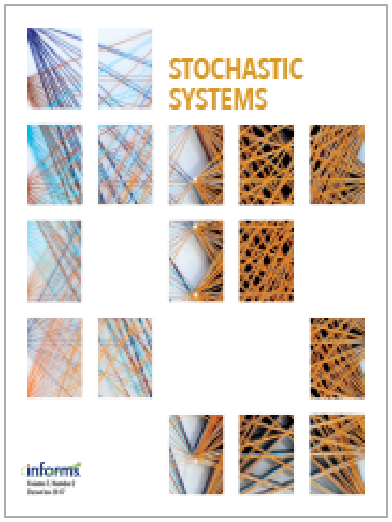


Stochastic Systems

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

Stochastic Monotonicity of Markovian Multiclass Queueing Networks

Haralambie Leahu, Michel Mandjes

To cite this article:
Haralambie Leahu, Michel Mandjes (2019) Stochastic Monotonicity of Markovian Multiclass Queueing Networks. Stochastic Systems 9(2):141-154. https://doi.org/10.1287/stsy.2018.0022

informs
http://pubsonline.informs.org/journal/stsy/

STOCHASTIC SYSTEMS
Vol. 9, No. 2, June 2019, pp. 141–154
ISSN 1946-5238 (online)

Stochastic Monotonicity of Markovian Multiclass Queueing Networks

Haralambie Leahu,[a] **Michel Mandjes**[b]

[a] Korteweg-de Vries Institute for Mathematics, University of Amsterdam, 1098 XG Amsterdam, Netherlands; [b] Center for Mathematics and Computer Science, 1098 XG Amsterdam, Netherlands

Contact: haralambie.leahu@cwi.nl (HL); m.r.h.mandjes@uva.nl, https://orcid.org/0000-0001-6783-4833 (MM)

Received: November 1, 2017
Accepted: August 27, 2018
Published Online in Articles in Advance: May 23, 2019

MSC2010 Subject Classification: Primary 60J28; secondary 65C40, 93D20

https://doi.org/10.1287/stsy.2018.0022

Abstract. Multiclass queueing networks (McQNs) extend the classical concept of the Jackson network by allowing jobs of different classes to visit the same server. Although such a generalization seems rather natural, from a structural perspective, there is a significant gap between the two concepts. Nice analytical features of Jackson networks, such as stability conditions, product–form equilibrium distributions, and stochastic monotonicity, do not immediately carry over to the multiclass framework. The aim of this paper is to shed some light on this structural gap, focusing on monotonicity properties. To this end, we introduce and study a class of Markov processes, which we call *Q-processes*, modeling the time evolution of the network configuration of any open, work-conservative McQN having exponential service times and Poisson input. We define a new monotonicity notion tailored for this class of processes. Our main result is that we show monotonicity for a large class of McQN models, covering virtually all instances of practical interest. This leads to interesting properties that are commonly encountered for "traditional" queueing processes, such as (i) monotonicity with respect to external arrival rates and (ii) star-convexity of the stability region (with respect to the external arrival rates); such properties are well known for Jackson networks but had not been established at this level of generality. This research was partly motivated by the recent development of a simulation-based method that allows one to numerically determine the stability region of a McQN parameterized in terms of the arrival-rates vector.

History: The former designation of this paper was SSY-2017-235.

Funding: The research for this paper is partly funded by the NWO Gravitation Programme Networks [Grant 024.002.003].

Keywords: Markov processes • stability • monotonicity

1. Introduction

Multiclass queueing networks (McQNs) arise as natural generalizations of conventional Jackson networks: although in Jackson networks each station (server) acts as a $\cdot/M/1$ *single-class* queue, in McQNs each network station is a *multiclass* queue. McQNs are particularly suitable for describing complex manufacturing systems (to be thought of as assembly lines) as they allow jobs (or, in queuing lingo, customers) visiting the same station multiple times to have different service requirements and/or a different routing scheme. One can think of situations in which a piece entering the system undergoes some physical transformation during the process; hence, its processing time at a given station (as well as its next destination) might depend on the processing stage. Other possible applications include packet transmission models in telecommunication networks and distributed systems in computer science, in which packets/tasks of different types can be routed to the same server to control resource utilization.

