

Limit theorems for bandwidth sharing networks with rate constraints

Josh Reed	Bert Zwart
Stern School of Business	CWI
New York University	The Netherlands
New York, NY	Amsterdam

December 22, 2010

Abstract

Bandwidth-sharing networks as considered by Massoulié & Roberts [33] 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 as their analysis requires techniques from both 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 each user is allowed to submit is bounded above by some rate, as is standard in practice. 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 function of the size of the network.

1 Introduction

Bandwidth-sharing networks as considered by Massoulié & Roberts [29, 33] 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 [26]. From an Operations Research 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, stochastic models.

Contemporary research has devoted a significant amount effort to analyzing bandwidth sharing networks and a considerable amount of this effort has been devoted to deriving stability conditions

for bandwidth sharing networks. This question is still not settled in general and is not the subject matter of the present paper although a variety of results may be found in De Veciana *et al.* [34, 35], Bonald & Massoulié [5], Mo & Walrand [30], Massoulié [28], Bramson [13], Gromoll & Williams [21], and Chiang *et al.* [14]. 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.* [8].

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 & Williams [25] and Gromoll and Williams [20]. Properties of overloaded bandwidth sharing networks have subsequently been derived by Borst *et al.* [11, 15]. A diffusion approximation for bandwidth sharing networks was derived in Kang *et al.* [23]. Ye and Yao [36, 37] 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.

This 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. For instance, it is common in applications (see Bonald & Proutiere [7]) 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 work in which system capacities and user download speeds are assumed to be comparable. A consequence of our limit on the maximum individual download speed is that a significant number of users are required in order 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 [16] for a survey 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 & Whitt [22] for many-server queues. In fact, the model we consider is a highly non-trivial

example of a *Markovian Service Network*, as considered by Mandelbaum *et al.* [27]. Unfortunately, we could not directly fit our assumptions into theirs, so we verify the necessary details from scratch in an appendix.

In the call center queuing 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)*. As we shall see in the present paper, in a multi-class multi-node 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, as 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 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 & Mandjes [1]. 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, our framework makes the fluid and diffusion approximations in [1] rigorous, and does not require explicit knowledge of the bandwidth allocation function. In fact, all that is necessary is a directional differentiability property that is established in complete generality in Section 2 of this paper, using results from the sensitivity analysis of nonlinear programs as developed in Bonnans & Shapiro [10]. 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 service times and patience times of users. Extensions to general distributions is currently under investigation and is challenging but promising since one artefact of our scaling is that our system never empties, thus eliminating many of the problems faced by studies in conventional bandwidth sharing networks. This is substantiated by the fact that 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 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.

The remainder of this paper is organized as follows. A model description is provided in Section 2. Section 3 contains a detailed sensitivity analysis of the bandwidth allocation function. Fluid and diffusion approximations are presented in Section 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 in this invariant point, leading to a multivariate normal law for the diffusion approximation. In Section 6, we illustrate our results with some examples. Proofs can be found in the remainder of the paper.

2 The Model

Consider a network consisting of J resources and I routes. Each resource is given an index $j \in \{1, 2, \dots, J\}$ and a route i is a non-empty subset of $\{1, 2, \dots, J\}$. We 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 direction in which information is being passed through the resources. 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 processing rates assigned to flows on a particular route is the bandwidth devoted to route i , which we denote by Λ_i . Each resource j is assigned a limited amount of total bandwidth, c_j , which it must distribute to each of the routes passing through it. We therefore obtain the matrix inequality $A\Lambda \leq c$, where $\Lambda = (\Lambda_1, \Lambda_2, \dots, \Lambda_I)'$ and $c = (c_1, c_2, \dots, c_J)'$. In addition, we assign to each route i a maximum rate m_i at which flows on that route may be processed. Thus, if z_i represents the total number of flows on route i , we require that $\Lambda_i \leq m_i z_i$. In matrix notation, this may be written as $Imz \geq \Lambda$, where I is the $I \times I$ identity matrix and $m = (m_1, m_2, \dots, m_I)'$ and $z = (z_1, z_2, \dots, z_I)'$.

Now let $Z_i(t)$ be the number of flows on route i at time t and set $Z(t) = (Z_1(t), Z_2(t), \dots, Z_I(t))'$. In what follows, we suppose that there exists some bandwidth sharing policy $\Lambda : \mathbb{R}^I \mapsto \mathbb{R}^I$ such that if the total number of flows on each route at time t is given by $Z(t)$, then the bandwidth devoted to route i by each resource on route i is given by $\Lambda_i(Z(t))$. In Section 3, detailed prescriptions of bandwidth allocation policies Λ are given. However, we first state how the system operates given an arbitrarily chosen bandwidth allocation policy Λ .

For each $i = 1, \dots, I$, let $(E_i(t), t \geq 0)$ be a rate one Poisson processes and let $\eta_i > 0$ be the average arrival rate of flows to route i . We furthermore make the assumption that $(E_i(t), t \geq 0)$, $i = 1, \dots, I$, are independent of one another. The number of flows which have arrived to route i by time t is then given by $E_i(\eta_i t)$. For $k \geq 1$, the size of the k th job to arrive to route i is an exponential random variable with mean $1/\mu_i$. We assume that job sizes are independent of one another both within a specific route and between routes. At time zero, we assume that there are $Z_i(0)$ flows on route i each of whose unprocessed information sizes are i.i.d. exponential random variables with mean $1/\mu_i$. 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 $\gamma_i \geq 0$ to have all of its information processed before abandoning from the system. Note that this implies that some flows

may depart prematurely from the system before having all of their information processed.

At a given point in time $t \geq 0$, each of the $Z_i(t)$ flows along route i must equitably share the total bandwidth allocated to the route. Since the total bandwidth assigned to route i is $\Lambda_i(t)$, this implies that the bandwidth allocated to each of the $Z_i(t)$ flows along route i is given by $\Lambda_i(Z(t))/Z_i(t)$, which we define to be 0 if $Z_i(t) = 0$.

Now note that since the information sizes for each flow are exponentially distributed, it follows that the departure process of service completions of flows of type i is a doubly stochastic Poisson process with instantaneous rate $Z_i(t)(\Lambda_i(Z(t))/Z_i(t))\mu_i = \Lambda_i(Z(t))\mu_i$. Moreover, the departure of flows of type i abandoning from the system is also a doubly stochastic Poisson process with rate $\gamma_i Z_i(t)$. Letting $N_i^s = (N_i^s(t), t \geq 0)$ and $N_i^a = (N_i^a(t), t \geq 0)$, $i = 1, \dots, I$, be independent, rate one Poisson processes, the process $Z = ((Z_1(t), \dots, Z_I(t)), t \geq 0)$ tracking the total number of flows on each route at each point time t , is given by the solution to the system of equations,

$$Z_i(t) = Z(0) + E_i(\eta_i t) - N_i^s \left(\mu_i \int_0^t \Lambda_i(Z(s)) ds \right) - N_i^a \left(\gamma_i \int_0^t Z_i(s) ds \right), \quad (1)$$

for $i = 1, \dots, I$, and $t \geq 0$. This is a set of I equations and it is straightforward to show that there exists a unique solution Z . In the parlance of [27], (1) is referred to as a time-homogeneous Markovian service network.

3 The bandwidth allocation mechanism and its properties

We now assign to each flow on route i a utility function U_i . We assume that the utility of a flow is a function of the bandwidth allocated to that flow. Hence, if Λ_i total units of bandwidth have been allocated to route i and there are $z_i > 0$ flows on route i at a given point in time, then each flow will receive a utility of $U_i(\Lambda_i/z_i)$. We assume that U_i is a strictly increasing, strictly concave, twice differentiable function, such that $U_i'(0) = \infty$. One important family of utility functions are the weighted α fair utility functions, given by $U_i'(x) = \kappa_i x^{-\alpha}$, with $\alpha > 0$.

For a given choice of utility functions U_i , $i = 1, \dots, I$, we next consider the bandwidth sharing policy Λ which arise as the solution to the following global utility maximization problem.

$$\begin{aligned} (P_z) \quad & \max && \sum_{i=1}^{\mathbf{I}} z_i U_i \left(\frac{\Lambda_i}{z_i} \right) \\ & \text{subject to} && A\Lambda \leq C \\ & && \Lambda \leq mIz. \end{aligned}$$

Note that since the criterium function in the above nonlinear program is strictly concave, it follows that for each $z = (z_1, \dots, z_I)$ on the interior of the positive orthant, there exists a unique optimal solution to this problem, which we denote by $\Lambda(z) = (\Lambda_1(z), \dots, \Lambda_I(z))$.

In this section, we show that our choice of $\Lambda(\cdot)$ as defined above is a directionally differentiable, locally Lipschitz function on $(0, \infty)^I$, which in some cases is also differentiable. The results in this section may be viewed as extensions of Lemma A.3 of Kelly & Williams (2004), and Proposition 4.1 of Borst *et al.* (2009), who respectively established continuity and Lipschitz continuity of $\Lambda(\cdot)$ as defined above for special cases. Our results in this section can be seen as relatively straightforward examples of the perturbation analysis of nonlinear optimization problems in Banach spaces, a theory which is nowadays well developed. These techniques however do not seem to have found widespread use in the Applied Probability community so far. Our main reference for this section is Bonnans & Shapiro (2000) and all proofs for this section may be found in the appendix.

In order to be clear, we note that throughout this section we assume that z lies in the interior of the positive orthant, i.e. $z_i > 0$ for all i . Furthermore, we allow the possibility that $m_i = \infty$. In our examples later on, A is an incidence matrix, i.e. $A_{ji} = 1$ if resource j is used by class of type i , but we note that our results still hold if the elements of A would be nonnegative, or if A is such that the interior of the Polyhedral capacity set $\{\Lambda : A\Lambda \leq C, \Lambda \geq 0\}$ is non-empty. This implies the validity of several constraint qualifications, in particular the Fromovitz-Mangasarian constraint qualification.

Now let $p(z)$ be a J dimensional vector of Lagrange multipliers corresponding to the capacity constraints $A\Lambda \leq C$, and let $q(z)$ be a I dimensional vector of Lagrange multipliers corresponding to the rate constraints $\Lambda \leq mz$. It then follows $p(z)$, $q(z)$ and $\Lambda(z)$ form a solution of the Karush-Kuhn Tucker (KKT) conditions for P_z , i.e.

$$(A\Lambda(z) - C)p(z) = 0, \quad (\Lambda(z) - mIz)q(z) = 0, \quad U'_i(\Lambda_i(z)/z_i) = q_i(z) + \sum_j p_j(z)A_{ji}.$$

Let $\mathcal{I}(z)$ be the set of active rate constraints, and let $\mathcal{J}(z)$ be the set of active capacity constraints of (P_z) . Denote by $\gamma(\Lambda, z)$ the set of Lagrange multipliers of (P_z) and from now on, let $(p(z), q(z))$ be Lagrange multipliers of (P_z) that also solve the optimization problem

$$\max_{(p,q) \in \gamma(\Lambda, z)} - \sum_{i=1}^I d_i m_i q_i. \quad (2)$$

The following is now the first main result of this section and serves as one of the major tools in establishing the main limit theorem of this paper. Recall that the directional derivative of Λ in a direction d at a point z is defined as $\lim_{t \downarrow 0} (\Lambda(z + td) - \Lambda(z))/t$, assuming that this limit exists.

Theorem 3.1. *Let $z \in (0, \infty)^I$. Define*

$$v_i(z) = \frac{1}{z_i} U''_i(\Lambda_i/z_i), \quad u_i(z) = \frac{\Lambda_i(z)}{z_i} v_i(z). \quad (3)$$

$\Lambda(\cdot)$ is directionally differentiable in any direction d , and the directional derivative $H_d(z)$ is the

unique solution to the following quadratic programming problem.

$$\begin{aligned}
(D_{z,d}) \quad & \max && -2 \sum_{i=1}^I u_i(z) d_i h_i + \sum_{i=1}^I v_i(z) h_i^2 \\
\text{subject to} \quad & (Ah)_j = 0, && \text{if } p_j(z) > 0, \\
& (Ah)_j \leq 0, && \text{if } (A\Lambda(z))_j = C_j, \\
& h_i = d_i m_i, && \text{if } q_i(z) > 0, \\
& h_i \leq d_i m_i, && \text{if } \Lambda_i(z) = m_i z_i.
\end{aligned}$$

An implication of this result is the following:

Theorem 3.2. $\Lambda(\cdot)$ is locally Lipschitz on $(0, \infty)^I$, i.e. for every compact subset E of $(0, \infty)^I$ there exists a constant K_E such that $\|\Lambda(x) - \Lambda(y)\| \leq K_E \|x - y\|$ for all $x, y \in E$.

