} = 1\) if resource \(j\) is an element of route \(i\) and zero otherwise. Intuitively, one may think of a resource \(j\) as a server on a network and a route \(i\) as a series of resources through which information is passed. Note, however, that a route is an unordered set and so we do not distinguish the order in which information is being passed through the resources. An example is provided in Figure 1.

A flow represents a specific transfer of information along a route. Each flow in the network is assigned a processing rate and the sum of all of the processing rates assigned to the flows on a particular route \(i\) is referred to as the bandwidth devoted to route \(i\), which we denote by \(A_i\). At the same time, each resource \(j\) is assigned a limited amount of total bandwidth, \(0 < C_j < \infty\), which it must distribute to each of the routes passing through it. We therefore obtain the matrix inequality \(\Lambda \leq C\), where \(\Lambda = (A_1, A_2, \ldots, A_J)\)' and \(C = (C_1, C_2, \ldots, C_J)\)' . In addition, we also assign to each route \(i\) a maximum rate \(0 < m_i \leq \infty\) at which flows on that route may be processed. In particular, if \(z_i\) represents the total number of flows on route \(i\), we require that \(\Lambda_i \leq m_i z_i\). Thus, letting \(I_m\) be the \(I \times 1\) diagonal matrix with the \(m_i\)'s on the diagonal, we have that in matrix notation \(\Lambda \leq I_m z\), where \(z = (z_1, z_2, \ldots, z_I)'\). Note also that, in general, we require \(C_j\) to be finite but allow for the possibility of \(m_i = \infty\).

We next proceed to describe the stochastic assumptions, which we place on our bandwidth-sharing network. At time \(t = 0\), we assume that there are \(Z(t)\) flows already present on route \(i\), each of whose unprocessed information sizes are i.i.d. (independent identically distributed) exponential random variables with mean \(1/\mu_i\). Next, for each \(i = 1, \ldots, I\), let \((E_i(t), t \geq 0)\) be a rate one Poisson process and let \(0 < \eta_i < \infty\) be the average arrival rate of flows to route \(i\). Also, assume that \((E_i(t), t \geq 0), i = 1, \ldots, I\) are independent of one another. The number of flows that have arrived externally to route \(i\) by time \(t \geq 0\) is then given by \(E_i(\eta_i t)\). Hence, flows arrive to route \(i\) according to a Poisson process with rate \(\eta_i\). Next, for each \(k \geq 1\), we assume that the information size of the \(k\)th flow to arrive to route \(i\) is an exponential random variable with mean \(1/\mu_i\). Moreover, we assume that the information sizes of flows are independent of one another both within a specific route and between routes as well. We also assume that each flow arriving to route \(i\) is impatient and is only willing to wait an exponentially distributed amount of time with rate \(0 \leq \gamma_i < \infty\) to have all of its information processed before abandoning from the network. Note that this implies that some flows may depart prematurely from the network before having all of their information processed. This does not happen if \(\gamma_i = 0\), in which case flows on route \(i\) are not impatient.

Now, let \(Z_i(t)\) be the number of flows on route \(i\) at time \(t \geq 0\) and set \(Z(t) = (Z_1(t), Z_2(t), \ldots, Z_I(t))'\). We refer to \(Z(t)\) as the user population vector at time \(t\). Next, we suppose that there exists some bandwidth allocation policy \(\Lambda: \mathbb{R}^I_+ \mapsto \mathbb{R}^J_+\) such that if the user population vector at time \(t\) is given by \(Z(t)\), then the bandwidth devoted to route \(i\) by each resource on route \(i\) is given by \(\Lambda_i(Z(t))\). Moreover, we assume that each of the \(Z_i(t)\) flows along route \(i\) must equitably share the total bandwidth allocated...
to the route. Hence, since the total bandwidth assigned to route \( i \) is \( \lambda_i(Z(t)) \), this implies that the bandwidth allocated to each of the \( Z_i(t) \) flows on route \( i \) is given by \( \lambda_i(Z(t))/Z_i(t) \), which we define to be 0 if \( Z_i(t) = 0 \). In §3, we provide a detailed prescription of how the bandwidth allocation policy \( \Lambda \) is determined.

We now conclude this section by providing a sample-path equation describing the dynamics of the user population process \( Z = (Z_i(t), t \geq 0) \). This equation will turn out to be useful in our proofs that follow. First, note that, since the information sizes of each flow on route \( i \) are assumed to be i.i.d. exponential random variables with rate \( \mu_i \), it follows that the departure process of service completions of flows of type \( i \) from the network is a doubly stochastic Poisson process with an instantaneous rate at time \( t \geq 0 \) given by \( Z_i(t)(\lambda_i(Z(t))/Z_i(t))\mu_i = \lambda_i(Z(t))\mu_i \). Next, since the patience times of flows on route \( i \) are assumed to be i.i.d. exponential random variables with rate \( \gamma_i \), it follows that the departure process of abandonments of flows of type \( i \) from the network is a doubly stochastic Poisson process with instantaneous rate at time \( t \geq 0 \) given by \( \gamma_iZ_i(t) \). Thus, letting \( N_i^t = (N_i^t(t), t \geq 0) \) and \( N_i^n = (N_i^n(t), t \geq 0), i = 1, \ldots, I \) be independent, rate one Poisson processes, it follows that the user population process \( Z \) is given by the solution to the system of equations,

\[
Z_i(t) = Z_i(0) + E_i(\eta_i t) - N_i^t\left( \mu_i \int_0^t \lambda_i(Z(s)) \, ds \right) - N_i^n\left( \gamma_i \int_0^t Z_i(s) \, ds \right), \quad t \geq 0
\]

for \( i = 1, \ldots, I \). Note that (1) represents a system of \( I \) equations and it is straightforward to show that there exists a unique solution \( Z \).

3. The Bandwidth Allocation Mechanism and Its Properties

In this section, we provide a more detailed description of how the bandwidth allocation function \( \Lambda \) of §2 is determined. In particular, we cast \( \Lambda \) as the solution to an optimization problem and we show how one may compute the directional derivative of \( \Lambda \) in the interior of its domain. We then proceed to provide a result on the Lipschitz continuity of \( \Lambda \) as well as providing conditions under which \( \Lambda \) is fully differentiable. The results of this section are interesting in their own right, while also serving as crucial elements of our fluid and diffusion approximations in the sections that follow. Our main reference for this section is Bonnans and Shapiro (2000) and we note that all of the proofs for the results in this section may be found in §7.

We begin by assigning to each flow on a route \( i = 1, \ldots, I \) a utility function \( U_i: (0, \infty) \rightarrow \mathbb{R} \). The utility of a flow is assumed to be a function of the amount of bandwidth allocated to that flow. In particular, if \( \lambda_i \) is the total amount of bandwidth that has been allocated to route \( i \), and if there are \( z_i > 0 \) flows on route \( i \) at a given point in time, then each flow will receive \( \lambda_i/z_i \) units of bandwidth, and consequently, each flow will have a utility of \( U_i(\lambda_i/z_i) \). The total utility across all flows on route \( i \) will therefore be given by \( z_iU_i(\lambda_i/z_i) \). We assume that \( U_i \) is a strictly increasing, strictly concave, twice differentiable function on \((0, \infty)\) such that \( \lim_{x \to 0} U_i'(x) = \infty \). We also make the convention that \( z_iU_i(\lambda_i/z_i) = 0 \) for all \( x > 0 \) if \( z_i = 0 \). Finally, we note that one important family of utility functions are the weighted \( \alpha \)-fair utility functions, which are given by the functional form \( U_i'(x) = \kappa_i x^{-\alpha} \) with \( \kappa_i, \alpha > 0 \).

We now proceed to construct our bandwidth allocation function \( \Lambda \) as the solution to a nonlinear programming problem. In particular, given a fixed number of flows \( z = (z_1, \ldots, z_I) \) on each route and a set of utility functions \( U_i, i = 1, \ldots, I \), the system will attempt to allocate bandwidth in such a manner to maximize the total utility of all flows on the network, subject to the capacity and individual rate constraints described in §2. Mathematically speaking, the bandwidth allocation \( \Lambda = (\Lambda_1, \ldots, \Lambda_I) \) is given by the solution to the following global utility maximization problem:

\[
(P_z) \quad \max_{\Lambda \in \mathbb{R}^I} \sum_{i=1}^I z_iU_i\left( \frac{\lambda_i}{z_i} \right)
\]

subject to \( \Lambda \leq C \),

\[
\Lambda \leq \Lambda_{\mu} z,
\]

\[
\Lambda \geq 0.
\]

Note that since the criterion function in the above nonlinear program is strictly concave, it follows that for each \( z = (z_1, \ldots, z_I) \) in the interior of the positive orthant, there exists a unique optimal bandwidth allocation, which we denote by \( \Lambda(z) = (\Lambda_1(z), \ldots, \Lambda_I(z)) \). Moreover, for such \( z \), it is straightforward from our assumptions on the utility functions that \( \Lambda(z) \) lives in the interior of the positive orthant, making the constraint \( \Lambda \geq 0 \) superfluous. Therefore we can and will ignore the constraint \( \Lambda \geq 0 \) whenever \( z \in (0, \infty)^I \).