1.1. Motivation and Background

This generalization involves a number of complications. For instance, the following mathematical challenges arise:

I. Although the network-configuration process is expected to be Markovian (provided that all service and exogenous interarrival times are independent, exponentially distributed), the Markovian structure (the state-space and the transition dynamics) of an McQN is by no means straightforward as it depends on the service disciplines of the underlying stations. Although for some particular service disciplines the structure becomes quite simple, a unified (Markovian) framework is lacking, and this makes the analysis of such networks rather difficult.

The Markovian modeling of McQNs is important for both theoretical and practical reasons. On one hand, it allows one to use the powerful Markov process machinery to derive analytical properties of the underlying network-configuration process, and on the other hand, it facilitates the use of standard simulation methods.

II. Stochastic monotonicity is, in general, a desirable property that is widely used in applications, for example, optimization. Jackson networks satisfy the following monotonicity condition: if one increases any flow of jobs entering the network, then the resulting queues (at each station at any given time) will not decrease (in a stochastic sense) (Shanthikumar and Yao 1989). However, a straightforward generalization of this result to the multiclass framework is impeded by the inhomogeneous nature of the queues. While single-class queues can be naturally identified and ordered by means of their length, in the multiclass framework a total ordering of the queue configurations is not possible.

III. Stability is arguably a crucial (asymptotic) property of a queueing network. In this context, stability refers to positive recurrence of the associated Markov process. For Jackson networks, stability is equivalent to subcriticality, that is, traffic rate below one at every queue. In the multiclass framework, however, this is not the case as illustrated by various counter examples; see Kumar and Seidman (1990), Rybko and Stolyar (1993), Bramson (1994a), Seidman (1994), and Dai (1995). As such, stability conditions are not available in closed form in general.

In addition, stability is closely related to monotonicity properties. More specifically, validity of monotonicity properties is expected to imply that stability of a queueing network is a monotone property. This is indeed the case for Jackson networks as stability (to be thought of as subcriticality) is a monotone property with respect to (external) arrival, service and traffic rates. For McQNs, however, some results in the literature indicate that this is not necessarily the case when parameterized with respect to service rates (Bramson 1994b, Dai et al. 1999) or traffic rates (Dumas 1997). This naturally raises the question (but also casts some doubts on) whether stability of a McQN is monotone with respect to (external) arrival rates.

I. Challenge I has been addressed in Bramson (2008) in which a Markovian formalism has been proposed in a fairly general framework (with no restrictions over the underlying distributions).

II. Challenge II is a classical topic in applied probability and has attracted considerable attention over the past decades. We refer to Shanthikumar and Yao (1989) for monotonicity results pertaining to queueing networks and Massey (1987) (the references therein) for a standard theory tailored to Markov processes. Both approaches are suited for Jackson networks but fail to provide satisfactory results in the multiclass setup.

Finally, much research was invested in the 1990s into challenge III. The most successful approach to studying the stability of McQNs is based on fluid (model) limits, an asymptotic technique originally introduced in Rybko and Stolyar (1993) and further expanded in Dai (1995), which relates the stability of an McQN to that of its associated fluid model. Namely, stability of the fluid model implies that of the McQN; see Bramson (2008), section 5.5, for more background, including counterexamples showing that the converse is not true.

Stochastic monotonicity of (Markovian) McQNs became relevant recently when, motivated by the lack of closed-form stability conditions, the authors investigated simulation-based methods for approximating stability regions (Leahu and Mandjes 2017). Extensive simulation experiments indicated that a certain form of (stochastic) monotonicity with respect to external arrival rates could still be expected even for McQNs for which no stability conditions are known. Such a property would be enough to guarantee, for instance, that stability is a monotone property with respect to arrival rates.