Some of our results require the stronger result of $\Lambda(\cdot)$ being differentiable at a point z , which, in view of the previous theorem, is equivalent to $H_d(z)$ being a linear function in d . A sufficient condition for differentiability is given in the following theorem. Recall that the *strict complementarity condition* holds if all active constraints 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 $i \in \mathcal{I}(z)$. Also, recall that a set of constraints are linearly independent if the coefficient vectors on the l.h.s. of these constraints cannot be written as a combination of one another.

Theorem 3.3. $\Lambda(\cdot)$ is differentiable at a point $z \in (0, \infty)^I$ if the constraints in $\mathcal{I}(z) \cup \mathcal{J}(z)$ are linearly independent, and if the strict complementarity condition holds. In this case, $H_d(z)$ is the solution of

$$\begin{aligned}
(D'_{z,d}) \quad & \max && -2 \sum_{i=1}^I u_i(z) d_i h_i + \sum_{i=1}^I v_i(z) h_i^2 \\
\text{subject to} \quad & (Ah)_j = 0, && j \in \mathcal{J}(z), \\
& h_i = d_i m_i, && i \in \mathcal{I}(z).
\end{aligned}$$

Let $p^d(z)$ and $q^d(z)$ be the vectors of Lagrange multipliers. Then $H^d(z), p^d(z), q^d(z)$ form the unique solution of the system of $I + |\mathcal{I}(z)| + |\mathcal{J}(z)|$ linear equations

$$\begin{aligned}
2H_i^d(z)v_i(z) &= 2u_i(z) + \sum_{j \in \mathcal{J}(z)} p_j^d(z) A_{ji} + q_i^d(z) I(i \in \mathcal{I}(z)), \quad i = 1, \dots, I, \\
(Ah^d(z))_j &= 0, \quad j \in \mathcal{J}(z), \\
H_i^d &= d_i m_i, \quad i \in \mathcal{I}(z).
\end{aligned}$$

The proof of Theorem 3.3 follows immediately from Theorem 3.1, exploiting strictly complementarity and linear independence of the constraints.

Note that since active constraints are required to be independent, we have that $|\mathcal{I}(z)| + |\mathcal{J}(z)| \leq I$ and so if the derivative of $\Lambda(\cdot)$ exists, it may be found by solving a system of at most $2I$ equations. Thus, from a computational standpoint, finding the derivative of Λ is not much more difficult than finding Λ itself.

4 Fluid and diffusion limits

We now introduce what we refer to as the ‘large capacity scaling’ regime. Recall that as mentioned in the introduction, it is natural to consider a regime in which the capacity of each resource in the network and arrival rates to the network are arbitrarily large, while individual user rate constraints remain bounded. In order to model this, we introduce a sequence of networks indexed by n where the bandwidth allocation policy Λ^n in the n th system is given by the solution of (P_z) , with the capacity vector C replaced by nC . One may easily verify that this implies that

$$\Lambda^n(z) = n\Lambda(z/n), \quad (4)$$

and hence the bandwidth allocation policy scales in a natural way as well in this regime. We also denote by η_i^n the arrival rate of flows to route i in the n th network, which we assume grows to infinity at a linear rate as n grows large, i.e. $\eta_i^n = n\eta_i$.

Our main results in this section are to provide fluid and diffusion limits for the user population process introduced in Section 2. Define first the fluid scaled quantities, $\bar{Z}^n(0) = n^{-1}Z^n(0)$ and $\bar{Z}^n(t) = n^{-1}Z^n(t)$ for $t \geq 0$, and set $\bar{Z}^n = (\bar{Z}^n(t), t \geq 0)$. We then have the following fluid limit result.

Theorem 4.1. *If $\bar{Z}^n(0) \Rightarrow \bar{Z}(0) \in (0, \infty)^I$ as $n \rightarrow \infty$, and $n^{-1}\eta_i^n \rightarrow \eta_i$ as $n \rightarrow \infty$, for $i = 1, \dots, I$, then $\bar{Z}^n \Rightarrow \bar{Z}$ as $n \rightarrow \infty$, where \bar{Z} is the unique, strong solution to the system of equations,*

$$\bar{Z}_i(t) = \bar{Z}_i(0) + \eta_i t - \mu_i \int_0^t \Lambda_i(\bar{Z}(s)) ds - \gamma_i \int_0^t \bar{Z}_i(s) ds, \quad (5)$$

for $t \geq 0$, for $i = 1, \dots, I$.

Note that the system of equations (5) represents an autonomous system of ordinary differential equations and it may be solved numerically using an iterative approach.

We now assume that $\bar{Z}(0)$ in the statement of Theorem 4.1 is constant. By (5), this then implies that \bar{Z} is a deterministic function. In our next result, we center the user population process Z^n by \bar{Z} and then rescale by a factor of \sqrt{n} . For each $n \geq 1$, define the diffusion scaled quantities $\tilde{Z}^n(t) = n^{1/2}(\bar{Z}^n(t) - \bar{Z}(t))$ for $t \geq 0$ and set $\tilde{Z}^n = (\tilde{Z}^n(t), t \geq 0)$. The following result provides a weak limit for the diffusion scaled user population process \tilde{Z}^n and it is the main result of this section.

Theorem 4.2. Let $\tilde{\xi}_i = (\tilde{\xi}_i(t), t \geq 0)$, $i = 1, \dots, I$, be independent, Brownian motions with infinitesimal variance

$$\sigma_i(t) = \sqrt{\eta_i + \mu_i \Lambda_i(\bar{Z}(t)) + \gamma_i \bar{Z}_i(t)},$$

for $t \geq 0$.

If $\tilde{Z}^n(0) \Rightarrow \tilde{Z}(0)$ as $n \rightarrow \infty$, and $\sqrt{n}(n^{-1}\eta_i^n - \eta_i) \rightarrow \beta_i$ as $n \rightarrow \infty$, for $i = 1, \dots, I$, then $\tilde{Z}^n \Rightarrow \tilde{Z}$ as $n \rightarrow \infty$, where \tilde{Z} is the unique, strong solution to the stochastic differential equation

$$\tilde{Z}_i(t) = \tilde{Z}_i(0) + \tilde{\xi}_i(t) + \beta_i e - \mu_i \int_0^t (H_{\bar{Z}(s)})_i(\bar{Z}(s)) ds - \gamma_i \int_0^t \tilde{Z}_i(s) ds, \quad (6)$$

for $t \geq 0$, and $i = 1, \dots, I$.

In general, the limit process \tilde{Z} in Theorem 4.2 may exhibit complex behavior depending on \bar{Z} and the smoothness of $\Lambda(\cdot)$ at each point along $\bar{Z}(t)$ along the fluid limit path \bar{Z} . We now, however, present several cases in which the diffusion limit (6) may be explicitly solved. We first assume that the bandwidth allocation policy $\Lambda(\cdot)$ is differentiable at each point along the path of its fluid limit \bar{Z} . This then implies that $H_d(\bar{Z}(t))$ is a linear function of d for each $t \geq 0$ and so we may write $H_d(\bar{Z}(t)) = H(\bar{Z}(t))d$ for a particular matrix $H(\bar{Z}(t))$. Let $\tilde{\xi}(t) = (\tilde{\xi}_1(t), \dots, \tilde{\xi}_I(t))'$ for $t \geq 0$ and let $\beta = (\beta_1, \dots, \beta_I)'$. Also, let I_μ be the $I \times I$ matrix such that $(I_\mu)_{ii} = \mu_i$ for $i = 1, \dots, I$, and zero otherwise, and, similarly, let I_γ be the $I \times I$ matrix with $(I_\gamma)_{ii} = \gamma_i$ for $i = 1, \dots, I$, and zero otherwise. We then have the following result. Its proof is standard and an example of it may be found, for instance, in [24].

Proposition 4.3. Suppose that the conditions of Theorem 3.3 hold at $\bar{Z}(t)$ for each $t \geq 0$ and let $\Phi(t)$ be solution to the matrix-valued ODE

$$\dot{\Phi}(t) = -(I_r H(\bar{Z}(t)) + I_\gamma) \Phi(t),$$

for $t \geq 0$, with initial condition $\Phi(0) = I$. The solution to (6) is then given by

$$\tilde{Z}(t) = \Phi(t) \left(\tilde{Z}(0) + \int_0^t \Phi^{-1}(s) d(\tilde{\xi}(s) + \beta s) \right), \quad (7)$$

for $t \geq 0$. Moreover, if $E[||\tilde{Z}(0)||^2] < \infty$, then

$$E[\tilde{Z}(t)] = \Phi(t) \left[E[\tilde{Z}(0)] + \int_0^t \Phi^{-1}(s) \beta ds \right],$$

and, for $0 \leq s \leq t$,

$$\begin{aligned} & E[(Z(s) - E[\tilde{Z}(s)])(Z(t) - E[\tilde{Z}(t)])'] \\ &= \Phi(s) \left[E[(Z(0) - E[\tilde{Z}(0)])(Z(0) - E[\tilde{Z}(0)])'] + \int_0^{s \wedge t} (\Phi^{-1}(u) \sigma(u)) (\Phi^{-1}(u) \sigma(u))^T du \right] \Phi'(t), \end{aligned}$$

where $\sigma(u)$ is the I -dimensional column vector with i th component

$$\sigma_i(u) = \sqrt{\eta_i + \mu_i \Lambda_i(\bar{Z}(u)) + \gamma_i \bar{Z}_i(u)}.$$

Also, if $\tilde{Z}(0)$ is Gaussian distributed, then $\tilde{Z}(t)$ is Gaussian as well.

Suppose now that $\bar{Z}(0)$ is an invariant point for the fluid limit ODE (5) so that $\bar{Z}(t) = \bar{Z}(0)$ for $t \geq 0$. This then implies that $H(\bar{Z}(t)) = H(\bar{Z}(0))$ for all $t \geq 0$ and, that

$$\sigma_i(u) = \sigma_i = \sqrt{\eta_i + \mu_i \Lambda_i(\bar{Z}(0)) + \gamma_i \bar{Z}_i(0)},$$

for $u \geq 0$. From (6), we then conclude that \tilde{Z} is a time-homogeneous, multi-dimensional Ornstein-Uhlenbeck process. In Section 5, a precise characterization of all invariant points for (5) is given. We close this section with the following result whose proof is also standard and may be found in [24].

Proposition 4.4. *Assume that $\bar{Z}(0)$ is an invariant point for (5) and let $A = I_r H(\bar{Z}(0)) + I_\gamma$. If each eigenvalue of A has a positive real part, then $\tilde{Z}(t) \Rightarrow \tilde{Z}(\infty)$ as $t \rightarrow \infty$, where $\tilde{Z}(\infty)$ is a normal random variable with mean*

$$E[\tilde{Z}(\infty)] = \beta \int_0^\infty e^{-tA} dt,$$

and variance-covariance matrix

$$E[(\tilde{Z}(\infty) - E[\tilde{Z}(\infty)])(\tilde{Z}(\infty) - E[\tilde{Z}(\infty)])'] = \int_0^\infty e^{-tA} \sigma \sigma' e^{-tA'} dt.$$

5 Invariant points

Recall now that η_i is the arrival rate of flows of class i , μ_i is the inverse of the mean service time of flows of type i , and γ_i is the reneging rate of flows of type i . Consider those points $z \in (0, \infty)^I$ which satisfy

$$\eta_i = \mu_i \Lambda_i(z) + \gamma_i z_i. \tag{8}$$

We refer to such points as invariant points for the fluid limit and we sometimes also call them fixed points. Our main result in this section is to show that assuming the system is not overloaded, there exists a unique fixed point.

Consider the first the case of $\gamma_i = 0$ for all i , i.e. no impatience. We then have that (8) reduces to

$$\Lambda(z) = \rho. \tag{9}$$

From (9), it follows that there do not exist invariant points unless $A\rho \leq C$. In other words, the system is not overloaded. We therefore assume that $A\rho \leq C$. In order to determine which values of z satisfy (9), we observe that the KKT conditions for (P_z) , combined with (9), yield

$$U'_i(\rho_i/z_i) = q_i + \sum_j A_{ji}p_j, \quad p_j(A\rho - C)_j = 0, \quad q_i(\rho_i - m_i z_i) = 0.$$

From these equations and the non-negativity constraints on the Lagrange multipliers, it follows that z is an invariant point for (P_z) if and only if there exists constants non-negative constants p_j and q_i such that

$$z_i = \frac{\rho_i}{U_i'^{-1}\left(q_i + \sum_j A_{ji}p_j\right)}. \quad (10)$$

Note also p_j and q_i must further satisfy $p_j = 0$ if $(A\rho)_j < C_j$, and $q_i = (U'_i(m_i) - \sum_j p_j A_{ji})^+$. The sufficiency of the above statement is straightforward. The only non-trivial part for necessity is to observe the fact the the KKT conditions are necessary, and to verify the expression of q_i . The latter follows by examining the three cases where $U'_i(m_i) - \sum_j p_j A_{ji}$ is strictly positive, zero, or strictly negative separately, and by the necessity of the KKT conditions. Note that in the underloaded case of $(A\rho)_j < C_j$ for every j , then $p_j = 0$ for all j , so that $q_i = U'_i(m_i)$, implying $z_i = \rho_i/m_i$.