Now, for each point \( z \in (0, \infty)^I \) in the interior of the positive orthant and for each direction \( d \in \mathbb{R}^I \), define

\[
H^d(z) = \lim_{t \downarrow 0} (\Lambda(z + td) - \Lambda(z))/t
\]

(2)
to be the directional derivative (in the direction \( d \)) of \( \Lambda(\cdot) \) at the point \( z \), assuming that this limit exists. We now proceed in Theorem 3.1 to show that our choice of the bandwidth allocation function \( \Lambda(\cdot) \), defined by \((P_z)\) above, leads to a directionally differentiable function at each point \( z \in (0, \infty)^I \) in each direction \( d \in \mathbb{R}^I \). Moreover, we also show that \( H^d(z) \) may be determined for each \( z \in (0, \infty)^I \) by solving a specific quadratic programming problem. To prove this result, we will specialize some of the more general results found in §5.2.3 of Bonnans and Shapiro (2000). In particular, our Theorem 3.1 may be viewed
as a relatively straightforward application of the perturbation analysis of nonlinear optimization problems in Banach spaces, a theory which is nowadays well developed. Still, techniques from this theory do not seem to have found widespread use in the applied probability community so far.

Before providing the statement of Theorem 3.1, we must first adopt some additional notation. For each \( z \in (0, \infty)^I \), let \( p(z) \) be a \( J \)-dimensional vector of Lagrange multipliers corresponding to the capacity constraints \( \Lambda \Lambda \leq C \), and let \( q(z) \) be an \( I \)-dimensional vector of Lagrange multipliers corresponding to the individual rate constraints \( \Lambda \leq \Gamma_i z \), and recall that \( p(z), q(z) \), and \( \Lambda(z) \) jointly form a solution to the Karush-Kuhn Tucker (KKT) conditions for \((P_i)\). That is, we have that
\[
(\Lambda(z) - C)p(z) = 0, \quad (\Lambda(z) - I_m z)q(z) = 0, \quad \text{and}
\]
\[
U_i'(\Lambda_i(z)/z_i) = q_i(z) + \sum_{j=1}^J p_j(z)A_{ji}
\]
for \( i = 1, \ldots, I \). One may also consult (3.8) on page 147 of Bonnans and Shapiro (2000) for this result. We now denote by \( \gamma(z) \) the set of all possible Lagrange multipliers of \((P_i)\).

Next, suppose that we pick a direction \( d = (d_1, \ldots, d_I) \in \mathbb{R}^I \). We will then denote by \((p^d(z), q^d(z))\), the specific set of Lagrange multipliers of \((P_i)\) that also solve the optimization problem
\[
\max_{(p, q) \in \gamma(z)} - \sum_{i=1}^I d_i m_i q_i,
\]
for motivation for this will be given in §7. Finally, we let \( \mathcal{J}(z) \) denote the set of active individual rate constraints of \((P_i)\), and we denote by \( \mathcal{J}(z) \) the set of active capacity constraints of \((P_i)\). Note that these sets do not depend on the choice of direction \( d \). The following is now our first main result concerning the directional differentiability of \( \Lambda(\cdot) \).

**Theorem 3.1.** Let \( z \in (0, \infty)^I \) and for each \( i = 1, \ldots, I \), define
\[
v_i(z) = \frac{1}{z_i} U_i'(\Lambda_i(z)/z_i) \quad \text{and} \quad u_i(z) = \frac{\Lambda_i(z)}{z_i} v_i(z).
\]
Then, \( \Lambda(\cdot) \) is directionally differentiable in any direction \( d \in \mathbb{R}^I \), and its directional derivative \( H^d(z) \) is the unique solution to the following quadratic programming problem:
\[
(D_{z,d}) \max_{h \in \mathbb{R}^I} -2 \sum_{i=1}^I u_i(z)d_i h_i + \sum_{i=1}^I v_i(z) h_i^2
\]
subject to \( (Ah)_j = 0 \) if \( p_j^d(z) > 0 \),
\[
(Ah)_j \leq 0 \quad \text{if} \quad (\Lambda(z))_j \leq C_j,
\]
\[
h_i = d_i m_i, \quad \text{if} \quad q_i^d(z) > 0
\]
\[
h_i \leq d_i m_i, \quad \text{if} \quad \Lambda_i(z) = m_i z_i.
\]
We remark that the above result would continue to hold if \( m_i = \infty \) for some \( i = 1, \ldots, I \), so that flows on some routes are not subject to individual rate constraints. Moreover, Theorem 3.1 would also continue to hold if, rather than assuming that \( A \) is an incidence matrix as described in §2, we simply assumed that the elements of \( A \) were nonnegative, or if \( A \) were assumed to be such that the interior of the polyhedral capacity set \( \{ \Lambda: \Lambda \Lambda \leq C, \Lambda \geq 0 \} \) is nonempty. This would imply the validity of several constraint qualifications, in particular, the Fromovitz-Mangasarian constraint qualification, see pages 71–72 of Bonnans and Shapiro (2000). Such extensions are relevant to cover more general network models involving multipath routing, see §5.5 of Kang et al. (2009).

We next proceed to show that \( \Lambda \) as defined by the solution to \((P_i)\) above is, in fact, a locally Lipschitz function on \((0, \infty)^I \). Results of a similar nature may also be found in Lemma A.3 of Kelly and Williams (2004) and Proposition 4.1 of Borst et al. (2014). These authors, respectively, established continuity and Lipschitz continuity of \( \Lambda(\cdot) \) defined above for special cases, with no rate constraints. That is, \( m_i = \infty \) for \( i = 1, \ldots, I \). The following result will also be helpful in §4 when proving our main fluid limit result, Theorem 4.1.

**Theorem 3.2.** The bandwidth allocation function \( \Lambda(\cdot) \) is locally Lipschitz on \((0, \infty)^I \). That is, for every compact subset \( \mathcal{C} \subset (0, \infty)^I \), there exists a constant \( K_{\mathcal{C}} \) such that
\[
\|\Lambda(x) - \Lambda(y)\| \leq K_{\mathcal{C}}\|x - y\| \quad \text{for all} \quad x, y \in \mathcal{C}.
\]

It turns out that some of our results in §4 will also require the stronger condition of \( \Lambda(\cdot) \) being differentiable at a point \( z \), which, in view of Theorem 3.1, is equivalent to \( H^d(z) \) being a linear function of \( d \). A sufficient condition for differentiability is given in the following theorem. Recall that the strict complementarity condition holds for \((P_i)\) if all of the active constraints of \((P_i)\) have strictly positive Lagrange multipliers, i.e., \( p_j(z) > 0 \) for all \( j \in \mathcal{J}(z) \) and \( q_i(z) > 0 \) for all \( i \in \mathcal{J}(z) \). Also, recall that a set of linear constraints are linearly independent if the coefficient vectors on the left-hand side of these constraints cannot be written as a linear combination of one another.

**Theorem 3.3.** The bandwidth allocation function \( \Lambda(\cdot) \) is differentiable at a point \( z \in (0, \infty)^I \) if the constraints in \( \mathcal{J}(z) \cup \mathcal{J}(z) \) are linearly independent, and if the strict complementarity condition holds. In this case, \( H^d(z) \) is the solution of
\[
(D_{z,d}) \max_{h \in \mathbb{R}^I} \left\{ -2 \sum_{i=1}^I u_i(z)d_i h_i + \sum_{i=1}^I v_i(z) h_i^2 \right\}
\]
subject to \( (Ah)_j = 0, \quad j \in \mathcal{J}(z), \)
\[
h_i = d_i m_i, \quad i \in \mathcal{J}(z),
\]
\[
h_i \leq d_i m_i, \quad \text{if} \quad \Lambda_i(z) = m_i z_i,
\]
Moreover, \( H^d(z), p^d(z), \) and \( q^d(z) \) form the unique solution of the system of \( I + |\mathcal{J}(z)| + |\mathcal{J}(z)| \) linear equations,
\[
2(H^d(z))_j v_j(z) = 2u_j(z) + \sum_{j \in \mathcal{J}(z)} p_j^d(z)A_{ji} + q^d_j(z)I(i \in \mathcal{J}(z)), \quad i = 1, \ldots, I,
\]
\[
(AH^d(z))_j = 0, \quad j \in \mathcal{J}(z),
\]
\[
(H^d(z))_j = d_j m_j, \quad i \in \mathcal{J}(z).
\]
The proof of Theorem 3.3 follows immediately from Theorem 3.1, exploiting strict complementarity and linear independence of the constraints. Note also that, since active constraints are required to be independent of one another, we have that \(|f(x)| + |f(y)| \leq I\) and so if the derivative of \(\Lambda(\cdot)\) exists, then it may be found by solving a system of at most 2I equations. Thus, from a computational point of view, finding the derivative of \(\Lambda\) is not much more difficult than finding \(\Lambda\) itself.