1.2. Contributions

In this paper, we restrict our analysis to *Markovian* McQNs. To this end, we introduce a new class of Markovian processes, called Q-processes, formalizing the network-configuration process associated with a Markovian McQN. Concerning questions I–III, the main contributions of this paper are the following:

1. We develop a novel stochastic monotonicity concept tailored to Q-processes, called $\mathcal{F}$-monotonicity, and identify a set of rather general conditions (fulfilled by virtually all McQNs of practical interest) that guarantee $\mathcal{F}$-monotonicity; see Theorem 1. Furthermore, we show that $\mathcal{F}$-monotonicity implies monotonic behavior with respect to both (external) arrival rates and (when started empty) with time, extending in this way the well-known properties of the M/M/1 queue.

2. Second, we prove that for $\mathcal{F}$-monotone Q-processes stability is a monotone property with respect to external arrival rates; see Theorem 2. In particular, the stability region is an open, star-shaped domain, having the origin as a vantage point. This result formally validates the numerical findings in Leahu and Mandjes (2017).

1.3. Approach

Restricting our analysis to *Markovian* McQNs is necessary for several reasons. In the first place, it makes the corresponding network-configuration process Markovian, which significantly simplifies the formalism. Second, deviating from the Markovian framework leads to serious complications as stability (Vande Vate et al. 2004) and monotonicity (Whitt 1993) are sensitive to changing the shape of the underlying distributions.

A prerequisite for our investigation is an appropriate modeling framework that is general enough to cover a wide range of Markovian queue-related processes. More specifically, we need to define a state-space that is wide enough to accommodate various types of queue configurations together with an underlying structure that allows one to define the (Markovian) transitions performed by the network-configuration process. We formalize the queue at a given station as an ordered sequence of digits, in which the class of each job in the queue is represented by a specific digit; the order of the digits is interpreted as the order in which they are due to receive service as long as no other arrivals occur in the queue. Furthermore, to accommodate various queueing disciplines, for example, priority rules, we need to introduce a family of *insertion operators* indicating how an extradigit (representing the class of a new job arriving in the queue) is placed in the sequence. Finally, to cover processor-sharing disciplines, we shall also introduce the concept of *service allocation*, that is, a probability distribution on the set of digits specifying the fraction of service allocated by the server to each class present in the queue. The space of ordered sequences endowed with a family of insertion operators and a service allocation is called a *space of multiclass configurations* and is the building block for defining a Q-process; these facts are formalized in Section 3.

Furthermore, we consider the concept of *$\mathscr{F}$-monotonicity*, a weaker form of stochastic monotonicity for Markov processes on ordered spaces. To be more specific, consider a Markov process on a (partially) ordered space. The ordering induces a class of (real-valued) increasing functions and a stochastic ordering on the probability distributions on the underlying state-space (Massey 1987). Standard stochastic monotonicity— compare with Massey (1987)—presumes that two versions of the process started in a pair of ordered states remain (stochastically) ordered at any time. Such a property does not hold, in general, for Q-processes. On the other hand, $\mathscr{F}$-monotonicity, for some given (sub)class of increasing functions $\mathscr{F}$, still presumes an ordering between the two versions but with respect to a different (weaker) stochastic ordering, determined by the subclass $\mathscr{F}$. Although $\mathscr{F}$-monotonicity is weaker than the standard stochastic monotonicity for Markov processes on ordered spaces and requires essentially different proof techniques, it still retains most of the relevant properties of stochastic monotonicity and, as it turns out, it is better suited in the context of Q-processes.

Finally, we consider the vector of (external) arrival rates as a parameter of the Q-process corresponding to a given McQN and express the corresponding stability region (the set of parameters that make the process stable) as the support of some limiting functional of the process, defined on the parameter space. Based on analytical properties of this functional, we derive relevant properties of the corresponding stability region.

1.4. Organization of the Paper

The paper is organized as follows. In Section 2, we provide a brief account of the mathematical concept of McQN. In Section 3, we define the concept of Q-process, that is, the general stochastic process model for the dynamics of a Markovian McQN. Then, in Section 4, we introduce and elaborate on the (novel) concept of $\mathscr{F}$-monotonicity, and in Section 5, we investigate stability properties of $\mathscr{F}$-monotone Q-processes. Finally, in Section 6, we illustrate the practical importance of our results by pointing out their relevance in developing numerical methods for evaluating the stability region associated with a Q-process (respectively, McQN).