The characterization of invariant points in the presence of impatience is more challenging. In what follows, we assume $\eta_i > 0$ for all i . Suppose that z is an invariant point for (P_z) , so that it satisfies (8). Our approach now is to substitute this expression for z into the KKT conditions for (P_z) , and to identify a concave programming problem for which the new set of equations forms the KKT conditions. This idea has been applied to a separate problem in Borst *et al.* (2009). This substitution results in the equations

$$U'_i(\Lambda_i \gamma_i / (\eta_i - \mu_i \Lambda_i)) = q_i + \sum_j p_j A_{ji}, \quad p_j((A\Lambda)_j - C_j) = 0, \quad q_i(\Lambda_i - \frac{d_i \eta_i}{\gamma_i + d_i \mu_i}) = 0.$$

Define the function G_i by $G'_i(x) = U'_i(x\gamma_i/(\eta_i - \mu_i x))$. We see that G'_i is strictly increasing, hence G_i is strictly concave.

Define now the optimization problem (Q) as follows:

$$\max_{\Lambda} \sum_i G_i(\Lambda_i)$$

subject to

$$A\Lambda \leq C, \quad \Lambda_i \in \left[0, \frac{m_i \eta_i}{\gamma_i + m_i \mu_i}\right].$$

Note that the RHS is smaller than ρ_i . Given the fixed point z , set $\Lambda^* = \Lambda(z)$, $p^* = p(z)$ and $q^* = q(z)$.

It follows easily that these parameters form a solution to the KKT conditions for (Q) (since the conditions can be rewritten in terms of those of (P) , using the fixed point equation for z). Thus, Λ^* solves (Q) .

Thus, if z is a fixed point, then $\Lambda(z)$ is an optimal solution of (Q) . Since (Q) must have a unique solution (the criterium function is strictly concave), $\Lambda(z)$ is unique. Since z can be expressed explicitly in terms of $\Lambda(z)$ (using the fixed point equation), it follows that z is unique as well. Actually, a reverse argument ensures that z exists and is computable: simply solve (Q) , and its solution Λ yields the desired fixed point.

In addition, $\Lambda(\cdot)$ 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 since the Lagrange multipliers for (P_z) can be expressed in those of (Q) , and the active constraints are identical in both programs.

6 Examples

We now provide examples illustrating how the methodology of this paper may be used to analyze different bandwidth sharing networks.

6.1 Connection with queues in the Halfin-Whitt regime

We first provide a basic example that connects our work with some result from the call center queueing literature. We consider a network with a single resource and single route so that $I = J = 1$. Furthermore, we assume that $\eta_1 = 1$ and that $C_1 = m_1 = \mu = 1$. In this case, it is straightforward to see that the bandwidth allocation function is simply $\Lambda(z) = \min\{z, 1\}$ for any utility function $U_1(\cdot)$ and so our model is equivalent to an $M/M/n$ queue. Also, since $\rho = 1$, the set of invariant points is $[1, \infty)$. In order to simplify matters, we now take $U_1(x) = \log x$ and, in the ‘large capacity scaling’ regime, we assume that $\eta^n = n - \beta\sqrt{n} + o(\sqrt{n})$.

We now determine the directional derivative of Λ and the point $z = 1$. It is clear from the explicit expression $\Lambda(z) = \min\{z, 1\}$ that $H^d(1) = d1(d < 0)$, however, it is instructive to obtain this result from the general procedure outlined in Theorem 3.1. Since both the capacity and the rate constraints are active at $z = 1$, all non-negative pairs (p, q) with $p + q = 1$ are Lagrange multipliers. Problem (2) is to minimize dq and results in $p(1), q(1) = (0, 1)$ if $d < 0$ and $p(1), q(1) = (1, 0)$ if $d > 0$. Consequently, $H^d(1)$ is the maximizing value of the quadratic function $2dh - h^2$ 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.

Applying Theorem 4.2, we now conclude that the diffusion limit for the user population process becomes the solution to the SDE

$$d\tilde{Z}(t) = -(\beta + \tilde{Z}(t)^-)dt + \sqrt{2}dW(t), \quad t \geq 0, \quad (11)$$

where $W(\cdot)$ is a standard Brownian motion. Assuming $\beta > 0$, the steady-state distribution of this diffusion is not Gaussian but is still computable. We refer to Halfin & Whitt [22] for further details.

6.2 A single link with multiple customer classes

We now consider a network with a single resource but multiple routes so that $J = 1$ and $I \geq 1$. We assume that the resource operates at unit capacity so that $c = 1$. We assume that the utility functions are given by weighted proportional fairness, i.e. $U_i(x) = \kappa_i \log x$. As it turns out, the ratio κ_i/d_i plays a significant role in our analysis, suggesting that users with a higher maximal service rate should be given more weight in the service allocation. The computations in this section may be used for pricing schemes. For example, users may have to pay a price for a certain value of d_i and a service provider can optimize over κ_i to maximize profit.

The program (Q) determining the equilibrium fluid point can be solved explicitly in this case as follows. Set $\bar{\rho}_i = \frac{d_i \eta_i}{\theta_i + d_i \mu_i}$. If $\sum_i \bar{\rho}_i \leq 1$, then the equilibrium fluid point is given by $\Lambda_i^* = \bar{\rho}_i$, and $z_i^* = \frac{\eta_i}{\theta_i d_i + \mu_i}$, implying that all classes are served on maximal scale when served in isolation. If $\sum_i \bar{\rho}_i = 1$, then the active constraints are not independent, implying that $\Lambda(\cdot)$ is not differentiable at z^* .

We next focus on the overloaded case $\sum_i \bar{\rho}_i > 1$. In this case, it is not difficult to verify that the solution to (Q) can be computed as follows:

1. Order the indices i such that $\frac{\kappa_1}{d_1} \leq \frac{\kappa_2}{d_2} \dots$
2. Solve p^* from

$$\sum_{i=1}^I \frac{\eta_i}{\frac{\theta_i}{\kappa_i} \min\{\frac{\kappa_i}{d_i}, p\} + \mu_i} = 1,$$

and set $i^* = \max\{i : \frac{\kappa_i}{d_i} < p^*\}$.

3. Set $\Lambda_i^* = \frac{\eta_i}{\frac{\theta_i}{\kappa_i} \min\{\frac{\kappa_i}{d_i}, p^*\} + \mu_i}$, and note that $\Lambda_i^* = \rho_i^*$ if $i > i^*$.

Note that $p^* \geq \frac{\kappa_{i^*+1}}{d_{i^*+1}}$. If equality holds, then $q_{i^*+1} = 0$, implying that the strict complementarity condition will not hold, and $\Lambda(\cdot)$ will not be differentiable in the point z^* given by $z_i^* = \frac{1}{\theta_i} (\eta_i - \mu_i \Lambda_i^*)$. If $p^* > \frac{\kappa_{i^*+1}}{d_{i^*+1}}$, then $\Lambda(\cdot)$ is differentiable at z^* , and its derivative can be computed by solving a simple set of linear equations.

The connection with the QD, ED and QED regimes mentioned in the introduction is the following. One may say that all flows of type $1 \leq i \leq i^*$ operate in the ED regime while all other flows of type $i > i^*$ operate in the QD regime, with the exception of flow $i^* + 1$, which operates in the QED regime when $p^* = \frac{\kappa_{i^*+1}}{d_{i^*+1}}$. Note also that all flows operate in the ED regime when κ_i is chosen proportional to d_i .

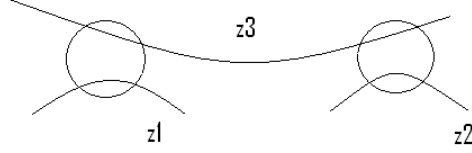


Figure 1: A linear network with two nodes and three routes.

6.3 A linear network

As our final example, we consider a linear network which consists of I nodes and $J = I + 1$ routes. Routes 1 through I are the singletons $\{1\}$ through $\{I\}$, respectively, while route $I + 1$ is equal to $\{1, \dots, I\}$. An illustration of a linear network with $I = 2$ is provided in the figure above. Intuitively, one expects that if route $I + 1$ is assigned a high utility function, it will tend to draw capacity away from the remaining routes.

In order to begin our analysis, we assume that we are at a point $z \in \mathbb{R}_+^{I+1}$ such that all constraints in (P_z) are active. We then have that the KKT conditions for (P_z) reduce to the system of equations $q_i(z) = U'_1(\Lambda_i/z_i) - p_i(z)$ for $i = 1, \dots, I$ and $q_{I+1}(z) = U'_{I+1}(\Lambda_{I+1}/z_{I+1}) - \sum_{i=1}^I p_i(z)$. This then implies that the solution to problem (2) may be expressed as the solution to the linear program

$$\begin{aligned}
 R_z \quad & \max && \sum_{i=1}^I (d_i m_i + d_{I+1} m_{I+1}) p_i(z) \\
 & \text{subject to} && p_i(z) \leq U'_i(\Lambda_i/z_i), i = 1, \dots, I, \\
 & && \sum_{i=1}^I p_i(z) \leq U'_{I+1}(\Lambda_{I+1}/z_{I+1}) \\
 & && p_i(z) \geq 0, i = 1, \dots, I.
 \end{aligned}$$

Now relabel indices such 1 through I such that $d_i m_i \leq d_{i+1} m_{i+1}$, for $i = 1, \dots, I - 1$, and let

$$\pi(z) = \max \left\{ j > 0 : \sum_{i=j}^I U'_i(\Lambda_i/z_i) \geq U'_{I+1}(\Lambda_{I+1}/z_{I+1}) \right\}. \quad (12)$$

Solving the dual of (R_z) and then relating the solution back to the primal problem, one obtains

that the solution to (R_z) is given by $p_i(z) = 0$ for $1 \leq i < \pi(z)$, $p_i(z) = U'_i(\Lambda_i/z_i)$ for $i > \pi(z)$, and

$$p_{\pi(z)} = U'_{I+1}(\Lambda_i/z_{I+1}) - \sum_{i=\pi(z)+1}^I U'_i(\Lambda_i/z_i).$$

Now using the KKT conditions for (P_z) , one sees that $q_i(z) = U'_i(\Lambda_i/z_i)$ for $1 \leq i \leq \pi(z) - 1$, $q_i(z) = 0$ for $i > \pi(z)$ and, by complimentary slackness, that

$$q_{\pi(z)} = \sum_{i=\pi(z)}^I U'_i(\Lambda_i/z_i) - U'_{I+1}(\Lambda_i/z_{I+1}).$$

Now note that by (12) we have that $p_{\pi(z)} > 0$ and let us assume first that $q_{\pi(z)} > 0$. The linear program $(D_{z,d})$ in Theorem 3.1 is then given by

$$\begin{aligned} (D_{z,d}) \quad & \max && -2 \sum_{i=1}^{I+1} u_i(z) d_i h_i + \sum_{i=1}^{I+1} v_i(z) h_i^2 \\ \text{subject to} \quad & h_j = -h_{I+1} = 0, && j = \pi(z), \dots, I, \\ & h_j \leq -h_{I+1}, && j = 1, \dots, \pi(z) - 1, \\ & h_i = d_i m_i, && i = 1, \dots, \pi(z), \\ & h_i \leq d_i m_i, && i = \pi(z) + 1, \dots, I + 1. \end{aligned}$$

It is straightforward to see that the only feasible solution to this linear program is $h_i = d_i m_i$ for $i = 1, \dots, \pi(z)$, $h_i = d_{\pi(z)} m_{\pi(z)}$ for $\pi(z) < i \leq I$ and $h_{I+1} = -d_{\pi(z)} m_{\pi(z)}$.