### 4. Fluid and Diffusion Limits

In this section, we provide our main fluid and diffusion limits for bandwidth-sharing networks operating under the bandwidth allocation function \(\Lambda(\cdot)\) described in §3. The context in which our limits will be obtained is an asymptotic regime in which the network topology, individual user rate constraints, and information and patience size distributions remain fixed, while the capacities of the resources in the network and the arrival rates of flows to the network grow arbitrarily large. We refer to this regime as the “large-capacity scaling” regime and, mathematically, it is defined as follows.

We consider a sequence of bandwidth-sharing networks indexed by some parameter \(n \geq 1\). As mentioned in the preceding paragraph, the topology of each network remains fixed with respect to \(n\). In particular, we assume that each network has \(I \geq 1\) routes and \(J \geq 1\) resources with an incidence matrix \(A\) as described in §2. We also assume that the individual user rate constraints for flows on the network remain fixed with respect to \(n\). That is, flows on route \(i = 1, \ldots, I\) have an individual rate constraint given by \(m_i\). Finally, we assume that the capacity vector of the \(n\)th network grows linearly with \(n\) and is given by \(nC\), where \(C = (C_1, \ldots, C_J)^T\) is the original capacity vector as described in §2.

We next assume that in the \(n\)th network at time \(t = 0\), there are \(Z_i^n(0)\) flows already present on route \(i = 1, \ldots, I\), and we set \(Z^n(0) = (Z_1^n(0), \ldots, Z_I^n(0))\). Flows of type \(i\) arrive externally to the \(n\)th network after time \(t = 0\) according to a Poisson process with rate \(\eta_i^n\), and we assume that \(\eta_i^n\) grows roughly at a linear rate as \(n\) grows large. In particular, we assume that \(\eta_i^n/n \to \eta_i > 0\) as \(n \to \infty\). Finally, we assume that the information size distribution and that the patience time distribution of flows arriving to the \(n\)th network does not change with \(n\). In particular, flows of type \(i\) have information sizes, which are i.i.d. exponential random variables with rate \(0 < \mu_i < \infty\), and patience times, which are i.i.d. exponential random variables with rate \(0 < \gamma_i < \infty\). For the remainder of the paper, unless otherwise noted, all relevant quantities associated with the \(n\)th network will be denoted by a superscript \(n\).

One interesting and useful consequence of the definition of the large-capacity scaling regime is that the bandwidth allocation function of the \(n\)th network, \(\Lambda^n\), scales in a natural way with \(n\). In particular, note that \(\Lambda^n\) is given by the solution to the original global utility maximization problem \((P)_n\), but where the capacity vector \(C\) has been replaced by \(nC\). It is then straightforward to verify that this implies that

\[
\Lambda^n(z) = n\Lambda(z/n), \quad z \in \mathbb{R}_+^I.
\]

It turns out that the scaling provided by (5) will play a pivotal role in proving our main limit theorems of the present section.

Now, for each \(n \geq 1\) and \(t \geq 0\), let \(Z^n(t) = (Z_1^n(t), \ldots, Z_I^n(t))\) be the user population vector in the \(n\)th network at time \(t \geq 0\). We then define

\[
\bar{Z}^n(t) = Z^n(t)/n
\]

to be the fluid scaled user population vector at time \(t \geq 0\) and we set \(\bar{Z}^n = (\bar{Z}^n(t), t \geq 0)\) to be the fluid scaled user population process. Our first main result of this section, Theorem 4.1, provides a weak limit, under appropriate initial conditions, for the sequence \((\bar{Z}^n, n \geq 1)\) in the large-capacity scaling regime. We refer to this weak limit as the fluid limit of \((\bar{Z}^n, n \geq 1)\). We now show, the fluid limit of \((\bar{Z}^n, n \geq 1)\) may be characterized as the solution to an \(I\)-dimensional ordinary differential equation (ODE). In particular, we have the following theorem.

**Theorem 4.1.** If \(\bar{Z}^n(0) \Rightarrow \bar{Z}(0) \in (0, \infty)^I\) as \(n \to \infty\), then \(\bar{Z}^n \Rightarrow \bar{Z}\) as \(n \to \infty\), where \(\bar{Z}\) is the unique, strong solution to the system of equations given by

\[
\begin{align*}
\dot{\bar{Z}}_i(t) &= \bar{Z}_i(0) + \eta_i t - \mu_i \int_0^t \bar{Z}_i(s) \, ds \\
&\quad - \gamma_i \int_0^t \bar{Z}_i(s) \, ds, \quad t \geq 0
\end{align*}
\]

for \(i = 1, \ldots, I\).

Our proof of Theorem 4.1 may be found in §EC.1 of the electronic companion. We remark, however, for the moment that, by Theorem 3.2, the bandwidth allocation function \(\Lambda\) is locally Lipschitz on the interior of the positive orthant \((0, \infty)^I\). Moreover, as demonstrated in the proof of Theorem 4.1, if \(\bar{Z}(0) \in (0, \infty)^I\), then the solution to (6) must lie in \((0, \infty)^I\) for all \(t \geq 0\). Hence, a standard successive approximations argument may be used to show that there does indeed exist a unique strong solution to the system of equations given by (6). Also, a more general, measure-valued version of Theorem 4.1 may be found in Remerova et al. (2014), which relaxes the Markovian assumptions of the present paper.

We next note that the fluid limit \(\bar{Z}\) of Theorem 4.1 provides a first-order approximation to the dynamics of the user population process. Indeed, conditional on \(\bar{Z}(0)\), the process \(\bar{Z}\) is entirely deterministic. However, in many situations, it is desirable to obtain a second-order stochastic approximation to the user population process. This is the motivation behind our next result, Theorem 4.2, which provides a diffusion limit approximation to the user population process. To state our result, we first need to set up
the following notation. As a matter of convenience, for the remainder of this section, we assume that
\[ \tilde{Z}_n(t) = \tilde{Z}(t) \] as \( n \to \infty \),

(7)

where \( \tilde{Z}(t) \) is a constant. By (6), this then implies that the fluid limit \( \tilde{Z} \) in Theorem 4.1 is a deterministic function. Thus, for each \( n \geq 1 \) and \( t \geq 0 \), we may define
\[ \tilde{Z}_n(t) = n^{1/2}(\tilde{Z}(t) - \tilde{Z}(t)) \]
to be the diffusion scaled user population vector at time \( t \geq 0 \) and set \( \tilde{Z}_n = (\tilde{Z}_n(t), t \geq 0) \) to be the diffusion scaled user population process. We now proceed to characterize, under appropriate initial conditions, the weak limit of \( (\tilde{Z}_n, n \geq 1) \) as the solution to an \( I \)-dimensional stochastic differential equation (SDE). We refer to this weak limit as the diffusion limit of \( (\tilde{Z}_n, n \geq 1) \). We have the following theorem.

**Theorem 4.2.** If \( \tilde{Z}_n(t) \to \tilde{Z}(t) \) as \( n \to \infty \), and \( \sqrt{n} \cdot (n^{-1} \eta_i - \eta_i) \to \beta_i \in \mathbb{R} \) as \( n \to \infty \) for \( i = 1, \ldots, I \), then \( \tilde{Z} \to \tilde{Z} \) as \( n \to \infty \), where \( \tilde{Z} \) is the unique strong solution to
\[
\tilde{Z}_i(t) = \tilde{Z}_i(0) + \tilde{\xi}_i(t) + \beta_i e - \mu_i \int_0^t \tilde{H}_i(s) \tilde{Z}(s) ds + \gamma_i \tilde{Z}(t), \quad t \geq 0,
\]
for \( i = 1, \ldots, I \), where \( \tilde{\xi}_i = (\tilde{\xi}_i(t), t \geq 0) \), \( i = 1, \ldots, I \) is a sequence of independent Brownian motions with infinitesimal variances given, respectively, by
\[
\sigma_i^2(t) = \eta_i + \mu_i \Lambda_i(\tilde{Z}(t)) + \gamma_i \tilde{Z}(t), \quad t \geq 0
\]

(9)

The proof of Theorem 4.2 may be found in §EC.2 of the electronic companion. We remark, however, that by Proposition 7.2 of §7, the directional derivative \( H^d(c) \) of \( \Lambda \) is locally Lipschitz continuous in \( d \) for each fixed \( c \in (0, \infty)^I \), with a Lipschitz constant that is uniformly bounded (in terms of \( c \)) over compact subsets of \( (0, \infty)^I \). Hence, since by Theorem 4.1, it follows that \( \tilde{Z} \) is a continuous function, which over compact intervals of time \( [0, T] \), lies in compact subsets of \( (0, \infty)^I \), a standard successive approximations technique may be used to show that there exists a unique strong solution to the system of equations given by (8).