1.5. Notations and Conventions

In this paper, we employ the following notation. In the first place, by $\mathbb{N}$ we denote the set of nonnegative integers $\mathbb{N} := \{0, 1, 2, \ldots\}$ and by $\mathbb{R}$ the set of real numbers. For a denumerable set $\mathscr{J}$ and $\mathbb{S} = \mathbb{N}, \mathbb{R}$, we denote by $\mathbb{S}^{\mathscr{J}}$ the set of $\mathscr{J}$-labeled vectors over $\mathbb{S}$; alternatively, $\mathbb{S}^{\mathscr{J}}$ defines the space of all mappings $u : \mathscr{J} \longrightarrow \mathbb{S}$. When $\mathscr{J} = \{1, \ldots, d\}$, we use the simplified notation $\mathbb{S}^d$. Moreover, $\mathbf{0}$ denotes the null vector $(0, \ldots, 0)$, and $\mathbf{1}_\Gamma$, for $\Gamma \subseteq \mathscr{J}$, denotes the characteristic vector (mapping) of Γ, defined via

$$(\mathbf{1}_\Gamma)_\jmath = \begin{cases} 1, & \jmath \in \Gamma, \\ 0, & \jmath \notin \Gamma. \end{cases}$$

When $\Gamma = \mathscr{J}$, we use the simplified notation $\mathbf{1}$ instead $\mathbf{1}_\Gamma$.

For an arbitrary vector $\boldsymbol{x} = (x_\jmath : \jmath \in \mathscr{J}) \in \mathbb{S}^{\mathscr{J}}$, we define the 1-norm,

$$\|\boldsymbol{x}\| := \sum_{\jmath \in \mathscr{J}} |x_\jmath|,$$

whenever the right-hand side (r.h.s.) is finite, and we extend this notation to bounded linear operators $\boldsymbol{U}$ defined on (subspaces of) $\mathbb{S}^{\mathscr{J}}$, namely $\|\boldsymbol{U}\| := \sup\{|\boldsymbol{U}\boldsymbol{x}| : \|\boldsymbol{x}\| \leq 1\}$. On $\mathbb{S}^{\mathscr{J}}$, we denote the natural (component-wise) ordering

$\boldsymbol{x} \le \boldsymbol{z}$, respectively, $\boldsymbol{x} < \boldsymbol{z}$, if $x_\jmath \le z_\jmath$, respectively $x_\jmath < z_\jmath$, for $\jmath \in \mathscr{J}$. Finally, $\varsigma[\boldsymbol{x}] := \{\jmath \in \mathscr{J} : x_\jmath \neq 0\}$ denotes the support of $\boldsymbol{x} \in \mathbb{S}^{\mathscr{J}}$, and $\boldsymbol{I}$ denotes the identity operator on $\mathbb{S}^{\mathscr{J}}$.

2. Markovian Multiclass Queueing Networks

Following the exposition in Dai (1995), we consider a multiserver network, comprised of $\aleph \ge 1$ single servers, labeled $1, \ldots, \aleph$. The network is used by $d \ge 1$ classes of jobs in such a way that each class k job, at any time, requires service at a fixed server, denoted by $\mathscr{S}(k)$. Once the service at $\mathscr{S}(k)$ is finished, it either becomes a job of class l with probability R_{kl} (independently of all routing history) or leaves the system with (exit) probability

$$R_{k0} := 1 - \sum_{l=1}^{d} R_{kl}.$$

The routing matrix $\boldsymbol{R} = \{R_{kl}\}_{k,l=1,\ldots,d}$ is assumed transient (substochastic); that is,

$$\boldsymbol{I} + \boldsymbol{R} + \boldsymbol{R}^2 + \ldots = (\boldsymbol{I} - \boldsymbol{R})^{-1}. \tag{1}$$

This guarantees that any job entering the network visits a finite number of classes before leaving the network almost surely; in standard queueing language, the network is said to be *open*.

Each class k has its own exogenous (possibly null) arrival stream, regulated by a Poisson process with rate $\theta_k \ge 0$ and requires independent and identically distributed service times, exponentially distributed with rate $\beta_k > 0$, independent of everything else; a null arrival process corresponds to a class with no external input, which models, for instance, intermediate processing stages of a certain class. Note that any class is identified with its server, its specific routing probabilities, its specific exogenous arrival process, and its specific service-time distribution.