On the other hand, suppose that $q_{\pi(z)} = 0$. Then, the linear program $(D_{z,d})$ in Theorem (3.1) becomes

$$\begin{aligned} (D_{z,d}) \quad & \max && -2 \sum_{i=1}^{I+1} u_i(z) d_i h_i + \sum_{i=1}^{I+1} v_i(z) h_i^2 \\ \text{subject to} \quad & h_j = -h_{I+1} = 0, && j = \pi(z), \dots, I, \\ & h_j \leq -h_{I+1}, && j = 1, \dots, \pi(z) - 1, \\ & h_i = d_i m_i, && i = 1, \dots, \pi(z) - 1, \\ & h_i \leq d_i m_i, && i = \pi(z), \dots, I + 1. \end{aligned}$$

Noting that the feasible region to this linear program may be reduced, we then see that the optimal solution to $(D_{z,d})$ is given by $h_i = d_i m_i$ for $i = 1, \dots, \pi(z) - 1$, $h_j = -h^*$ for $j = \pi(z), \dots, I$, and $h_{I+1} = h^*$, where h^* is the optimal solution to

$$\begin{aligned} (D_{z,d}) \quad & \max && 2 \left(-u_{I+1}(z) d_{I+1} + \sum_{i=1}^I u_i(z) d_i \right) h + \sum_{i=1}^{I+1} v_i(z) h^2 \\ \text{subject to} \quad & -d_{\pi(z)} m_{\pi(z)} \leq h \leq -d_{\pi(z)-1} m_{\pi(z)-1}. \end{aligned}$$

7 Proofs of properties of the bandwidth sharing function

In this section, we provide proofs of the results in Section 3. In particular, we provide results on the differentiability of Λ as a function of z . We first provide the proof of Theorem 3.1. We follow Section 5.2.3 of Bonnans & Shapiro (2000).

Proof of Theorem 3.1. Let

$$\begin{aligned} \max \quad & \sum_{i=1}^I U_i' \left(\frac{\Lambda_i}{z_i} \right) h_i + \left(U_i \left(\frac{\Lambda_i}{z_i} \right) - \frac{\Lambda_i}{z_i} U_i' \left(\frac{\Lambda_i}{z_i} \right) \right) d_i \\ \text{subject to} \quad & (Ah)_j \leq 0, \quad j \in \mathcal{J}(z), h_i \leq m_i d_i, \quad i \in \mathcal{I}(z) \end{aligned}$$

be the linearization of P_z at $(\Lambda(z), z)$.

The Lagrangian associated with P_z may be written as

$$L(\Lambda, z, p, q) = \sum_{i=1}^I z_i U_i \left(\frac{\Lambda_i}{z_i} \right) + \sum_{j=1}^J p_j ((A\Lambda)_j - C_j) + \sum_{i=1}^I q_i (\Lambda_i - m_i z_i).$$

Denote the set $\gamma(\Lambda, z)$ of Lagrange multipliers of P_z . The dual to the linearization of the above linear program may be written as

$$\max_{(p,q) \in \gamma(\Lambda,z)} D_z L(\Lambda, z, p, q) d.$$

which may be written explicitly as

$$\max_{(p,q) \in \gamma(\Lambda,z)} \sum_{i=1}^I \left(U_i \left(\frac{\Lambda_i}{z_i} \right) - \frac{\Lambda_i}{z_i} U_i' \left(\frac{\Lambda_i}{z_i} \right) \right) q_i - \sum_{i=1}^I d_i m_i q_i. \quad (13)$$

Now let $\mathcal{S}(DL_d)$ denote the set of optimal solution to the dual of the linearized problem. It then follows that the set of optimal solution to the linearized problem may be written as

$$\mathcal{S}(PL_d) = \left\{ h : \begin{array}{l} (Ah)_j = 0, \text{ if } p_j(z) > 0, \\ (Ah)_j \leq 0, \text{ if } , (A\Lambda(z))_j = C_j \\ h_i = d_i m_i \text{ if } q_i(z) > 0 \\ h_i \leq d_i m_i \text{ if } \Lambda_i(z) = m_i z_i \end{array} \right.$$

Now let $O(\Lambda, z)$ be the value of the objective function of P_z and note that the Hessian of O with respect to Λ and z is equivalent to that of $L(\Lambda, z, p, q)$. In particular, we have that

$$\frac{\partial O}{\partial z_i \partial \Lambda_j} = 0$$

for $i \neq j$ and

$$\frac{\partial O}{\partial z_i \partial z_i} = \frac{\Lambda_i^2}{z_i^3} U_i'' \left(\frac{\Lambda_i}{z_i} \right), \quad \frac{\partial O}{\partial \Lambda_i \partial \Lambda_i} = \frac{1}{z_i} U_i'' \left(\frac{\Lambda_i}{z_i} \right), \quad \frac{\partial O}{\partial \Lambda_i \partial z_i} = -\frac{\Lambda_i}{z_i^2} U_i'' \left(\frac{\Lambda_i}{z_i} \right).$$

By (5.2.125) of Bonnans and Shapiro we now have that the directional derivative of Λ in a direction d is given by the optimal solution h to the optimization problem given by the Theorem. \square

We next provide the proof of Theorem 3.2.

Proof of Theorem 3.2. Let E be a compact subset of $(0, \infty)^I$. We show that there exists a constant K_E such that for any d with $\|d\| \leq 1$ and for each $z \in E$, $\|H_d(z)\| \leq K_E$. This then implies that Λ is Lipschitz continuous on E .

In order to begin, note that $H_d(z)$ is the optimal solution to the quadratic programming problem $D_{z,d}$ given by Theorem 3.1. Moreover, note that for a fixed d , there can exist at most $2^{2(I+J)}$ different combinations of constraints for $D_{z,d}$ and hence, $2^{2(I+J)}$ different feasible regions. Let us label these feasible regions by $f(d; 1), f(d; 2), \dots, f(d; 2^{2(I+J)})$ where d is meant to indicate the dependence on the choice of direction d . Note also that by the existence of the directional derivative, it follows that $f(d; k)$ is always non-empty. We now claim that there exists a compact set $\mathcal{C} \in \mathbb{R}^I$ such that for any choice of direction d with $\|d\| \leq 1$ and $k = 1, \dots, 2^{2(I+J)}$, there exists an element $x(d; k) \in \mathcal{C}$ such that $x(d; k) \in f(d; k)$.

Let e_l be the I -dimensional vector of all zeros except a 1 in the l th position and let $(e_l; k)$ be a point arbitrarily chosen from $f(e_l; k)$. Also, let $-e_l$ be the I -dimensional vector of all zeros except a -1 in the l th position and let $x(-e_l; k)$ be a point arbitrarily chosen from $f(-e_l; k)$. Next, let \mathcal{C} be the convex hull of the points $x(e_l; k)$ and $x(-e_l; k)$ for $l = 1, \dots, I$ and $k = 1, \dots, 2^{2(I+J)}$. Clearly \mathcal{C} is compact and so we now claim that \mathcal{C} has the required property above.

Let $d = (\alpha_1(d), \dots, \alpha_I(d)) \in \mathbb{R}^I$ such that $\|d\| \leq 1$ and let $k \in \{1, \dots, 2^{2(I+J)}\}$. We now claim that

$$x(d; k) = \sum_{l=1}^I (\max(\alpha_l(d), 0)x(e_l; k) + \max(-\alpha_l(d), 0)x(-e_l; k)) \quad (14)$$

is an element of both \mathcal{C} and $f(d; k)$. Since $\|d\| \leq 1$, $|\alpha_l(d)| \leq 1$ for $l = 1, \dots, I$ and so $x(d; k) \in \mathcal{C}$. Next note that if $(Ax(e_l; k))_j = 0$ and $(Ax(-e_l; k))_j = 0$ for $j \in \{1, \dots, J\}$ and $l = 1, \dots, I$, then

since by (14)

$$\begin{aligned}
(Ax(d; k))_j &= \left(A \left(\sum_{l=1}^I (\max(\alpha_l(d), 0)x(e_l; k) + \max(-\alpha_l(d), 0)x(-e_l; k)) \right) \right)_j \\
&= \sum_{l=1}^I (A (\max(\alpha_l(d), 0)x(e_l; k) + \max(-\alpha_l(d), 0)x(-e_l; k)))_j \\
&= \sum_{l=1}^I \max(\alpha_l(d), 0)(Ax(e_l; k))_j + \sum_{l=1}^I \max(-\alpha_l(d), 0)(Ax(-e_l; k))_j,
\end{aligned}$$

it follows that $(Ax(d; k))_j = 0$ as well. In a similar manner, if $(Ax(e_l; k))_j \leq 0$ and $(Ax(-e_l; k))_j \leq 0$ for $l = 1, \dots, I$, then since $\max(\alpha_l(d), 0) \geq 0$ and $\max(-\alpha_l(d), 0) \geq 0$, it follows that $(Ax(d; k))_j \leq 0$ as well. We therefore see that the first two sets of constraints in the quadratic programming problem will be satisfied. Next note that for $i = 1, \dots, I$,

$$x_i(d; k) = \sum_{l=1}^I (\max(\alpha_l(d), 0)x_i(e_l; k) + \max(-\alpha_l(d), 0)x_i(-e_l; k)).$$

Thus, since

$$d_i = \sum_{l=1}^I (\max(\alpha_l(d), 0)(e_l)_i + \max(-\alpha_l(d), 0)(-e_l)_i),$$

it follows that if $x_i(e_l; k) = m_i(e_l)_i$ and $x_i(-e_l; k) = m_i(-e_l)_i$ for $i \in \{1, \dots, I\}$ and $l = 1, \dots, I$, then $x_i(d; k) = m_i d_i$. In a similar manner, if $\max(\alpha_l(d), 0) \geq 0$ and $\max(-\alpha_l(d), 0) \geq 0$, if $x_i(e_l; k) \leq m_i(e_l)_i$ and $x_i(-e_l; k) \leq m_i(-e_l)_i$ for $i \in \{1, \dots, I\}$ and $l = 1, \dots, I$, then $x_i(d; k) \leq m_i d_i$. Thus, we see that the second two sets of constraints in the quadratic programming problem are satisfied as well and so the set \mathcal{C} satisfies the desired properties.

Now for each $i = 1, \dots, I$, let

$$v_i^u(E) = \sup_{z \in E} |v_i(z)| \text{ and } u_i^u(E) = \sup_{z \in E} |u_i(z)|,$$

and

$$v_i^l(E) = \inf_{z \in E} |v_i(z)| \text{ and } u_i^l(E) = \inf_{z \in E} |u_i(z)|.$$

Note that these values are finite since v_i and u_i are continuous on E and E is compact. Moreover, they are strictly positive. Let $z \in E$ and $d = (d_1, \dots, d_I) \in \mathbb{R}^I$ with $\|d\| \leq 1$. It then follows

that for any $h \in \mathbb{R}^I$, since U_i is strictly concave for each $i = 1, \dots, I$ and hence $v_i(z) < 0$ for all $z \in E \subset (0, \infty)^I$ and also since $|d_i| \leq 1$ for $i = 1, \dots, I$, we have that

$$-2 \sum_{i=1}^I u_i(z) d_i h_i + \sum_{i=1}^I v_i(z) h_i^2 \geq -2 \sum_{i=1}^I u_i^u(E) |h_i| - \sum_{i=1}^I v_i^u(E) h_i^2. \quad (15)$$

Next, define

$$\kappa(E) = \inf_{h \in \mathcal{C}} -2 \sum_{i=1}^I u_i^u(E) |h_i| - \sum_{i=1}^I v_i^u(E) h_i^2,$$

which is finite since \mathcal{C} is compact. Also note that $\kappa(\mathcal{E}) < 0$ as it is easily shown that $\mathcal{C} \neq \{0\}$. It then follows from (15) that the optimal value of $D_{z,d}$ must be at least $\kappa(E)$ since there exists at least one feasible point of $D_{z,d}$ in \mathcal{C} .

Now for each $i \in \{1, \dots, I\}$ note that

$$\sup_{h_i \in \mathbb{R}} -2u_i(z)h_i + v_i(z)h_i^2 \leq \sup_{h_i \in \mathbb{R}} 2u_i^u(E)|h_i| - w_i^l(E)h_i^2$$

where the inequality follows since $v_i(z) < 0$ and since both $u_i^u(E)$ and $w_i^l(E)$ are finite and $w_i^l(E)$ is strictly positive. Set

$$b_i(E) = \sup_{h_i \in \mathbb{R}} 2u_i^u(E)|h_i| - w_i^l(E)h_i^2,$$

and note that $b_i(E) > 0$. We may now write that if $h^*(z, d)$ is the optimal solution to $D_{z,d}$, then

$$\begin{aligned} \sum_{l=1, l \neq i}^I b_l(E) + (-2u_i(z)d_i h_i^* + v_i(z)(h_i^*)^2) &\geq -2 \sum_{l=1}^I u_l(z) d_l h_l^* + \sum_{l=1}^I v_l(z) (h_l^*)^2 \\ &\geq \kappa(\mathcal{C}), \end{aligned}$$

from which it follows that

$$2|u_i(z)|h_i^* + v_i(z)(h_i^*)^2 \geq -2u_i(z)d_i h_i^* + v_i(z)(h_i^*)^2 \geq \kappa(\mathcal{C}) - \sum_{m=1, m \neq i}^I b_m(E). \quad (16)$$

From (16) we may obtain the inequality

$$\begin{aligned} h_i^*(z, d) &\leq -2 \frac{|u_i(z)|}{v_i(z)} + \frac{1}{h_i^*(z, d)} \frac{(\kappa(\mathcal{C}) - \sum_{l=1, l \neq i}^I b_l(E))}{v_i(z)} \\ &= 2 \frac{|u_i(z)|}{|v_i(z)|} + \frac{1}{h_i^*(z, d)} \frac{|\kappa(\mathcal{C}) - \sum_{l=1, l \neq i}^I b_l(E)|}{|v_i(z)|}, \end{aligned}$$

which implies that

$$h_i^* \leq \max \left\{ 1, 2 \frac{|u_i(z)|}{|v_i(z)|} + \frac{|(\kappa(C) - \sum_{m=1, m \neq i}^I b_m(E))|}{|v_i(z)|} \right\}.$$

Taking the supremum of the righthand side above over all $z \in E$, we find that

$$h_i^*(z, d) \leq \max \left\{ 1, 2 \frac{|u_i^u(E)|}{|v_i^l(E)|} + \frac{|(\kappa(C) - \sum_{m=1, m \neq i}^I b_m(E))|}{|v_i^l(E)|} \right\},$$

for all $z \in E$. Since the quantity on the righthand side above is independent of d , this completes the proof. \square

We also have the following result which is used in the appendix.