We now conclude this section by conducting both a transient as well as a steady-state analysis of the diffusion limit \( \tilde{Z} \) provided by Theorem 4.2. The results of these analyses may then be used in a straightforward manner to obtain stochastic approximations to the transient and steady-state behavior of the user population process itself.

Let us begin by assuming that the bandwidth allocation policy \( \Lambda(\cdot) \) is differentiable at each point \( \tilde{Z}(t) \) along the path of the fluid limit \( \tilde{Z} \). This then implies that for each \( t \geq 0 \), the directional derivative \( H^d(\tilde{Z}(t)) \) is a linear function of \( d \) and so, for each \( t \geq 0 \), we may write \( H^d(\tilde{Z}(t)) = H(\tilde{Z}(t))d \) for a particular matrix \( H(\tilde{Z}(t)) \in \mathbb{R}^{I \times I} \). It then follows that we may apply standard results from the theory of stochastic calculus to obtain an explicit solution to (8), and hence the transient dynamics of \( \tilde{Z} \). One may consult, for instance, §5.6 of Karatzas and Shreve (1991) for the proof of this result. To state this result, however, we first need to set up the following notation. For each \( t \geq 0 \), let \( \tilde{\xi}(t) = (\tilde{\xi}_1(t), \ldots, \tilde{\xi}_I(t))' \) and \( \sigma(t) = (\sigma_1(t), \ldots, \sigma_I(t))' \). Next, set \( \beta = (\beta_1, \ldots, \beta_I)' \) and let \( I_i \) be the \( I \times I \) diagonal matrix such that \( (I_i)_{ii} = \mu_i \) for each \( i = 1, \ldots, I \), and, similarly, let \( I_i \) be the \( I \times I \) diagonal matrix such that \( (I_i)_{ii} = \gamma_i \) for each \( i = 1, \ldots, I \). We then have the following proposition.

**Proposition 4.3.** Suppose that the conditions of Theorem 3.3 hold at \( \tilde{Z}(t) \) for each \( t \geq 0 \) and let \( \Phi(t) \) be the solution to the matrix-valued ODE
\[
\frac{d\Phi(t)}{dt} = -(I_i \hat{H}(\tilde{Z}(t)) + I_i)\Phi(t), \quad t \geq 0,
\]

with initial condition \( \Phi(0) = I \). Then, the solution to (8) is given by
\[
\tilde{Z}(t) = \Phi(t) \left( \tilde{Z}(0) + \int_0^t (\Phi(s))^{-1} (\hat{\xi}(s) + \beta s) ds \right)
\]

(10)

for \( t \geq 0 \). Moreover, if \( E[\|\tilde{Z}(t)\|^2] < \infty \), then
\[
E[\tilde{Z}(t)] = \Phi(t) \left( E[\tilde{Z}(0)] + \int_0^t (\Phi(s))^{-1} \beta ds \right),
\]

and for \( 0 \leq s \leq t \),
\[
E[(Z(s) - E[\tilde{Z}(s)])(Z(t) - E[\tilde{Z}(t)])']
\]

\[
= \Phi(s) \left( E[(Z(0) - E[\tilde{Z}(0)])(Z(0) - E[\tilde{Z}(0)])']
\]
\[
+ \int_0^t ((\Phi(u))^{-1} \sigma(u))(\Phi(u))^{-1} \sigma(u)' du \right) \Phi(t).
\]

Moreover, if \( \tilde{Z}(0) \) is Gaussian distributed, then \( \tilde{Z}(t) \) is Gaussian distributed as well for each \( t \geq 0 \).

For our last result of this section, we provide a characterization of the steady-state behavior of the diffusion limit \( \tilde{Z} \). Suppose first that \( \tilde{Z}(0) \) is such that \( \tilde{Z}(t) = \tilde{Z}(0) \) for all \( t \geq 0 \). We refer to such points \( \tilde{Z}(0) \) as invariant points for the fluid limit Equation (6) and conduct a thorough analysis of such points in the section that follows. However, for the time being, we simply note that if \( \tilde{Z}(0) \) is an invariant point for (6) and if \( \Lambda(\cdot) \) is differentiable at \( \tilde{Z}(0) \), then it follows immediately that we may write \( H(\tilde{Z}(t)) = H(\tilde{Z}(0)) \) for all \( t \geq 0 \). Moreover, for each \( i = 1, \ldots, I \), we have that \( \sigma_i^2(t) = \sigma_i = \eta_i + \mu_i \Lambda_i(\tilde{Z}(0)) + \gamma_i \tilde{Z}(0), \quad t \geq 0 \).

We therefore conclude from (8) that if \( \tilde{Z}(0) \) is an invariant point for (6) and if \( \Lambda(\cdot) \) is differentiable at \( \tilde{Z}(0) \), then the diffusion limit \( \tilde{Z} \) of Theorem 4.2 reduces to a time-homogeneous, \( I \)-dimensional Ornstein-Uhlenbeck process. In this case, we may provide an explicit characterization of the steady-state distribution of \( \tilde{Z} \). In particular, we have the following result, whose proof may be found, for instance, in the proof Theorem 5.6.7 of Karatzas and Shreve (1991).
Proposition 4.4. Assume that \( \bar{Z}(0) \) is an invariant point for (6) and that the conditions of Theorem 3.3 hold at \( \bar{Z}(0) \). Then, let \( A = I_n H(\bar{Z}(0)) + I_n \). If each eigenvalue of \( A \) has a positive real part, then \( \bar{Z}(t) \Rightarrow \bar{Z}(\infty) \) as \( t \to \infty \), where \( \bar{Z}(\infty) \) is a normal random vector with mean

\[
E[\bar{Z}(\infty)] = \beta \int_0^\infty e^{-\lambda t} dt,
\]

and variance-covariance matrix given by

\[
E[(\bar{Z}(\infty) - E[\bar{Z}(\infty)])(\bar{Z}(\infty) - E[\bar{Z}(\infty)]')] = \int_0^\infty e^{-\lambda t} \sigma \sigma' e^{-\lambda t} dt.
\]

5. Invariant Points

Recall from §4 that a point \( z \in (0, \infty)^I \) is said to be an invariant point for the fluid limit Equation (6) if conditional on \( \bar{Z}(0) = z \), one has that \( \bar{Z}(t) = z \) for all \( t \geq 0 \). Note, however, that this condition implies that \( \bar{Z}(0) \) is an invariant point for the fluid limit Equation (6) if and only if one has that \( d\bar{Z}(t)/dt = 0 \) for all \( t \geq 0 \) for \( i = 1, \ldots, I \). Moreover, upon closer inspection of the fluid limit Equation (6), it is evident that

\[
d\bar{Z}(t)/dt = \eta_i - \mu_i + (\gamma_i - \lambda) \bar{Z}(t) + \bar{Z}(t).
\]

Hence, an alternative characterization of the set of invariant points for the fluid limit Equation (6) is those \( z \in (0, \infty)^I \) such that

\[
\eta_i = \mu_i + \lambda z_i \quad \text{for } i = 1, \ldots, I.
\]

In this section, we study the fixed-point Equation (14) and provide a characterization of the set of invariant points for the fluid limit Equation (6) both in the case when all flows are patient, that is, \( \gamma_i = 0 \) for \( i = 1, \ldots, I \) as well as the case when all flows are impatient, that is, \( \gamma_i > 0 \) for \( i = 1, \ldots, I \). Our main objective throughout the section will be to establish existence and uniqueness results for such invariant points. We begin with the case that all flows are patient.

5.1. Patience

Suppose that \( \gamma_i = 0 \) for each \( i = 1, \ldots, I \), so that each flow arriving to the network is willing to wait an unlimited amount of time before completing service. In this case, the fixed-point Equation (14) for the set of invariant points reduces to the simpler form

\[
\Lambda(z) = \rho,
\]

where \( \rho \) is an \( I \)-dimensional column vector such that \( \rho_i = \eta_i/\mu_i \) for each \( i = 1, \ldots, I \). The quantity \( \rho_i \) represents the incoming rate at which work arrives to route \( i \). In addition, recall that, by the capacity constraints of the network, we must have that \( \Lambda \Lambda \leq C \). Hence, it is evident from (15) that in the case where all flows arriving to the network are patient, there cannot exist an invariant point for the fluid limit Equation (6) unless \( \Lambda \Lambda \leq C \). We refer to this condition as the system not being overloaded. In other words, the total rate of incoming work at each resource is no greater than the capacity of the resource. The following is now our main result of this section regarding the case where flows arriving to the network are patient.