We let $\mathscr{K}_i := \mathscr{S}^{-1}(i)$ denote the set of classes served by station i and assume (without loss of generality) that $\mathscr{K}_i \neq \emptyset$ for all i (equivalently, the mapping $\mathscr{S}$ is surjective); that is, any server is used. In particular, it holds that $\aleph \le d$.

Finally, each server employs its own nonidling service discipline (i.e., different servers may have different service disciplines) and has infinite buffer capacity.

The previously introduced McQN concept extends many known classes of queueing networks. For instance, when the mapping $\mathscr{S}$ is bijective (in particular, $\aleph = d$) one obtains a *Jackson network*. Moreover, if the service rate for any class k only depends on the server $\mathscr{S}(k)$, then we recover the concept of the *Kelly network*; see Kelly (1975). Finally, if there exists only one class with a nonnull exogenous arrival process and all jobs have the same (deterministic) routing, visiting all classes exactly once (in the same order), then the network is called a *reentrant line*. Reentrant lines provide popular instances of McQNs as they can be used to model (assembly) manufacturing lines.

3. Q-Processes

The objective of this section is to introduce the concept of the Q-process, a Markovian process modeling the network configuration dynamics (over time) of a Markovian McQN as described in Section 2. To construct a model general enough to accommodate all the usual service disciplines, we need a rather intricate formalism, which we briefly introduce here. Technical details and definitions can be found in Appendix A.

The building blocks of the Q-process model are the *spaces of multiclass configurations*, which formalize the basic structure required for modeling the dynamics of a single-server, multiclass queue. More specifically, we consider a finite set $\mathscr{K}$ of job classes to be processed by a server. Because jobs of the same class are (probabilistically) exchangeable, the class of all possible queue configurations is modeled by the augmented space $\overline{\mathbb{Q}}[\mathscr{K}] = \mathbb{Q}[\mathscr{K}] \cup \{\emptyset\}$, where

$$\mathbb{Q}[\mathscr{K}] := \{p = (k_1, \ldots, k_n) : n \ge 1, k_1, \ldots, k_n \in \mathscr{K}\} \tag{2}$$

denotes the space of all finite (ordered) sequences with elements in $\mathscr{K}$.

Furthermore, to formalize the queue configuration dynamics, we need to introduce some additional structure on $\overline{\mathbb{Q}}[\mathscr{K}]$, modeling arrival and departure events. More specifically, we need

- Insertion operators $I_k : \overline{\mathbb{Q}}[\mathscr{K}] \longrightarrow \mathbb{Q}[\mathscr{K}]$, which specify how an incoming job of class $k \in \mathscr{K}$ is placed in the (new) queue configuration; that is, there exists a representation (concatenation) $p = (p', p'')$ such that $I_k(p) = (p', k, p'')$; both p' and p'' are allowed to be $\emptyset$, but p'' may not contain any k-digits (jobs in the same class may not overtake each other). The family of all insertion operators $\{I_k : k \in \mathscr{K}\}$ defines a *queue policy* on $\overline{\mathbb{Q}}[\mathscr{K}]$.
- Deletion operators $D_k : \mathbb{Q}[\mathscr{K}] \longrightarrow \overline{\mathbb{Q}}[\mathscr{K}]$, which specify how a job of class $k \in \mathscr{K}$ is removed from the queue configuration.

- A service allocation (mapping) $W := (W_k : k \in \mathcal{K}) : \mathbb{Q}[\mathcal{K}] \longrightarrow \mathcal{P}[\mathcal{K}]$, where $\mathcal{P}[\mathcal{K}]$ denotes the simplex of probability vectors $\boldsymbol{w} := (w_k : k \in \mathcal{K})$ on $\mathcal{K}$, specifying which class(es) receive service and the corresponding fraction of server capacity in a given (nonempty) configuration. Because $W_k(p)$ denotes the fraction of server capacity allocated to class k, we impose that $W_k(p) = 0$ if there is no k-digit in the configuration p.