Proposition 7.1. *For each $z \in (0, \infty)^I$, $H_d(z)$ is Lipschitz continuous as a function of d with Lipschitz constant ζ_z . Moreover, ζ_z is uniformly bounded over each compact subset of $(0, \infty)^I$.*

Proof. Let $z \in (0, \infty)^I$ and recall that by definition, $H_d(z) = \lim_{t \rightarrow 0} t^{-1}(\Lambda(z + td) - \Lambda(z))$. Also recall by Theorem 3.2 that Λ is locally Lipschitz on $(0, \infty)^I$. Thus, suppose that Λ has Lipschitz constant K_r^z on $B_r(z)$, the open ball of radius r centered at z , where r is sufficiently small such that $B_r(z) \subset (0, \infty)^I$. It then follows that for $d_1, d_2 \in \mathbb{R}^I$,

$$\begin{aligned} \|H_{d_1}(z) - H_{d_2}(z)\| &= \left\| \lim_{t \rightarrow 0} t^{-1}(\Lambda(z + td_1) - \Lambda(z)) - \lim_{t \rightarrow 0} t^{-1}(\Lambda(z + td_2) - \Lambda(z)) \right\| \\ &= \left\| \lim_{t \rightarrow 0} t^{-1}(\Lambda(z + td_1) - \Lambda(z + td_2)) \right\| \\ &= \lim_{t \rightarrow 0} t^{-1} \|\Lambda(z + td_1) - \Lambda(z + td_2)\| \\ &\leq K_r^z \lim_{t \rightarrow 0} t^{-1} \|td_1 - td_2\| \\ &= K_r^z \|d_1 - d_2\|. \end{aligned}$$

Thus, since by Theorem 3.2 Λ is locally Lipschitz on $(0, \infty)^I$, it follows that for fixed $r > 0$, K_R^z is uniformly bounded over E , which completes the proof. \square

8 Appendix: Additional proofs

In the appendix, we provide the proofs of the fluid and diffusion limits stated in Section 4. We relegate these proofs to the appendix as they are of a more standard nature. Although our setup does not appear to allow a direct application of the results found in [27] on Markovian service networks, it is not too far apart either. In Subsection 8.1, we provide the proofs of our fluid limit results, and in Subsection 8.2 we provide the proofs of our diffusion limit results.

8.1 Fluid Limit Proof

In this subsection, we provide the proof of Theorem 4.1. We begin with the following two lemmas. For each $n \geq 1$, $t \geq 0$ and $i \in \{1, \dots, I\}$, let $\bar{E}_i^n(t) = n^{-1}E_i(nt)$, $\hat{N}_i^{s,n}(t) = n^{-1}(N_i^s(nt) - nt)$ and $\hat{N}_i^{a,n}(t) = n^{-1}(N_i^a(nt) - nt)$. Also define the processes $\bar{E}_i^n = \{\bar{E}_i^n(t), t \geq 0\}$, $\hat{N}_i^{s,n} = \{\hat{N}_i^{s,n}(t), t \geq 0\}$ and $\hat{N}_i^{a,n} = \{\hat{N}_i^{a,n}(t), t \geq 0\}$. We then have the following result.

Lemma 8.1. *If $\bar{Z}^n(0) \Rightarrow \bar{Z}(0)$ as $n \rightarrow \infty$ and $n^{-1}\eta_i^n \rightarrow \eta_i$ as $n \rightarrow \infty$ for $i = 1, \dots, I$, then for $i = 1, \dots, I$,*

$$\hat{N}_i^{s,n} \left(\mu_i \int_0^e \Lambda_i(\bar{Z}^n(s)) ds \right) \Rightarrow 0, \quad \text{as } n \rightarrow \infty,$$

and

$$\hat{N}_i^{a,n} \left(\gamma_i \int_0^t \bar{Z}_i^n(s) ds \right) \Rightarrow 0, \quad \text{as } n \rightarrow \infty.$$

Proof. Let $i \in \{1, \dots, I\}$ and note that by (1), we have that for each $t \geq 0$, $\bar{Z}_i^n(t) \leq \bar{Z}_i^n(0) + \bar{E}_i^n(n^{-1}\eta_i^n t)$. Thus, since $\Lambda_i(z) \leq m_i z_i$, it follows that for each $t \geq 0$ and $n \geq 1$,

$$\int_0^t \Lambda_i(\bar{Z}^n(s)) ds \leq t m_i (\bar{Z}^n(0) + \bar{E}_i^n(n^{-1}(\eta_i^n t))). \quad (17)$$

Now, for each $i \in \{1, \dots, I\}$ and $t \geq 0$, by the functional weak law of large numbers and the assumption that $n^{-1}\eta_i^n \rightarrow \eta_i$ as $n \rightarrow \infty$, it follows that $\bar{E}_i^n(n^{-1}\eta_i^n t) \Rightarrow \eta_i t$ as $n \rightarrow \infty$. From (17), this then implies that for each $t \geq 0$ the sequence

$$\left\{ \int_0^t \Lambda_i(\bar{Z}^n(s)) ds, n \geq 1 \right\} \quad (18)$$

is tight. However, by the functional weak law of large numbers, $\hat{N}_i^{s,n} \Rightarrow 0$ as $n \rightarrow \infty$, and so it follows by the fact that (18) that

$$\hat{N}_i^{s,n} \left(\mu_i \int_0^e \Lambda_i(\bar{Z}^n(s)) ds \right) \Rightarrow 0,$$

as $n \rightarrow \infty$. Similar reasoning shows that

$$\hat{N}_i^{a,n} \left(\gamma_i \int_0^t \bar{Z}_i^n(s) ds \right) \Rightarrow 0,$$

as $n \rightarrow \infty$, which completes the proof. \square

Our next result shows that the fluid scaled process \bar{Z}^n may be bounded away from the origin when n is large.

Lemma 8.2. *If $\bar{Z}^n(0) \Rightarrow \bar{Z}(0) \in (0, \infty)^I$ as $n \rightarrow \infty$ and $n^{-1}\eta_i^n \rightarrow \eta_i$ as $n \rightarrow \infty$ for $i = 1, \dots, I$, then for each $T \geq 0$ and $\varepsilon > 0$, there exists constants $K_\varepsilon^T > 0$ and $n_\varepsilon^T \geq 1$ such that for sufficiently large $n \geq n_\varepsilon^T$,*

$$P \left(\inf_{0 \leq t \leq T} \min_{i=1, \dots, I} |\bar{Z}_i^n(t)| > K_\varepsilon^T \right) > 1 - \varepsilon.$$

Proof. Let $i \in \{1, \dots, I\}$, $n \geq 1$ and $t \geq 0$. By equation (1), the definition of $\bar{Z}^n(t)$, $\hat{N}_i^{s,n}(t)$ and $\hat{N}_i^{a,n}(t)$, and the fact that $n^{-1}\Lambda_i^n(Z^n(t)) = \Lambda_i(\bar{Z}^n(t))$, it follows that

$$\bar{Z}_i^n(t) = \bar{Z}_i^n(0) + \bar{E}_i^n(n^{-1}\eta_i^n t) - n^{-1}N_i^s \left(\mu_i \int_0^t \Lambda_i(\bar{Z}^n(s)) ds \right) - n^{-1}N_i^a \left(\gamma_i \int_0^t \bar{Z}_i^n(s) ds \right) \quad (19)$$

Now consider the solution $\bar{Y}_i^n = \{\bar{Y}_i^n(t), t \geq 0\}$ to the equation

$$\begin{aligned} \bar{Y}_i^n(t) &= \bar{Z}_i^n(0) + \bar{E}_i^n(n^{-1}\eta_i^n t) - \hat{N}^{s,n} \left(\mu_i \int_0^t \Lambda_i(\bar{Z}^n(s)) ds \right) - \hat{N}^{a,n} \left(\gamma_i \int_0^t \bar{Z}_i^n(s) ds \right) \\ &\quad - (\gamma_i + \mu_i m_i) \int_0^t \bar{Y}_i^n(s) ds, \end{aligned} \quad (20)$$

for $t \geq 0$. Since the function $g_i(x) = (\gamma_i + \mu_i m_i)x$ is Lipschitz continuous, by Lemma 1 of [32] one express the unique solution to this equation by writing

$$\begin{aligned} &\bar{Y}_i^n \\ &= \Psi_{\gamma_i + \mu_i m_i} \left(\bar{Z}_i^n(0) + \bar{E}_i^n(n^{-1}\eta_i^n e) - \hat{N}^{s,n} \left(\mu_i \int_0^e \Lambda_i(\bar{Z}^n(s)) ds \right) - \hat{N}^{a,n} \left(\gamma_i \int_0^e \bar{Z}_i^n(s) ds \right) \right), \end{aligned} \quad (21)$$

where for each $a \in \mathbb{R}$, $\Psi_a : D([0, \infty), \mathbb{R}) \mapsto D([0, \infty), \mathbb{R})$ is a Lipschitz continuous map such that for each $x \in D([0, \infty), \mathbb{R})$, $\Psi_a(x)$ is the unique solution to integral equation

$$z(t) = x(t) - a \int_0^t z(s) ds, \quad (22)$$

for $t \geq 0$. By the assumptions of the lemma and Lemma 8.1,

$$\bar{Z}_i^n(0) + \bar{E}_i^n(n^{-1}\eta_i^n) - \hat{N}^{s,n} \left(\mu_i \int_0^e \Lambda_i(\bar{Z}^n(s)) ds \right) - \hat{N}^{a,n} \left(\gamma_i \int_0^e \bar{Z}_i^n(s) ds \right) \Rightarrow \bar{Z}_i(0) + \eta_i e,$$

as $n \rightarrow \infty$. Thus, by the continuous mapping Theorem and (21),

$$\bar{Y}_i^n \Rightarrow \bar{Y}_i = \Psi_{\gamma_i + \mu_i m_i}(\bar{Z}(0) + \eta_i e),$$

as $n \rightarrow \infty$. However, by (22), we may explicitly write that

$$\bar{Y}_i(t) = \frac{\eta_i}{\gamma_i + \mu_i m_i} + \left(\bar{Z}(0) - \frac{\eta_i}{\gamma_i + \mu_i m_i} \right) e^{-(\gamma_i + \mu_i m_i)t} > 0,$$

for $t \geq 0$, where the inequality follows from the fact that $\bar{Z}(0) \in (0, \infty)^I$. It is then straightforward to show that this then implies that for each $T \geq 0$ and $\varepsilon > 0$, there exists a $K_\varepsilon^T > 0$ such that for sufficiently large $n \geq n_\varepsilon^T$

$$P \left(\inf_{0 \leq t \leq T} \min_{i=1, \dots, I} |\bar{Y}_i^n(t)| > K_\varepsilon^T \right) > 1 - \varepsilon. \quad (23)$$

We now show that $\bar{Z}_i^n(t) \geq \bar{Y}_i^n(t)$ for each $n \geq 0$ and $t \geq 0$ and $i = 1, \dots, I$, which, by virtue of (23), completes the proof. First note that since $\Lambda_i(z) \leq m_i z_i$, it follows from (19) that

$$\begin{aligned} \bar{Z}_i^n(t) &\geq \bar{Z}^n(0) + \bar{E}_i^n(n^{-1}\eta_i^n t) - \hat{N}^{s,n} \left(\mu_i \int_0^t \Lambda_i(\bar{Z}^n(s)) ds \right) - \hat{N}^{a,n} \left(\gamma_i \int_0^t \bar{Z}_i^n(s) ds \right) \\ &\quad - (\gamma_i + \mu_i m_i) \int_0^t \bar{Z}_i^n(s) ds. \end{aligned} \quad (24)$$

Subtracting (20) from (24), one obtains that

$$(\bar{Z}_i^n(t) - \bar{Y}_i^n(t)) \geq -(\gamma_i + \mu_i m_i) \int_0^t (\bar{Z}_i^n(s) - \bar{Y}_i^n(s)) ds.$$

Thus, by Gronwall's inequality, it follows that

$$(\bar{Z}_i^n(t) - \bar{Y}_i^n(t)) \geq 0 \cdot \exp(-(\gamma_i + \mu_i m_i)t) = 0,$$

which completes the proof. \square

We now present the proof of Theorem 4.1.