Proposition 5.1. Suppose that \( \gamma_i = 0 \) for each \( i = 1, \ldots, I \), and that \( \Lambda \Lambda \leq C \). Then \( z \in (0, \infty)^I \) is an invariant point for the fluid limit Equation (6) if and only if there exists nonnegative constants \( p_j, \; j = 1, \ldots, J \), and \( q_i, \; i = 1, \ldots, I \), such that

\[
z_i = \frac{p_i}{(U'_i)^{-1}(q_i + \sum_{j=1}^J A_{ij})}, \quad i = 1, \ldots, I,
\]

and where \( p_j = 0 \) if \( (A\rho)_j < C_j \) for \( j = 1, \ldots, J \), and \( q_i = (U'_i(m_i) - \sum_{j=1}^J p_j A_{ji})^+ \) for \( i = 1, \ldots, I \).

Proof of Proposition 5.1. We begin with the case of necessity. Suppose that \( z \in (0, \infty)^I \) is an invariant point for the fluid limit Equation (6). It then follows by (15) and the KKT conditions for \( (P_\lambda') \) that there must exist a set of nonnegative constants \( p_j, \; j = 1, \ldots, J \), and \( q_i, \; i = 1, \ldots, I \) such that

\[
U'_i(p_i/z_i) = q_i + \sum_{j=1}^J A_{ij}p_j \quad \text{and} \quad q_i(p_i - m_i z_i) = 0, \quad i = 1, \ldots, I,
\]

and

\[
p_j(A\rho - C)_j = 0, \quad j = 1, \ldots, J,
\]

where in (17) and (18) above, we have dropped the dependency of \( p \) and \( q \) on \( z \). Now, note that the first equality in (17) implies (16) and that (18) implies \( p_j = 0 \) when \( (A\rho)_j < C_j, \; j = 1, \ldots, J \). To verify that \( q_i = (U'_i(m_i) - \sum_{j=1}^J p_j A_{ji})^+ \) for \( i = 1, \ldots, I \), we reason as follows. First, note that, by the first equality in (17), we have that \( q_i = U'_i(p_i/z_i) - \sum_{j=1}^J p_j A_{ji} \) for \( i = 1, \ldots, I \). Hence, if \( p_i/z_i < m_i \), then it is immediate that the desired relationship holds. On the other hand, if \( p_i/z_i \leq m_i \), then by the second equality in (17), we have that \( q_i = 0 \). Moreover, by the strict concavity of \( U'_i \), we have that \( U'_i(m_i) < U'_i(p_i/z_i) \) and so again, the desired relationship holds.

We next prove sufficiency. Suppose that \( z \in (0, \infty)^I \) satisfies (16). We now show that \( \Lambda(z) = \rho \), which, by (15), implies that \( z \) is an invariant point for the fluid limit Equation (6). However, it may be simply verified that \( \Lambda(z) = \rho \) by noting that the conditions of the proposition imply that the KKT conditions for \( \Lambda(z) \) hold when choosing \( (\Lambda(z), p(z), q(z)) = (\rho, \rho, q) \).
We remark that in the underloaded case of \((Ap)_j < C_j\) for each \(j = 1, \ldots, J\), one has that \(p_j = 0\) for each \(j = 1, \ldots, J\), and so by Proposition 5.1, it follows that \(q_i = U'_i(m_i)\) for each \(i = 1, \ldots, I\). By the strict concavity of \(U_i\), this then implies that \(z_i = \rho_i/m_i\) for each \(i = 1, \ldots, I\), and so the invariant point \(z = (\rho_1/m_1, \ldots, \rho_I/m_I)\) is unique. On the other hand, in the critical case for which \((Ap)_j = C_j\) for each \(j = 1, \ldots, J\), then any choice of \(p_j \geq 0\) is feasible, and so there may exist multiple invariant points.

5.2. Impatience

We next cover the case in which \(\gamma_i > 0\) for each \(i = 1, \ldots, I\), so that each flow on the network is impatient and may abandon from the network if not served within a reasonable amount of time. It turns out, the proof of the characterization of invariant points in the presence of impatience is more challenging than in the previous subsection where all of the flows on the network were assumed to be patient. Our main result is the following.

**Proposition 5.2.** Suppose that \(\gamma_i > 0\) for each \(i = 1, \ldots, I\). Then, there exists a unique invariant point for the fluid limit Equation (6).

A more general version of Proposition 5.2 (relaxing the Markovian assumptions) is given by Theorem 2 of Remerova et al. (2014). Specializing that result to our particular case, we remark that the unique invariant point for the fluid limit Equation (6) can be characterized in terms of the solution to the following optimization problem. First, for each \(i = 1, \ldots, I\), define the function \(G_i\) by setting \(G_i(x) = U_i(x\gamma_i/(\eta_i - \mu_i) x)\) for \(0 < x < \rho_i\). It then follows that \(G_i\) is a strictly decreasing function, and hence \(G_i\) is strictly concave. Next, define the optimization problem \((Q)\) as follows:

\[
(Q) \quad \max_{\Lambda} \sum_{i=1}^{I} G_i(\Lambda_i)
\]

subject to \(A\Lambda \leq C\)

\[
A_i = \begin{bmatrix} \frac{m_i \eta_i}{\gamma_i + m_i \mu_i} \end{bmatrix}, \quad i = 1, \ldots, I.
\]

The unique invariant point for the fluid limit Equation (6) can now be found by first solving \((Q)\) and obtaining its solution \(\Lambda\), and then substituting \(\Lambda\) into the fixed-point Equation (14) to determine \(z\).

We now conclude this section by making a connection between the results above and those of Theorem 3.3 and Proposition 4.4. First, note that \(\Lambda(\cdot)\) is differentiable at the invariant point \(z\) if the active constraints of \((Q)\) are linearly independent and if the strict complementarity condition is satisfied for \((Q)\). This follows from Theorem 3.3 since the Lagrange multipliers for \((P)\) can be expressed in terms of those for \((Q)\), and since the set of active constraints are identical in both problems, see the proof of Theorem 2 of Remerova et al. (2014) for more details. Thus, thanks to the above characterization of the unique invariant point \(z\) in terms of \((Q)\), the condition in Proposition 4.4 can be checked algorithmically.

### 6. Examples

We now provide two examples illustrating how the methodology of the present paper may be applied to analyzing specific bandwidth-sharing networks.

#### 6.1. Connection with Queues in the Halfin-Whitt Regime

We begin with a basic example showing how our work is related to previous results from the call center queueing literature. Let us assume that we have a bandwidth-sharing network consisting of a single route and a single resource. That is, \(I = J = 1\). In addition, set \(\eta = C = m = \mu = 1\) and \(\gamma = 0\). Suppose further that in the \(n\)th network of the large-capacity scaling regime, we have that the arrival rate of flows to the single route is given by \(\eta^n = n - \beta \sqrt{n} + o(\sqrt{n})\) and that the capacity of the single resource is equal to \(n\). Moreover, assume that arriving flows have mean information sizes equal to 1 and that all flows are infinitely patient, that is, \(\gamma = 0\). It is then not difficult to see that, for any strictly increasing utility function, \(U\), the bandwidth allocation function \(U^*(\cdot)\) in the \(n\)th system is given by \(U^*(z) = \min\{z, n\}\). Consequently, the user population process in the \(n\)th system, \(Z^n\), will behave as a birth-death process with constant birth rates \(\eta^n\) and death rates equal to \(\min\{n, Z^n(t)\}\). In other words, the user population process will behave in an identical manner as the number of customers in an \(M/M/n\) queue with an arrival rate of \(\eta^n\) and a service rate of 1. Also, note that applying the results of Theorem 4.1 with \(\eta = C = m = \mu = 1\) and \(\gamma = 0\), it is straightforward to see that \(p = \eta/\mu = 1\) and that the resulting set of invariant points for the corresponding fluid limit Equation (6) is \([1, \infty)\).

Now, note that it is evident that in the setup described above, the solution to \((P)\) is given by \(\Lambda(z) = \min\{z, 1\}\). Hence, it is straightforward to deduce that the directional derivative of \(\Lambda\) at the point \(z = 1\) is given by \(H^\Lambda(1) = d1\) \((d < 0)\). Nevertheless, it is instructive to obtain this result using the general theory provided in Theorem 3.1. We proceed as follows. For the purposes of illustration, we assume that the utility function of the flows on the single route in our network is given by \(U(x) = \log x\). It is then evident that the KKT conditions at \(z = 1\) specialize to

\[
1/\Lambda(1) = p(1) + q(1), \quad p(1)(\Lambda(1) - 1) = 0, \quad \text{and} \quad q(1)(\Lambda(1) - 1) = 0.
\]

We then see by inspection that the solution to (19) is given by \(\Lambda(1) = 1\) along with all nonnegative pairs \((p, q)\) such that \(p + q = 1\). Next, Problem (3), which is to minimize \(d\) over all pairs \((p, q)\) solving (19), results in an optimal solution \((p(1), q(1)) = (0, 1)\) if \(d < 0\) and \((p(1), q(1)) = (1, 0)\) if \(d > 0\). Consequently, \(H^\Lambda(1)\) is the maximizing value of the quadratic function \(2d^2 - h^2\) subject to the set constraints \(h = 0\), \(h \leq d\) if \(d > 0\) and the set of constraints \(h \leq 0\), \(h = d\) if \(d < 0\). It is then straightforward to see that
this results in the directional derivative \( H^d(1) = d1(d < 0) \) given above.