These elements relate to an McQN model as follows. The insertion operators are set in accordance with the queue discipline of the server. The order of the digits in a sequence p does not necessarily reflect the order of arrivals but rather the order in which jobs will be considered for service. Furthermore, a deletion operator D_k removes (by convention) the first k-digit from the sequence; if no such digit exists, it leaves the sequence unchanged. Finally, a service allocation distributes the server capacity among the jobs/classes present in the queue; that is, each class receives a given fraction of server capacity (which, by convention, is assigned to its first representative) in accordance with the service discipline of the server.

Note that in this modeling paradigm the service discipline of a particular server is identified with a combination of queue policy and service allocation, which, in principle, can be chosen independently from each other. Note, however, that in some cases there are multiple ways to model a particular service discipline; for instance, when the service allocation is order insensitive (which is typically the case for processor-sharing disciplines), the queue policy becomes irrelevant (a default can be used). For specific examples of queue policies and service allocations and how they relate to the usual service disciplines, see Appendix A.

Remark 1. The mappings I_k, D_k, and W_k (which are defined on $\mathbb{Q}[\mathcal{K}]$) admit natural extensions to the augmented space $\overline{\mathbb{Q}}[\mathcal{K}]$ as follows: $I_k(\emptyset) := (k)$, $D_k(\emptyset) := \emptyset$, and $W_k(\emptyset) := 0$; note, however, that $W(\emptyset) = \mathbf{0} \notin \mathcal{P}[\mathcal{K}]$. ⋄

Definition 1. A space of multiclass configurations over $\mathcal{K}$ is the space,

$$\{\overline{\mathbb{Q}}[\mathcal{K}]; I_k, D_k, W_k : k \in \mathcal{K}\},$$

of finite ordered sequences over $\mathcal{K}$, endowed with a family of insertion/deletion operators $\{I_k/D_k : k \in \mathcal{K}\}$ and service allocation $W = (W_k : k \in \mathcal{K})$. ⋄

Let $1 \le \aleph \le d$ and assume that $\{\mathcal{K}_i\}_{i=1}^{\aleph}$ is a partition of the set $\{1, \dots, d\}$. Furthermore, we assume that $\{\overline{\mathbb{Q}}[\mathcal{K}_i]; I_k, D_k, W_k : k \in \mathcal{K}_i\}$ is a space of multiclass configurations over $\mathcal{K}_i$ for $i = 1, \dots, \aleph$ and let

$$\mathbb{X} := \overline{\mathbb{Q}}[\mathcal{K}_1] \times \cdots \times \overline{\mathbb{Q}}[\mathcal{K}_\aleph] = \{[p_1, \dots, p_\aleph], p_i \in \overline{\mathbb{Q}}[\mathcal{K}_i], i = 1, \dots, \aleph\} \tag{3}$$

denote the space of all possible configurations of an McQN with $\aleph$ stations and d classes, in which classes in $\mathcal{K}_i$ are assigned to queue/station i; obviously $\mathbb{X}$ is denumerable. Furthermore, we denote the empty configuration on $\mathbb{X}$ by $\emptyset := [\emptyset, \dots, \emptyset]$. Finally, for $\xi = [p_1, \dots, p_\aleph] \in \mathbb{X}$ we define $W_k(\xi) := W_k(p_i)$ provided that $k \in \mathcal{K}_i$.

On the class of real-valued functions $\mathbb{R}^{\mathbb{X}}$, we define the following linear operators:

- For $h : \mathbb{X} \longrightarrow \mathbb{R}$, we define the h-multiplication operator $\Pi[h]\phi := h \cdot \phi$.
- For $f : \mathbb{X} \longrightarrow \mathbb{X}$, we define the f-composition operator $\Phi[f]\phi := \phi \circ f$.

We are now in the position to formally define the concept of the Q-process.

Definition 2. Consider the following numerical elements:

- Two vectors $\boldsymbol{\theta} = (\theta_k : k = 1, \dots, d) \ge \mathbf{0}$, $\boldsymbol{\beta} = (\beta_k : k = 1, \dots, d) > \mathbf{0}$.
- A substochastic matrix $\boldsymbol{R} = \{R_{kl}\}_{k,l=1,\dots,d}$, satisfying (1).