Proof of Theorem 4.1. For each $\varepsilon > 0$, let C_ε be the compact subset of \mathbb{R}^I defined by $C_\varepsilon = \{z \in \mathbb{R}^I : \varepsilon < z_i < \varepsilon^{-1}, i = 1, \dots, I\}$. Next, let $\Lambda^\varepsilon : \mathbb{R}^I \mapsto \mathbb{R}^I$ be such that $\Lambda^\varepsilon(z) = \Lambda(z)$ for $z \in C_\varepsilon$ and extend Λ^ε to the remainder of \mathbb{R}^I in order to make Λ^ε Lipschitz on all of \mathbb{R}^I . Note that by standard Lipschitz extension theorems, such an extension is always possible since, by Theorem 3.2, Λ is Lipschitz on C_ε .

Now, for each $n \geq 1$, define the process $\bar{Z}^{n,\varepsilon} = \{(\bar{Z}_1^{n,\varepsilon}(t), \dots, \bar{Z}_I^{n,\varepsilon}(t)), t \geq 0\}$ to be the unique solution to

$$\begin{aligned} \bar{Z}_i^{n,\varepsilon}(t) &= \bar{Z}_i^n(0) + \bar{E}_i^n(n^{-1}\eta_i^n t) - \hat{N}^{s,n} \left(\mu_i \int_0^t \Lambda_i(\bar{Z}^n(s)) ds \right) - \hat{N}^{a,n} \left(\gamma_i \int_0^t \bar{Z}_i^n(s) ds \right) \\ &\quad - \mu_i m_i \int_0^t \Lambda_i^\varepsilon(\bar{Z}^{n,\varepsilon}(s)) ds - \gamma_i \int_0^t \bar{Z}_i^{n,\varepsilon}(s) ds, \end{aligned}$$

for $i = 1, \dots, I$. By Lemma 1 of Reed and Ward [32], it follows since Λ^ε is Lipschitz continuous that for each $x \in D([0, \infty), \mathbb{R}^I)$, there exists a unique $z \in D([0, \infty), \mathbb{R}^I)$ such that

$$z_i(t) = x_i(t) - \mu_i \int_0^t \Lambda_i^\varepsilon(z(s)) ds - \gamma_i \int_0^t z_i(s) ds, \quad (25)$$

for $t \geq 0$ and $i = 1, \dots, I$. Moreover, the map $\Psi^\varepsilon : D([0, \infty), \mathbb{R}^I) \mapsto D([0, \infty), \mathbb{R}^I)$ such that $z = \Psi^\varepsilon(x)$ is Lipschitz continuous. Hence, we may write

$$\bar{Z}^{n,\varepsilon} = \Psi^\varepsilon \left(\bar{Z}^n(0) + \bar{E}^n(n^{-1}\eta_i^n e) - \hat{N}^{s,n} \left(\mu_i \int_0^e \Lambda_i(\bar{Z}^n(s)) ds \right) - \hat{N}^{a,n} \left(\gamma_i \int_0^e \bar{Z}_i^n(s) ds \right) \right).$$

However, since by Lemma 8.1, the assumptions of the theorem and the functional weak law of large numbers

$$\bar{Z}^n(0) + \bar{E}^n(n^{-1}\eta_i^n) - \hat{N}^{s,n} \left(\mu_i \int_0^e \Lambda_i(\bar{Z}^n(s)) ds \right) - \hat{N}^{a,n} \left(\gamma_i \int_0^e \bar{Z}_i^n(s) ds \right) \Rightarrow \bar{Z}(0) + \bar{E}(\eta e),$$

as $n \rightarrow \infty$, it follows by the continuous mapping theorem that $\bar{Z}^{n,\varepsilon} \Rightarrow \bar{Z}^\varepsilon = \Psi^\varepsilon(\bar{Z}(0) + \bar{E}(\eta e))$ as $n \rightarrow \infty$.

Now note that by (19), we have that for each $i = 1, \dots, I$, and $T \geq 0$,

$$\sup_{0 \leq t \leq T} |\bar{Z}_i^n(t)| \leq \bar{Z}_i^n(0) + \bar{E}_i^n(n^{-1}\eta_i^n T).$$

However, since $n^{-1}\eta_i^n \rightarrow \eta_i$ and $\bar{Z}_i^n(0) \Rightarrow \bar{Z}_i(0)$, it follows that the sequence

$$\{\bar{Z}_i^n(0) + \bar{E}_i^n(n^{-1}\eta_i^n T), n \geq 1\}$$

is stochastically bounded. Hence, $\{\sup_{0 \leq t \leq T} |\bar{Z}_i^n(t)|, n \geq 1\}$ is stochastically bounded as well.

We now have that since the map $\bar{\Psi}^\varepsilon$ is unique, it follows by Lemma 8.2 and the fact that $\{\sup_{0 \leq t \leq T} |\bar{Z}_i^n(t)|, n \geq 1\}$ is stochastically bounded that for each $T \geq 0$,

$$\lim_{\varepsilon \rightarrow 0} \limsup_{n \rightarrow \infty} P \left(\sup_{0 \leq t \leq T} \|\bar{Z}^n(t) - \bar{Z}_\varepsilon^n(t)\| > 0 \right) = 0. \quad (26)$$

Since for each $\varepsilon > 0$, $\bar{Z}_\varepsilon^n \Rightarrow \bar{Z}_\varepsilon$ as $n \rightarrow \infty$, it follows by (26) that $\{\bar{Z}^n, n \geq 1\}$ is tight and hence, relatively compact. Consider an arbitrary subsequence along which \bar{Z}^n converges weakly to some limit \bar{Z} . By Lemma 8.2, it follows that for each $T \geq 0$,

$$P \left(\inf_{0 \leq t \leq T} \min_{i=1, \dots, I} |\bar{Z}_i(t)| > 0 \right) = 1. \quad (27)$$

Hence, by (26) and the definition of Λ_ε , it must be that \bar{Z} satisfies (5) almost surely. Since our chosen subsequence was arbitrary, it remains to show that \bar{Z} is the unique, strong solution to (5) and that (5) uniquely characterizes \bar{Z} in law order to complete the proof.

Suppose that \bar{Z} satisfies (5) almost surely. Clearly, for each $T \geq 0$,

$$\sup_{0 \leq t \leq T} \|\bar{Z}(t)\| \leq \|\bar{Z}(0)\| + T \sup_{i=1, \dots, I} \eta_i.$$

Moreover, by (27), $\inf_{0 \leq t \leq T} \min_{i=1, \dots, I} |\bar{Z}_i(t)| > 0$. Hence, for each ω , \bar{Z} satisfies (5) with Λ replaced by Λ_ε for ε sufficiently small where the choice of ε may depend on ω . Since Λ_ε is Lipschitz continuous on compact subsets of $(0, \infty)^I$, it follows by Lemma 1 of [32] that \bar{Z} is the unique, strong solution to (5). It is now straightforward to see that (5) uniquely characterizes \bar{Z} in law since for each $\delta > 0$, we have that $P(\bar{Z} = \Psi^\varepsilon(\bar{Z}(0) + \eta e)) > 1 - \delta$ for ε chosen sufficiently small, where Ψ^ε is a continuous map. □

8.2 Diffusion Limit Proof

In this section, we provide the proof of Theorem 4.2. In order to begin, for each $i = 1, \dots, I$, $t \geq 0$ and $n \geq 1$, let $\check{E}_i^n(t) = n^{-1/2}(E_i(nt) - nt)$, $\check{N}_i^{s,n}(t) = n^{-1/2}(N_i^s(nt) - nt)$ and $\check{N}_i^{a,n}(t) = n^{-1/2}(N_i^a(nt) - nt)$. Also, define the processes $\check{E}_i^n = \{\check{E}_i^n(t), t \geq 0\}$ for $i = 1, \dots, I$. Now let

$$\tilde{E}_i^n(t) = \check{E}_i^n(n^{-1}\eta_i^n t) \tag{28}$$

$$\tilde{S}_i^n(t) = \check{N}_i^{s,n} \left(\mu_i \int_0^t \Lambda_i(\bar{Z}^n(s)) ds \right), \tag{29}$$

$$\tilde{C}_i^n(t) = \check{N}_i^{a,n} \left(\gamma_i \int_0^t \bar{Z}_i^n(s) ds \right), \tag{30}$$

and define the \mathbb{R}^I -valued processes $\tilde{E}^n = ((\tilde{E}_1^n(t), \dots, \tilde{E}_I^n(t)), t \geq 0)$, $\tilde{S}^n = ((\tilde{S}_1^n(t), \dots, \tilde{S}_I^n(t)), t \geq 0)$ and $\tilde{C}^n = ((\tilde{C}_1^n(t), \dots, \tilde{C}_I^n(t)), t \geq 0)$.

Next, recall the definitions of $\tilde{Z}^n(t) = n^{1/2}(\bar{Z}^n(t) - \bar{Z}(t))$ and $\tilde{Z}^n = \{\tilde{Z}^n(t), t \geq 0\}$ from Section 4. By (1), (5), and (28)-(30), we may write that for each $t \geq 0$, $n \geq 1$ and $i \in \{1, \dots, I\}$,

$$\begin{aligned} \tilde{Z}_i^n(t) &= \tilde{Z}_i^n(0) + \tilde{E}_i^n(t) - \tilde{S}_i^n(t) - \tilde{C}_i^n(t) + \sqrt{n}(n^{-1}\eta_i^n - \eta_i)t \\ &\quad - \gamma_i \int_0^t \tilde{Z}_i^n(s) ds - \mu_i \int_0^t n^{1/2}(\Lambda_i(\bar{Z}^n(s)) - \Lambda_i(\bar{Z}(s))) ds. \end{aligned} \tag{31}$$

Now, for each $i = 1, \dots, I$, let $\tilde{E}_i^n = \{\tilde{E}_i^n(t), t \geq 0\}$, and let $\tilde{E}^n = (\tilde{E}_1^n, \dots, \tilde{E}_I^n)$,

$$\tilde{S}^n = \left(\tilde{N}_1^{s,n} \left(r_1 \int_0^e \Lambda_1(\bar{Z}^n(s)) ds \right), \dots, \tilde{N}_I^{s,n} \left(r_I \int_0^e \Lambda_I(\bar{Z}^n(s)) ds \right) \right)$$

and

$$\tilde{C}^n = \left(\tilde{N}_1^{a,n} \left(\gamma_1 \int_0^e \bar{Z}_1^n(s) ds \right), \dots, \tilde{N}_I^{a,n} \left(\gamma_I \int_0^e \bar{Z}_I^n(s) ds \right) \right).$$

The following is our first result. Recall that in our diffusion limit results, we are assuming that $\bar{Z}(0)$ is a constant.

Proposition 8.3. *If $\bar{Z}^n(0) \Rightarrow \bar{Z}(0)$ as $n \rightarrow \infty$, then $(\tilde{E}^n, \tilde{S}^n, \tilde{C}^n) \Rightarrow (\tilde{E}, \tilde{S}, \tilde{C})$ as $n \rightarrow \infty$, where $\tilde{E} = (\tilde{E}_1(\eta_1 e), \dots, \tilde{E}_I(\eta_I e))$ is such that \tilde{E}_i is a standard Brownian motion for each $i = 1, \dots, I$, and $\tilde{S} = (\tilde{S}_1, \dots, \tilde{S}_I)$ is such that \tilde{S}_i is a Brownian motion with infinitesimal variance at time t given by*

$$(\sigma_i^s)^2(t) = \mu_i \Lambda_i(\bar{Z}(t))$$

and $\tilde{C} = (\tilde{C}_1, \dots, \tilde{C}_I)$ is such that \tilde{C}_i is a Brownian motion with infinitesimal variance at time t given by

$$(\sigma_i^a)^2(t) = \gamma_i \bar{Z}_i(t).$$

Moreover, all of the Brownian motions in the above are independent of one another.