Now, suppose that we set \( \tilde{Z}(0) \) equal to the invariant point \( z = 1 \). That is, the limiting fluid scaled initial number of users on the network is equal to one. Then, applying Theorem 4.2, we conclude that the diffusion limit for the user population process is given by the solution to the SDE

\[
d\tilde{Z}(t) = -\left(\beta + \min\{\tilde{Z}(t), 0\}\right)dt + \sqrt{2}d\tilde{W}(t), \quad t \geq 0, \quad (20)
\]

where \( \tilde{W}(\tilde{W}(t), t \geq 0) \) is a standard Brownian motion. Assuming that \( \beta > 0 \), the steady-state distribution of this diffusion is not Gaussian but is still computable. We refer to Halfin and Whitt (1981) for further details.

6.2. A Single Link with Multiple Customer Classes

We next consider a network with a single resource but with multiple routes passing through it. That is, we assume that \( J = 1 \) but that \( I \geq 1 \), see Figure 2. We assume that the single resource operates at unit capacity, implying that \( C = 1 \) and that customers on each route in the network are impatient, so that \( \gamma_i > 0 \) for each \( i = 1, \ldots, I \). Moreover, we assume that the utility functions for the flows on route \( i = 1, \ldots, I \) are given by the weighted proportional fairness utility functions, so that \( U_i(x) = \kappa_i \log x \) for some \( \kappa_i > 0 \).

Our main focus will be on characterizing the unique invariant point of the fluid limit Equation (6). First, let us set \( \bar{\rho}_i = m_i \eta_i/(\gamma_i + m_i \mu_i) \) for each \( i = 1, \ldots, I \). It is then not too difficult to see that if \( \sum_i \bar{\rho}_i \leq 1 \), then the unique invariant point for the fluid limit Equation (6) is given by \( z_i = \eta_i/(\gamma_i + m_i \mu_i) \) with \( \Lambda_i = \bar{\rho}_i \) for each \( i = 1, \ldots, I \). This implies that flows on each route \( i \) are served according to their maximum possible rate \( m_i \). On the other hand, if \( \sum_i \bar{\rho}_i > 1 \), then the optimization problem (Q) characterizing the unique invariant point may be explicitly solved according to the following procedure:

1. Order the indices \( i = 1, \ldots, I \) such that \( \kappa_i/m_i \leq \kappa_2/m_2 \leq \cdots \leq \kappa_I/m_I \).

Figure 2. A single resource \((J = 1)\) with \( I = 3 \) routes running through it.

(2) Solve for \( p^* \) such that

\[
\sum_{i=1}^I \frac{\eta_i}{\gamma_i} \max\{\kappa_i/\mu_i, p^*\} + \mu_i = 1.
\]

(3) Set \( \bar{\nu} = \max\{i: \kappa_i/\mu_i < p^*\} \).

(4) Set \( \Lambda_i = \eta_i/(\gamma_i + \kappa_i/\mu_i) \) and note that \( \Lambda_i = \bar{\rho}_i \) if \( i > \bar{\nu} \).

(5) Set \( \gamma_i = (\eta_i - \Lambda_i \mu_i)/\gamma_i \).

For the validity of the above procedure as well as a rigorous proof for the case of \( \sum_i \bar{\rho}_i \leq 1 \), we refer the reader to the proof of Theorem 4 of Remerova et al. (2014), and, in particular, Equations (25)–(31) of that paper.

We now highlight the implications of the invariant point described by the above procedure. Note that flows on routes \( i = 1, \ldots, i^* \) should be expected to almost always be served at strictly less than their maximum possible rate \( m_i \), while flows on routes \( i = i^* + 1, \ldots, I \) should be expected to almost always be served at their maximum possible rate \( m_i \). The only possible exception to this rule occurs when \( p^* = \kappa_{i^*+1}/m_{i^*+1} \), in which case flows on route \( i^* + 1 \) will, depending on the number of flows on each route in the system, switch back and forth between being served at their maximum possible rate \( m_i \) and being served at some lesser rate. Hence the ratio \( \kappa_i/m_i \) plays a critical role in the steady-state behavior of the system. This observation may then be used to construct certain pricing schemes. For example, users may be asked to pay a price for a certain maximum amount of bandwidth \( m_i \) and a service provider can then optimize over \( \kappa_i \) to maximize its profit. It is also interesting to note the similarity between the characterizations mentioned above and the ED, QD, and QED regimes, respectively, which were mentioned in the §1. Finally, note that all flows in the network operate in the overloaded regime when \( \kappa_i \) is set in constant proportion to \( m_i \).

7. Proofs of Properties of the Bandwidth-Sharing Function

In this section, we provide the proofs of Theorems 3.1 and 3.2 from §3. We also prove a proposition regarding the directional derivative of the bandwidth allocation function \( \Lambda \). We begin, however, with the following preliminary result regarding the continuity of \( \Lambda(\cdot) \).

**Lemma 7.1.** \( \Lambda(\cdot) \) is continuous on \((0, \infty)^I\).

**Proof of Lemma 7.1.** Our argument is a simplified version of the proof of Lemma 1 in Remerova et al. (2014). In particular, it suffices to show that for each vector \( z \in (0, \infty)^I \) and sequence \( \{z^k, k \geq 1\} \) such that \( z^k \to z \) as \( k \to \infty \), we have that \( \Lambda(z^k) \to \Lambda(z) \) as \( k \to \infty \).

We proceed by contradiction. Let \( z \in (0, \infty)^I \) and suppose that \( z^k \to z \) but that \( \Lambda(z^k) \not\to \Lambda(z) \). Moreover, note that, since the set \( \{\Lambda(z^k), k \geq 1\} \) is a subset of the compact set \( \{\Lambda \geq 0, AA \leq C\} \), we may assume without loss of generality that \( \Lambda(z^k) \to \Lambda' \neq \Lambda(z) \). In addition, note that, since
\( \Lambda^* \) is clearly a feasible point of \((P_2)\), and since \( \Lambda(z) \) is the unique optimal solution to \((P_2)\), we have that
\[
I := \sum_{i=1}^{l} z_i U_i(\Lambda(z)_i/z_i) > \sum_{i=1}^{l} z_i U_i(\Lambda^*_i/z_i) := r. \tag{21}
\]
For each \( k \geq 1 \), now define \( \Lambda^k = (\Lambda^k_1, \ldots, \Lambda^k_l) \) by setting \( \Lambda^k_i := \min \{ \Lambda(z)_i, m_i z_i \} \) for each \( i = 1, \ldots, l \), and note that \( \Lambda^k \) is a feasible point of \((P_2)\). Moreover, since \( z^k \to z \), \( \Lambda^k \to \Lambda(z) \), and \( \Lambda(z^k) \to \Lambda^* \), it follows that
\[
\sum_{i=1}^{l} z_i U_i(\Lambda^k_i/z_i) \to I \quad \text{and} \quad \sum_{i=1}^{l} z_i U_i(\Lambda(z^k)_i/z_i) \to r.
\]
Thus, by (21), for \( k \) sufficiently large, we have that
\[
\sum_{i=1}^{l} z_i U_i(\Lambda^k_i/z_i) \geq \sum_{i=1}^{l} z_i U_i(\Lambda(z^k)_i/z_i),
\]
which contradicts \( \Lambda(z^k) \) being optimal for \((P_2)\). This completes the proof. \( \square \)

We now proceed with the proof of Theorem 3.1.

**Proof of Theorem 3.1.** We follow closely §5.2.3 of Bonnans and Shapiro (2000). In particular, directional differentiability of \( \Lambda(\cdot) \) at each point \( z \in (0, \infty)^l \) will follow once we have established that each condition of Theorem 5.53 in Bonnans and Shapiro (2000) as well as each condition of Remark 5.55 of Bonnans and Shapiro (2000) hold for each \( z \in (0, \infty)^l \). We begin by verifying that each of the five conditions of Theorem 5.53 of Bonnans and Shapiro (2000) are satisfied for \((P_2)\) for each \( z \in (0, \infty)^l \).

(i) Clearly, \( \Lambda(\cdot) \) has a unique optimal solution \( \Lambda(z) \) since we are optimizing a strictly concave function over a convex closed set, hence assumption (i) is satisfied.

(ii) Since the set of feasible points \( \{ \Lambda \geq 0, AA \leq C, \Lambda \leq z, m \} \) is nonempty, the Fromovitz-Mangasarian constraint qualification holds (see pages 71–72 of Bonnans and Shapiro 2000), and so, by Proposition 5.50(v), Gollan’s condition holds. This verifies that assumption (ii) is satisfied.

(iii) Note that, since the criterion function of our problem is strictly concave, and the set of feasible points is convex, the set of Lagrange multipliers \( \gamma(z) \) is nonempty, and hence assumption (iii) is satisfied.