Proof of Proposition 8.3. By Theorem 4.1, $\bar{Z}^n \Rightarrow \bar{Z}$ as $n \rightarrow \infty$. Hence, for each $i = 1, \dots, I$,

$$\int_0^e \bar{Z}_i^n(s) ds \Rightarrow \int_0^e \bar{Z}_i(s) ds, \quad (32)$$

as $n \rightarrow \infty$. Moreover, in a similar fashion to as in the proof of Lemma 8.2, using the boundedness of \bar{Z} away from zero, the local Lipschitz continuity of Λ and the continuous mapping theorem [4], it may be shown that for each $i = 1, \dots, I$, $\Lambda_i(\bar{Z}^n(e)) \Rightarrow \Lambda_i(\bar{Z}(e))$ as $n \rightarrow \infty$, and hence

$$\int_0^e \Lambda_i(\bar{Z}^n(s)) ds \Rightarrow \int_0^e \Lambda_i(\bar{Z}(s)) ds, \quad (33)$$

as $n \rightarrow \infty$.

Now recall that by the functional central limit theorem [4], $(\tilde{E}^n, \tilde{N}^{s,n}, \tilde{N}^{a,n}) \Rightarrow (\tilde{E}, \tilde{N}^s, \tilde{N}^a)$ as $n \rightarrow \infty$, where \tilde{E}, \tilde{N}^s and \tilde{N}^a , are standard I -dimensional Brownian motions, independent of one another. The result now follows by that fact that $n^{-1}\eta_i^n \rightarrow \eta_i$ as $n \rightarrow \infty$, (32), (33), the random time change theorem [4] and the definitions of \tilde{E}^n, \tilde{S}^n and \tilde{C}^n . \square

In order to prove Theorem 4.2, we first must show that the family of processes $(\tilde{Z}^n, n \geq 1)$ is tight.

Proposition 8.4. *If $\bar{Z}^n(0) \Rightarrow \bar{Z}(0)$ as $n \rightarrow \infty$ and $\sqrt{n}(n^{-1}\eta_i^n - \eta_i) \rightarrow \beta_i$ for $i = 1, \dots, I$, then $\{\tilde{Z}^n, n \geq 1\}$ is tight.*

Proof. We must verify that $\{\tilde{Z}^n, n \geq 1\}$ satisfies conditions (i) and (ii) of Theorem xxx of [4]. We begin with condition (i). Let $T \geq 0$. Since $\bar{Z}^n \Rightarrow \bar{Z}$ as $n \rightarrow \infty$ by Theorem 4.1, it follows by Prohorov's theorem [4], that for each $\varepsilon > 0$, there exists a C_ε^T such that

$$P\left(\sup_{0 \leq t \leq T} \|\bar{Z}^n(t)\| \leq C_\varepsilon^T\right) > 1 - \varepsilon,$$

for all $n \geq 1$. Thus, by Lemma 8.2, there exists a second constant such K_ε^T such that for all $n \geq n_\varepsilon^T$,

$$P\left(\sup_{0 \leq t \leq T} \|\bar{Z}^n(t)\| < C_\varepsilon^T \text{ and } \inf_{0 \leq t \leq T} \|\bar{Z}^n(t)\| > K_\varepsilon^T\right) > 1 - \varepsilon. \quad (34)$$

Since by Theorem 3.2, Λ is locally Lipschitz on $(0, \infty)^I$, it follows that Λ is Lipschitz continuous on each compact subset of $(0, \infty)^I$. Let $\kappa_{r, \varepsilon}$ be the Lipschitz constant for Λ on the closed ball of radius r centered at the origin intersected with the set $z_\varepsilon = \{z \in \mathbb{R}_+^I : z_i > \varepsilon, i = 1, \dots, I\}$. It then follows by (31) and (34) that for $n \geq n_\varepsilon^T$, with probability at least $1 - \varepsilon$,

$$\begin{aligned} & |\tilde{Z}_i^n(t)| \quad (35) \\ & \leq \left| \tilde{Z}_i^n(0) + \tilde{E}_i^n(t) - \tilde{S}_i^n(t) - \tilde{C}_i^n(t) + \sqrt{n}(\eta_i^n - \eta_i)t \right| + m_i \int_0^t \left| n^{1/2}(\Lambda_i(\bar{Z}^n(s)) - \Lambda_i(\bar{Z}(s))) \right| ds \\ & \quad + \gamma_i \int_0^t \tilde{Z}_i^n(s) ds \\ & \leq \left| \tilde{Z}_i^n(0) + \tilde{E}_i^n(t) - \tilde{S}_i^n(t) - \tilde{C}_i^n(t) + \sqrt{n}(\eta_i^n - \eta_i)t \right| + (m_i \kappa_{C_\varepsilon^T, K_\varepsilon^T} + \gamma_i) \int_0^t \sum_{i=1}^I |\tilde{Z}_i^n(s)| ds. \end{aligned}$$

Thus, for $n \geq n_\varepsilon^T$, with probability at least $1 - \varepsilon$

$$\begin{aligned} & \sup_{0 \leq t \leq T} \sum_{i=1}^I |\tilde{Z}_i^n(t)| \\ & \leq \sup_{0 \leq t \leq T} \sum_{i=1}^I \left| \tilde{Z}_i^n(0) + \tilde{E}_i^n(t) - \tilde{S}_i^n(t) - \tilde{C}_i^n(t) + \sqrt{n}(\eta_i^n - \eta_i)t \right| \\ & \quad + (m_i \kappa_{C_\varepsilon^T, K_\varepsilon^T} + \gamma_i) \int_0^T \sup_{0 \leq s \leq t} \sum_{i=1}^I |\tilde{Z}_i^n(s)| dt, \end{aligned}$$

and hence, by Gronwall's inequality, we obtain that for $n \geq n_\varepsilon^T$, with probability at least $1 - \varepsilon$,

$$\begin{aligned} & \sup_{0 \leq t \leq T} \sum_{i=1}^I |\tilde{Z}_i^n(t)| \\ & \leq \sup_{0 \leq t \leq T} \sum_{i=1}^I \left| \tilde{Z}_i^n(0) + \tilde{E}_i^n(t) - \tilde{S}_i^n(t) - \tilde{C}_i^n(t) + \sqrt{n}(\eta_i^n - \eta_i)t \right| e^{MT \sum_{i=1}^I (\mu_i + \gamma_i)}. \end{aligned} \quad (36)$$

Now since

$$\left\{ \sup_{0 \leq t \leq T} \sum_{i=1}^I \left| \tilde{Z}_i^n(0) + \tilde{E}_i^n(n^{-1}\eta_i^n t) - \tilde{S}_i^n(t) - \tilde{C}_i^n(t) + \sqrt{n}(n^{-1}\eta_i^n - \eta_i)t \right|, n \geq 1 \right\}$$

is stochastically bounded by Proposition 8.3, it follows by (36) that

$$\left\{ \sup_{0 \leq t \leq T} \sum_{i=1}^I |\tilde{Z}_i^n(t)|, n \geq 1 \right\}$$

is stochastically bounded as well, which verifies that condition (i) is satisfied.

We now show that condition (ii) is satisfied. Let $\delta, \varepsilon > 0$ and then note that similar to (35), by (31) and (34) it follows that that for $n \geq n_\varepsilon^T$, with probability at least $1 - \varepsilon$,

$$\begin{aligned} & |\tilde{Z}_i^n(t + \delta) - \tilde{Z}_i^n(t)| \\ & \leq \left| \left(\tilde{Z}_i^n(0) + \tilde{E}_i^n(t + \delta) - \tilde{S}_i^n(t + \delta) - \tilde{C}_i^n(t + \delta) + \sqrt{n}(\eta_i^n - \eta_i)(t + \delta) \right) \right. \\ & \quad \left. - \left(\tilde{Z}_i^n(0) + \tilde{E}_i^n(t) - \tilde{S}_i^n(t) - \tilde{C}_i^n(t) + \sqrt{n}(\eta_i^n - \eta_i)t \right) \right| \\ & \quad + m_i \int_t^{t+\delta} \left| n^{1/2}(\Lambda_i(\bar{Z}^n(s)) - \Lambda_i(\bar{Z}(s))) \right| ds + \gamma_i \int_t^{t+\delta} \tilde{Z}_i^n(s) ds \\ & \leq \left| \left(\tilde{Z}_i^n(0) + \tilde{E}_i^n(t + \delta) - \tilde{S}_i^n(t + \delta) - \tilde{C}_i^n(t + \delta) + \sqrt{n}(\eta_i^n - \eta_i)(t + \delta) \right) \right. \\ & \quad \left. - \left(\tilde{Z}_i^n(0) + \tilde{E}_i^n(t) - \tilde{S}_i^n(t) - \tilde{C}_i^n(t) + \sqrt{n}(\eta_i^n - \eta_i)t \right) \right| \\ & \quad + (m_i \kappa_{C_\varepsilon^T, K_\varepsilon^T} + \gamma_i) \int_t^{t+\delta} \sum_{i=1}^I |\tilde{Z}_i^n(s)| ds. \end{aligned}$$

However, since by Proposition 8.3 and the assumptions of the proposition, the sequence

$$\left\{ \tilde{Z}_i^n(0) + \tilde{E}_i^n(e) - \tilde{S}_i^n(t) - \tilde{C}_i^n(t) + \sqrt{n}(\eta_i^n - \eta_i)e, n \geq 1 \right\}$$

is tight and since as was just proven, the sequence

$$\left\{ \sup_{0 \leq t \leq T} \sum_{i=1}^I |\tilde{Z}_i^n(t)|, n \geq 1 \right\}$$

is stochastically bounded, it follows that condition (ii) is satisfied. \square

Now for each $n \geq 1$ and $i = 1, \dots, I$, let

$$\tilde{\varepsilon}_i^n(t) = \mu_i \int_0^t n^{1/2} (\Lambda_i(\bar{Z}^n(s)) - \Lambda_i(\bar{Z}(s))) ds - \mu_i \int_0^t (H_{\bar{Z}^n(s)}^n)_i(\bar{Z}(s)) ds,$$

for $t \geq 0$ and set $\tilde{\varepsilon}^n = \{(\tilde{\varepsilon}_1^n(t), \dots, \tilde{\varepsilon}_I^n(t)), t \geq 0\}$. We next have the following result.

Proposition 8.5. *If $\tilde{Z}^n(0) \Rightarrow \tilde{Z}(0)$ as $n \rightarrow \infty$, then $\tilde{\varepsilon}^n \Rightarrow 0$ as $n \rightarrow \infty$.*

Proof. For each $n \geq 1$ and $i = 1, \dots, I$, let $\tilde{\varepsilon}_i^n = \{\tilde{\varepsilon}_i^n(t), t \geq 0\}$. It suffices to show that $\tilde{\varepsilon}_i^n \Rightarrow 0$ as $n \rightarrow \infty$ for $i = 1, \dots, I$. Let $\varepsilon > 0$ and recall the definition Λ^ε from the proof of Theorem 4.1 to be such that $\Lambda^\varepsilon(z) = \Lambda(z)$ for $z \in C_\varepsilon$, where C_ε be the compact subset of \mathbb{R}^I defined by $C_\varepsilon = \{z \in \mathbb{R}^I : \varepsilon < z_i < \varepsilon^{-1}, i = 1, \dots, I\}$, and Λ^ε is arbitrarily extended as a Lipschitz continuous function to the remainder of \mathbb{R}^I .

Now, for $T \geq 0$, let $\varepsilon > 0$ be such that $\bar{Z}(t) > \varepsilon$ for $0 \leq t \leq T$ and define

$$\tilde{\varepsilon}_i^{n,\varepsilon}(t) = \mu_i \int_0^t n^{1/2} (\Lambda_i^\varepsilon(\bar{Z}^n(s)) - \Lambda_i^\varepsilon(\bar{Z}(s))) ds - \mu_i \int_0^t (H_{\bar{Z}^n(s)}^n)_i(\bar{Z}(s)) ds,$$

for $0 \leq t \leq T$. Note that by (34) it follows that for each $\delta > 0$, there exists an n_δ^T such that for all $n \geq n_\delta^T$, $\sup_{0 \leq t \leq T} |\tilde{\varepsilon}_i^{n,\varepsilon}(t) - \tilde{\varepsilon}_i^n(t)| = 0$ with probability at least $1 - \delta$. Hence, it suffices to show that

$$\sup_{0 \leq t \leq T} |\tilde{\varepsilon}_i^{n,\varepsilon}(t)| \Rightarrow 0,$$

as $n \rightarrow \infty$, in order to complete the proof.

First, recall that by Proposition 8.4 the sequence $\{\tilde{Z}^n, n \geq 1\}$ is tight. Thus, by Prohorov's theorem, for every sequence there exists a subsequence $\{n_k\}$ such that $\tilde{Z}^{n_k} \Rightarrow \tilde{Y}$ as $n_k \rightarrow \infty$. Moreover, by the Skorohod representation theorem, there exists an alternative probability space (Ω, \mathcal{F}, P) on which are defined a sequence of random elements $\{\hat{Z}^{n_k}, n_k \geq 1\}$ such that \hat{Z}^{n_k} is equal in law to \tilde{Z}^{n_k} and, moreover, such that $\hat{Z}^{n_k} \rightarrow \hat{Y}$ almost surely.

Now for each $n \geq 1$, define the map $\Upsilon_i^{n,\varepsilon} : D([0, \infty), \mathbb{R}^I) \mapsto D([0, \infty), \mathbb{R})$ by setting

$$\Upsilon_i^{n,\varepsilon}(x) = \mu_i \int_0^\varepsilon (1/n^{-1/2}) (\Lambda_i^\varepsilon(\bar{Z}(s) + n^{-1/2}x(s)) - \Lambda_i^\varepsilon(\bar{Z}(s))) ds - \mu_i \int_0^\varepsilon H_i^{x(s)}(\bar{Z}(s)) ds.$$

Note that by Proposition 7.1 and Theorem 3.2, the map $\Upsilon_i^{n,\varepsilon}$ is continuous with respect to the uniform topology. Moreover, $\Upsilon_i^{n,\varepsilon}(\tilde{Z}^n) = \tilde{\varepsilon}_i^{n,\varepsilon}$. Hence, $\Upsilon_i^{n_k,\varepsilon}(\hat{Z}^{n_k})$ is equal in law to $\tilde{\varepsilon}_i^{n_k,\varepsilon}$ and so it suffices to show that

$$\sup_{0 \leq t \leq T} |\Upsilon_i^{n_k,\varepsilon}(\hat{Z}^{n_k})(t)| \rightarrow 0,$$

almost surely as $n_k \rightarrow \infty$.

Note that we may write,

$$\Upsilon_i^{n,\varepsilon}(\hat{Z}^n) = \mu_i \int_0^e ((1/n^{-1/2})(\Lambda_i^\varepsilon(\bar{Z}(s) + n^{-1/2}\hat{Z}^n(s)) - \Lambda_i^\varepsilon(\bar{Z}(s))) - (H_{\hat{Z}^n(s)})_i(\bar{Z}(s))) ds.$$

Also, recall that since $\hat{Z}^{n_k} \rightarrow \hat{Y}$, it follows that for each $T \geq 0$, $\sup_{n_k \geq 1} \sup_{0 \leq t \leq T} \|\hat{Z}^{n_k}(t)\| < \infty$. Thus, since for each $T \geq 0$ there exists a compact subset $E_T \subset (0, \infty)^I$ such that $\bar{Z}(t) \in E_T$ for all $0 \leq t \leq T$, it follows by Proposition 7.1 and Theorem 3.2 that

$$\begin{aligned} & \sup_{0 \leq t \leq T} \left| \left((1/n^{-1/2})(\Lambda_i^\varepsilon(\bar{Z}(t) + n^{-1/2}\hat{Z}^n(t)) - \Lambda_i^\varepsilon(\bar{Z}(t))) - (H_{\hat{Z}^n(t)})_i(\bar{Z}(t))) \right) \right| \\ & \leq \sup_{0 \leq t \leq T} \left| (1/n^{-1/2})(\Lambda_i^\varepsilon(\bar{Z}(t) + n^{-1/2}\hat{Z}^n(t)) - \Lambda_i^\varepsilon(\bar{Z}(t))) \right| + \sup_{0 \leq t \leq T} \left| (H_{\hat{Z}^n(t)})_i(\bar{Z}(t)) \right| \\ & \leq K_E \sup_{0 \leq t \leq T} |\hat{Z}^n(t)| \\ & < \infty. \end{aligned}$$

Thus, by the bounded convergence theorem, it suffices to show that

$$((1/n^{-1/2})(\Lambda_i^\varepsilon(\bar{Z}(t) + n^{-1/2}\hat{Z}^n(t)) - \Lambda_i^\varepsilon(\bar{Z}(t))) - (H_{\hat{Z}^n(t)})_i(\bar{Z}(t))) \rightarrow 0,$$

as $n \rightarrow \infty$. However, this is immediate by the definition of $H_d(z)$, Proposition 7.1 and the fact that $\hat{Z}^{n_k} \rightarrow \hat{Y}$ as $n_k \rightarrow \infty$. \square

We are now in a position to provide the proof of Theorem 4.2.

Proof of Theorem 4.2. By (31) and (37) we have that

$$\begin{aligned} \tilde{Z}_i^n(t) &= \tilde{Z}_i^n(0) + \tilde{E}_i^n(t) - \tilde{S}_i^n(t) - \tilde{C}_i^n(t) - \tilde{\varepsilon}_i^n(t) + \sqrt{n}(n^{-1}\eta_i^n - \eta_i)t \\ &\quad - \mu_i \int_0^t (H_{\tilde{Z}^n(s)})_i(\bar{Z}(s)) ds, \end{aligned} \tag{37}$$

for $t \geq 0$ and $i = 1, \dots, I$. Recall that by Proposition 7.1, for each $z \in \mathbb{R}_+^I$, $H_d(z)$ is Lipschitz continuous as a function of d . Moreover, the Lipschitz constant is uniformly bounded for all z in a compact subset of $(0, \infty)^I$. Hence, since by Lemma 8.2, $\inf_{0 \leq t \leq T} \bar{Z}(t) > 0$, it follows

that the Lipschitz constant of $H_d(\bar{Z}(t))$ is uniformly bounded over the interval $[0, T]$ for each $T \geq 0$. By (37) and a straightforward modification to Lemma 1 of [32] to allow for time varying drift functions, we may then write $\tilde{Z}^n = \mathcal{M}(\tilde{Z}^n(0) + \tilde{E}^n - \tilde{S}^n - \tilde{C}^n + \sqrt{n}(n^{-1}\eta^n - \eta)e - \tilde{\varepsilon}^n)$, where $\mathcal{M} : D([0, \infty), \mathbb{R}^I) \mapsto D([0, \infty), \mathbb{R}^I)$ is a continuous function. The result now follows by the hypothesis of the theorem, Propositions 8.3 and 8.5 and the Continuous Mapping Theorem [4]. \square

Acknowledgments

The second author of this paper is also affiliated with VU University Amsterdam, Eurandom and Georgia Tech, and is sponsored by an NWO-VIDI grant and an IBM faculty award. Part of this research has been carried out at the Newton Institute in Cambridge (UK). Both authors are grateful to Alex Shapiro for pointing them towards key details in his book.

References

- [1] Ayesta, U., Mandjes, M. (2009). Bandwidth-sharing networks under a diffusion scaling. *Annals of Operations Research* **170**, 41-58.
- [2] Bazaraa, M.S., Sherali, H.D., Shetty, C.M. (1993). *Nonlinear Programming: Theory and Algorithms*. John Wiley, New Jersey.
- [3] Ben-Tal, A., Nemirovski, A. (2001). *Lectures on Modern Convex Optimization*. MPS-SIAM Series on Optimization, SIAM, Philadelphia.
- [4] Billingsley, P. (1999). *Convergence of Probability Measures*. John Wiley and Sons, New York.
- [5] Bonald, T., Massoulié, L. (2001). Impact of fairness on Internet performance. In: *Proc. ACM Sigmetrics & Performance 2001 Conf.*, Boston MA, USA, 82–91.
- [6] Bonald, T., Proutière, A. (2003). Insensitive bandwidth sharing in data networks. *Queueing Systems* **44**, 69–100.
- [7] Bonald, T., Proutière, A. (2004). On stochastic bounds for monotonic processor sharing networks. *Queueing Systems* **47**, 81–106.
- [8] Bonald, T., Massoulié, L., Proutiere, A., Virtamo, J. (2006). A queueing analysis of max-min fairness, proportional fairness and balanced fairness *Queueing Systems* **53**, 65–84.
- [9] Bonald, T., Roberts, J.W. (2001). Performance modeling of elastic traffic in overload. In: *Proc. ACM Sigmetrics & Performance 2001 Conf.*, Boston MA, USA, 342–343.
- [10] Bonnans, J.F., Shapiro, A. (2000). *Perturbation Analysis of Optimization Problems*. Springer, New York.

- [11] Borst, S.C., Egorova, R., Zwart, A.P. (2009). Fluid limits of bandwidth-sharing networks in overload. Submitted for publication.
- [12] Boyd, S., Vanderberghe, L. (2004). *Convex Optimization*. University Press, Cambridge.
- [13] Bramson, M. (2005). Stability of networks for max-min fair routing. Presentation at the 13th INFORMS Applied Probability Conference, Ottawa.
- [14] Chiang, M., Shah, D., Tang, A. (2006). Stochastic stability of network utility maximization: General file size distribution. In: *Proc. Allerton 2006 Conf.*
- [15] Egorova, R., Borst, S.C., Zwart, A.P. (2007). Bandwidth-sharing networks in overload. *Performance Evaluation* **64**, 978–993.
- [16] Noah Gans, Ger Koole, & Avishai Mandelbaum (2003). Telephone call centers: Tutorial, review, and research prospects. *Manufacturing and Service Operations Management* **5**, 79–141.
- [17] Gromoll, H.C., Puha, A.L., Williams, R.J. (2002). Fluid limit of a heavily loaded processor sharing queue. *Ann. Appl. Prob.* **12**, 797–859.
- [18] Gromoll, H.C., Robert, Ph., Zwart, A.P. (2008). Fluid limits for processor sharing queues with impatience. *Math. Oper. Res.* **33**, 375–402.
- [19] Gromoll, H.C., Robert, Ph., Zwart, A.P., Bakker, R. (2006). The impact of reneging in processor sharing queues. In: *Proc. ACM Sigmetrics & Performance 2006 Conf.*, St. Malo, France.
- [20] Gromoll, H.C., Williams, R.J. (2007). Fluid limit of a network with fair bandwidth sharing and general document size distribution. *Ann. Appl. Prob.*, to appear.
- [21] Gromoll, H.C., Williams, R.J. (2007). Fluid model for a data network with alpha-fair bandwidth sharing and general document size distributions: two examples of stability. *Proceedings of Markov Processes and Related Topics*, to appear.
- [22] Halfin, S., Whitt, W. (1981). Heavy traffic limits for queues with many exponential servers. *Operations Research* **29**, 567–588.
- [23] Kang, W. N., Kelly, F. P., Lee, N. H. and Williams, R. J. State space collapse and diffusion approximation for a network operating under a fair bandwidth-sharing policy. *Annals of Applied Probability* **19**, 1719-1780, 2009.
- [24] Karatzas, I., Shreve, S.E. (1991). *Brownian Motion and Stochastic Calculus*. Springer, New York.

- [25] Kelly, F.P., Williams, R.J. (2004). Fluid model for a network operating under a fair bandwidth-sharing policy. *Ann. Appl. Prob.* **14**, 1055–1083.
- [26] Kelly, F.P., Williams, R.J. (2010). Heavy traffic on a controlled motorway. In: *Probability and Mathematical Genetics: Papers in Honour of Sir John Kingman*.
- [27] Mandelbaum, A., Massey, W.A. and Reiman, M.I. (1998). Strong approximations for Markovian service networks. *Queueing Systems* **30**, 149–201.
- [28] Massoulié, L. (2007). Structural properties of proportional fairness: stability and insensitivity. *Ann. Appl. Prob.* **17**, 809–839.
- [29] Massoulié, L., Roberts, J.W. (1999). Bandwidth sharing: objectives & algorithms. In: *Proc. IEEE Infocom '99*, New York NY, USA, 1395–1403.
- [30] Mo, J., Walrand, J. (2000). Fair end-to-end window-based congestion control. *IEEE/ACM Trans. Netw.* **8**, 556–567.
- [31] Puha, A.L., Stolyar, A.L., Williams, R.J. (2006). The fluid limit of an overloaded processor sharing queue. *Math. Oper. Res.* **31**, 316–350.
- [32] Reed, J., Ward, A (2004). A diffusion approximation for a generalized Jackson network with renegeing. In Proceedings of the 42nd Allerton Conference on Communication, Control, and Computing
- [33] Roberts, J.W., Massoulié, L. (1998). Bandwidth sharing and admission control for elastic traffic. In: *Proc. ITC Specialist Seminar*, Yokohama, Japan.
- [34] De Veciana, G., Lee, T.-L., Konstantopoulos, T. (1999). Stability and performance analysis of networks supporting services with rate control – could the Internet be unstable? In: *Proc. Infocom '99*, New York NY, USA, 802–810.
- [35] De Veciana, G., Lee, T.-L., Konstantopoulos, T. (2001). Stability and performance analysis of networks supporting elastic services. *IEEE/ACM Trans. Netw.* **9**, 2–14.
- [36] Ye, H. and Yao, D.D., Heavy-Traffic Optimality of a Stochastic Network under Utility-Maximizing Resource Control. *Operations Research*, 56 (2008), 453-470.
- [37] Ye, H. and Yao, D.D., Utility-Maximizing Resource Control: Diffusion Limit and Asymptotic Optimality for a Two-Bottleneck Model. *Operations Research*, forthcoming