(iv) Let \( Q = Q(\Lambda, z, p, q) \) be an \( l \times l \) matrix such that \( Q_j(\Lambda, z, p, q) = (\partial/\partial \Lambda_j) \Lambda_j(\partial/\partial \Lambda_j) L(\Lambda, z, p, q) \), where \( L \) denotes the Lagrangian associated with \((P_2)\), see (22) below. A simple computation then shows that \( Q_j(\Lambda, z, p, q) = I(\ast j) U''_i(\Lambda_i/z_i) \), which is independent of \( p \) and \( q \). Thus we have that for any \( I \) dimensional vector \( h \),
\[
(Q(\Lambda, z, p, q)h) h' = \sum_{i=1}^{l} h_i^2 U''_i(\Lambda_i/z_i),
\]
which is strictly positive as long as \( h \neq 0 \). This implies that the strong second-order conditions (5.120) in Bonnans and Shapiro (2000) are satisfied, and so assumption (iv) is verified.

(v) For all \( z^0 \) in a neighborhood of \( z \), the set of feasible points \( \{ \Lambda \geq 0, AA \leq C, \Lambda \leq z^0 m \} \) is nonempty and uniformly bounded. Note that this is true even if \( m_i = \infty \) for some \( i \), since \( \Lambda_i \leq \max_j C_j \). Hence, assumption (v) is satisfied.

Having now verified that each of the five conditions of Theorem 5.53 of Bonnans and Shapiro (2000) are satisfied at each \( z \in (0, \infty)^l \), we next proceed to verify that each of the conditions in Remark 5.55 of Bonnans and Shapiro (2000) hold as well. This will then imply the directional differentiability of \( \Lambda(\cdot) \) at each point \( z \in (0, \infty)^l \). However, note that, by properties (ii)–(v) above, as well as Lemma 7.1, both suppositions of the remark hold for each \( z \in (0, \infty)^l \) and direction \( d \in \mathbb{R}^l \). Hence \( \Lambda \) is directionally differentiable at each \( z \in (0, \infty)^l \).

We next proceed to derive an expression for the directional derivative of \( \Lambda(\cdot) \) at each point \( z \in (0, \infty)^l \). We begin by introducing some additional notation. For each \( z \in (0, \infty)^l \) and \( d \in \mathbb{R}^l \), let
\[
(P_{L_d}) \quad \max_{h \in \mathbb{R}^l} \sum_{i=1}^{l} \left\{ U_i'\left(\frac{\Lambda(z)_i}{z_i}\right) h_i + \left( U_i\left(\frac{\Lambda(z)_i}{z_i}\right) - \frac{\Lambda_i}{z_i} U_i'\left(\frac{\Lambda(z)_i}{z_i}\right) \right) d_i \right\}
\]
subject to \( (Ah)_j \leq 0, \quad j \in f(z), \)
\[ h_i \leq m_i d_i, \quad i \in f(z), \]
be the linearization (in the direction \( d \)) of \((P_2)\) at \( (\Lambda(z), z) \) (see, for instance, problem \((P_{L_d})\) on page 446 of Bonnans and Shapiro 2000). Next, recall that the Lagrangian associated with \((P_2)\) may be written as
\[
L(\Lambda, z, p, q) = \sum_{i=1}^{l} z_i U_i\left(\frac{\Lambda_i}{z_i}\right) + \sum_{j=1}^{J} p_j ((AA)_j - C_j)
\]
\[ + \sum_{i=1}^{l} q_i (\Lambda_i - m_i z_i). \tag{22}
\]
where, in the above, we note that \( \Lambda \) is not a function of \( z \). Also, let \( \gamma(z) \) denote a set of Lagrange multipliers of \((P_2)\). The dual of \((P_{L_d})\) may then be written as
\[
(D_{L_d}) \quad \max_{(p, q) \in \gamma(z)} D_LL(\Lambda(z), z, p, q) d.
\]
Here, \( D_LL(\Lambda, z, p, q) \) is the gradient of \( L(\Lambda, z, p, q) \), where the derivative of the Lagrangian is taken with respect to each coordinate of \( z \) when \( \Lambda \) is regarded as not depending on \( z \), and the multiplication with \( d \) is taken to be the dot product. The above dual problem may be written explicitly as
\[
\max_{(p, q) \in \gamma(z)} \sum_{i=1}^{l} d_i \left( U_i\left(\frac{\Lambda_i}{z_i}\right) - \frac{\Lambda_i}{z_i} U_i'\left(\frac{\Lambda_i}{z_i}\right) \right) - \sum_{i=1}^{l} d_i m_i q_i, \tag{23}
\]
which may then be further simplified to (3). Now, let \((p^*(z), q^*(z))\) be the set of Lagrange multipliers that
solve (23). It then follows that the set of optimal solutions to \((PL_d)\) may be written as
\[
\mathcal{P}(PL_d) = \begin{cases}
(\lambda h_j) = 0 & \text{if } p^d_j(z) > 0, \\
h_j: (\lambda h_j) \leq 0 & \text{if } (\Lambda(z)) = \lambda_j, \\
h_j = d_i m_i & \text{if } q^d_j(z) > 0, \\
h_j \leq d_i m_i & \text{if } \Lambda_j(z) = m_i z_j.
\end{cases}
\]

(24)

Having set up the above notation, we now recall that the directional derivative of \(\Lambda\) at \(z\) is a solution to the optimization problem (5.125) of Bonnans and Shapiro (2000), which, in our case, is unique since (5.126) of Bonnans and Shapiro (2000) is always satisfied. We therefore now proceed to specialize (5.125) in Bonnans and Shapiro (2000) to our particular situation. Let \(O(\lambda, z)\) be the value of the objective function of \((P_i)\) and note that the Hessian of \(O\) regarding \(\lambda\) and \(z\) is equivalent to that of \(L(\lambda, z, p, q)\) above. It therefore follows that problem (5.125) in Bonnans and Shapiro (2000) reduces to problem \((D_{\tilde{z}, d})\) in Theorem 3.1 upon noting the equality of the set (24) with the set determined by the constraints of \((D_{\tilde{z}, d})\) as well as the fact that
\[\frac{\partial O}{\partial z_i \partial \lambda_j} = 0\]
for \(i \neq j\), and
\[\frac{\partial O}{\partial z_i \partial z_j} = \frac{\lambda_i^2}{z_i} U_j''(\frac{\lambda_i}{z_i}), \quad \frac{\partial O}{\partial \lambda_j} = \frac{1}{z_i} U_j''(\frac{\lambda_i}{z_i}).\]
\[\frac{\partial O}{\partial \lambda_i} = -\frac{\lambda_i^2}{z_i} U_j''(\frac{\lambda_i}{z_i}).\]

We omit the lengthy computation of the above results. Note finally that, since the directional derivative of \(\Lambda(\cdot)\) is well defined, it does not matter which vector of Lagrange multipliers \((p^d(z), q^d(z))\) that solve (23) are used.

We next provide the proof of Theorem 3.2.

**Proof of Theorem 3.2.** Let \(E \subset (0, \infty)^I\) be compact. Then, by the Hahn-Banach separation theorem (see, for instance, Rudin 1991), there exists an open and bounded set \(\tilde{E} \subset (0, \infty)^I\) whose closure \(\tilde{E} \subset (0, \infty)^I\) and such that \(E \subset \tilde{E}\). In particular, we may set \(E = (E'_1, E'_2) \times \cdots \times (E'_I, E'_I)\) for some set of finite constants \(E'_1, E'_2, \ldots, E'_I\). Hence, to complete the proof, it suffices to show that there exists a constant \(K_E\) such that \(\|\Lambda(x) - \Lambda(y)\| \leq K_E \|x - y\|\) for all \(x, y \in E\). However, by the fundamental theorem of calculus (see, for instance, Edwards and Penney 1994), it suffices to show that there exists a constant \(K_E\geq 0\) such that for each \(z \in E\), we have that \(\|H^d(z)\| \leq K_E\) for each \(d \in \mathbb{R}^I\) with \(\|d\| = 1\).

For each \(i = 1, \ldots, I\), let \(e_i\) be the \(I\)-dimensional vector of all zeros except a 1 in the \(i\)th position. Then, for a direction \(d \in \mathbb{R}^I\) with \(\|d\| = 1\), we may write
\[d = \sum_{i=1}^I d_i e_i,\]
where \(-1 \leq d_i \leq 1\) for each \(i = 1, \ldots, I\), and
\[\sum_{i=1}^I d_i^2 = 1.\]

Now, for each \(x \in \mathbb{R}^I\) (0), let \(\text{sgn}(x) = x/|x|\) and set \(\text{sgn}(0) = 0\). It then follows that for each \(z \in E\) and \(d \in \mathbb{R}^I\) such that \(\|d\| = 1\), we have that
\[\|H^d(z)\| = \lim_{t \to 0^+} \frac{\|\sum_{i=1}^I (\Lambda(z + t d_i e_i) - \Lambda(z + t \sum_{j=1}^{i-1} d_j e_j))\|}{t}\]
\[= \lim_{t \to 0^+} \frac{\|\sum_{i=1}^I (\Lambda(z + t \sum_{j=1}^{i-1} d_j e_j) - \Lambda(z + t \sum_{j=1}^{i-1} d_j e_j))\|}{t}\]
\[\leq \sum_{i=1}^I \sup_{t > 0} \left| \left( \frac{\Lambda(z + t \sum_{j=1}^{i-1} d_j e_j) - \Lambda(z + t \sum_{j=1}^{i-1} d_j e_j)}{t} \right) \right| \leq \sum_{i=1}^I \sup_{t > 0} \left| H^{\text{sgn}(d)}(z) \right|.
\]

where in the second to last inequality above, we have used the fact that \(E\) is open as well as the fact that for any \(z \in E\), using a standard argument, we have that \(\|\Lambda(z + t d_i e_i) - \Lambda(z)\| \leq |t| d_i \sup_{z \in \tilde{E}} \|H^{\text{sgn}(d)}(z)\|\) for all \(i\) sufficiently small such that \(z + t d_i e_i \in E\). Hence, to complete the proof, it suffices to show that for each \(i = 1, \ldots, I\), there exists a \(K_E \geq 0\) such that \(\|H^d(z)\| \leq K_E\) for each \(z \in E\) and that for each \(i = 1, \ldots, I\), there exists a \(K_{-E} \geq 0\) such that \(\|H^{\text{sgn}(d)}(z)\| \leq K_{-E}\) for each \(z \in E\).

It suffices to prove the above set of claims for the case of \(e_i\). The proofs for the cases of \(e_2, \ldots, e_I\) as well as \(e_{-1}, \ldots, e_{-I}\), follow in a similar manner. We first show that the optimal value of the optimization problem \((D_{\tilde{z}, e_i})\) defined in Theorem 3.1 is uniformly bounded from below for all \(z \in E\). Let \(z \in (0, \infty)^I\) and note that the objective function for \((D_{\tilde{z}, e_i})\) may be written as
\[-2u(z) h_i + \sum_{i=1}^I v_i(z) h_i^2,\]

(25)

where \(u_i\) and \(v_i\), \(i = 1, \ldots, I\) are defined in (4). Next, note that, since by assumption \(E\) is bounded, it follows that \(E\) is compact. We then define
\[u_i^+(E) = \sup_{z \in \tilde{E}} |u_i(z)|\]
and
\[v_i^+(E) = \sup_{z \in \tilde{E}} |v_i(z)|, \quad i = 1, \ldots, I,\]
\[u_i^-(E) = \inf_{z \in \tilde{E}} |u_i(z)|\]
and
\[v_i^-(E) = \inf_{z \in \tilde{E}} |v_i(z)|, \quad i = 1, \ldots, I.\]

(27)
Note that each of the above quantities are finite since $\tilde{E}$ is compact and since both $u_i$ and $v_i$, $i = 1, \ldots, I$ are continuous on $\tilde{E}$. It therefore follows that for each $h \in \mathbb{R}^I$, we have the inequality

$$-2u_1(z)h_1 + \sum_{i=1}^{I} v_i(z)h_i^2 \geq -2u_1^*(E)|h_1| - \sum_{i=1}^{I} v_i^*(E)h_i^2. \quad (28)$$

Next, note that the feasible region of $(D_{z,e})$ may change, depending on the value $z$. In particular, upon inspecting the potential set of constraints of $(D_{z,e})$ and considering whether each constraint is required or not required to be a part of the constraint set, it is straightforward to see that there may be at most $2^{2(I+J)}$ different feasible regions for $(D_{z,e})$. We now define the set $\mathcal{E} \subset \mathbb{R}^I$ to be an arbitrary subset of $\mathbb{R}^I$ consisting of exactly one point from each feasible region of $(D_{z,e})$ that is nonempty. Note that $\mathcal{E}$ is nonempty since, by Theorem 3.1, the directional derivative of $\Lambda$ exists at each point $z \in \tilde{E}$. Moreover, note that we may restrict the cardinality of $\mathcal{E}$ to be at most $2^{2(I+J)}$. We now define the finite quantity

$$\kappa(E) = \min_{h \in \mathcal{E}} \left\{ -2u_1^*(E)|h_1| - \sum_{i=1}^{I} v_i^*(E)h_i^2 \right\}.$$

It then follows from (25) and (28) that at each point $z \in E$, the optimal value of $(D_{z,e})$ must be at least $\kappa(E)$, our desired lower bound.

Now, let $H^i(z) = (H^i_1(z), \ldots, H^i_I(z)) \in \mathbb{R}^I$ be the directional derivative of $\Lambda$ in the direction $e_i$ at the point $z \in \tilde{E}$. It then follows from Theorem 3.1 and the lower bound $\kappa(E)$ for the optimal value of $(D_{z,e})$, that we must have that

$$-2u_1(z)H^i_1(z) + \sum_{i=1}^{I} v_i(z)(H^i_i(z))^2 \geq \kappa(E). \quad (29)$$

However, note that, by (26) and (27) and the nonpositivity of both $u_i$ and $v_i$, $i = 1, \ldots, I$, we have that

$$2u_1^*(E)|H^i_1(z)| - \sum_{i=1}^{I} v_i^*(E)(H^i_i(z))^2 \geq -2u_1(z)H^i_1(z) + \sum_{i=1}^{I} v_i(z)(H^i_i(z))^2, \quad (30)$$

from which it follows by (29) that

$$2u_1^*(E)|H^i_1(z)| - \sum_{i=1}^{I} v_i^*(E)(H^i_i(z))^2 \geq \kappa(E). \quad (31)$$

Now, using the positiveness of $u_1^*(E)$ and $v_i^*(E)$, $i = 1, \ldots, I$ as well as the finiteness of $\kappa(E)$, it is straightforward to see that (31) implies that there exists some constant $K_E \geq 0$ such that $|H^i(z)| \leq K_E$ for all $z \in E$. This completes the proof. □

We now complete this section with the following proposition, which will be useful in proving our fluid and diffusion limit results in the electronic companion. We have the following.

**Proposition 7.2.** For each $z \in (0, \infty)^I$, $H^i(z)$ is Lipschitz continuous as a function of $d$ with Lipschitz constant $\zeta_i$. Moreover, $\zeta_i$ is uniformly bounded over each compact subset of $(0, \infty)^J$.

**Proof of Proposition 7.2.** Let $E \subset (0, \infty)^I$ be a set of the form $E = (E_1^0 \times \cdots \times (E_I^0 \times E_I^1)$, where $E_i^0 \times E_i^1 > 0$ for each $i = 1, \ldots, I$. Next, let $z \in (0, \infty)^I$ and recall that, by definition, $H^i(z) = \lim_{t \to 0} r^{-1}(\Lambda(z + td) - \Lambda(z))$. Also, recall by Theorem 3.2 that $\Lambda$ is locally Lipschitz on $(0, \infty)^I$. In particular, suppose that $\Lambda$ has Lipschitz constant $K_E$ on $E$. It then follows that for $z \in E$ and $d_1, d_2 \in \mathbb{R}^I$,

$$\|H^i(z) - H^i(z)\| = \lim_{t \to 0} r^{-1}(\Lambda(z + td) - \Lambda(z)) \leq K_E \lim_{t \to 0} r^{-1} \|td_1 - td_2\| = K_E \|d_1 - d_2\|.$$

Thus, for each $z \in E$, we have that $H^i(z)$ is Lipschitz continuous as a function of $d$ with Lipschitz constant $K_E$, which is independent of $z \in E$. Hence, using the Hahn-Banach separation theorem (see, for instance, Rudin 1991) in the same manner as in the proof of Theorem 3.2, this completes the proof. □

**Supplemental Material**

Supplemental material to this paper is available at http://dx.doi.org/10.1287/opre.2014.1321.

**Acknowledgments**

The second author of this paper is also affiliated with Vrije Universiteit (VU) University Amsterdam, Euromand and Georgia Tech, and is sponsored by a VIDI grant from NWO, the Netherlands Organisation for Scientific Research and an IBM faculty award. Part of this research has been carried out at the Newton Institute in Cambridge, United Kingdom. Both authors are grateful to Alex Shapiro for pointing them toward key details in his book Bonnans and Shapiro (2000). The authors would also like to thank the referee for their constructive and helpful comments on the paper.

**References**


Josh Reed is an associate professor of information, operations and management science at the Leonard N. Stern School of Business, New York University, New York, New York. His research interests are in applied probability and stochastic networks.

Bert Zwart leads the Stochastics Group at the Center of Mathematics & Computer Science (CWI) in Amsterdam. He also holds a professor position at VU Amsterdam, and is affiliated with Georgia Tech. His research is in applied probability and stochastic networks, and more recently in revenue management.