A Q-process defined by parameter $(\boldsymbol{\theta}, \boldsymbol{\beta}, \boldsymbol{R})$ is the continuous-time Markov chain $\mathcal{X} := \{X_t\}_{t \ge 0}$ on the space $\mathbb{X}$ given by (3), having generator

$$\boldsymbol{A} := \sum_{(k,l) \in \mathcal{T}} \Pi\big[h_{(k,l)}\big]\big(\Phi[f_{(k,l)}] - \boldsymbol{I}\big), \tag{4}$$

where $\mathcal{T} := \{0, 1, \dots, d\}^2 \setminus \{(0,0)\}$, and for $\xi = [p_1, \dots, p_\aleph] \in \mathbb{X}$, we define

- $f_{(0,k)}(\xi) = [p_1, \dots, I_k(p_i), \dots, p_\aleph]$ and $f_{(k,0)}(\xi) = [p_1, \dots, D_k(p_i), \dots, p_\aleph]$;
- $f_{(k,l)} = f_{(0,l)} \circ f_{(k,0)}$ for $l \neq 0$;
- $h_{(0,k)}(\xi) = \theta_k$ and $h_{(k,l)}(\xi) = \beta_k W_k(p_i) R_{kl}$ for $k \in \mathcal{K}_i$, $i = 1, \dots, \aleph$, and $l = 0, 1, \dots, d$. ⋄

Regarding Definition 2, a few remarks are in order.

- In the definition, $(0,k)$-transitions correspond to external arrivals to class k, $(k,0)$-transitions correspond to external departures from class k, and (k,l)-transitions correspond to switches from class k to class l.
- For any $(k,l) \in \mathcal{T}$, $h_{(k,l)}(\xi)$ gives the transition rate corresponding to a (k,l)-transition from state ξ, and $f_{(k,l)}(\xi)$ denotes the state-space transform after performing a (k,l)-transition from ξ.

- In some situations, depending on the individual service disciplines, the resulting Q-process is lumpable; that is, the state-space $\mathbb{X}$ can be partitioned in equivalence classes, and the resulting quotient process is still Markov. In such cases, the reduced models agree with the ones in Dai (1995); see Appendix A.
- For any $(k,l) \in \mathscr{T}$, $h_{(k,l)}(\xi)$ gives the transition rate corresponding to a (k,l)-transition from state ξ, and $f_{(k,l)}(\xi)$ denotes the state-space transform after performing a (k,l)-transition from ξ. A Q-process is called *elementary* if $\aleph = d$. Elementary Q-processes correspond to Jackson network models. They are lumpable with quotient space $\mathbb{N}^d$.

In accordance with the Markovian formalism, we introduce the following notations: for a Q-process $\mathscr{X} = \{X_t : t \geq 0\}$ on $\mathbb{X}$, we denote by $\boldsymbol{P}^t$ the associated transition operator, defined as (for suitable ϕ)

$$\forall t \geq 0, \xi \in \mathbb{X} : \boldsymbol{P}^t(\xi, \phi) := \mathbb{E}[\phi(X_t) | X_0 = \xi] = \mathbb{E}^{\xi}[\phi(X_t)].$$

We further set $\boldsymbol{P}^t(\xi, \Omega) := \boldsymbol{P}^t(\xi, \mathbf{1}_\Omega)$ for $\Omega \subseteq \mathbb{X}$. The reader should bear in mind that a Q-process depends (numerically) on the parameters $(\boldsymbol{\theta}, \boldsymbol{\beta}, \boldsymbol{R})$; to emphasize the dependence on $\boldsymbol{\theta}$, we use the notation $\boldsymbol{P}^t_{\boldsymbol{\theta}}$ and $\mathbb{E}_{\boldsymbol{\theta}}$.

Note that the generator in (4) defines a bounded linear operator on $\mathscr{C}_0[\mathbb{X}]$, that is, the space of functions $\phi \in \mathbb{R}^{\mathbb{X}}$ vanishing at infinity,[1] having norm

$$\|\boldsymbol{A}\| = \sum_{k=1}^{d} \theta_k + \sum_{i=1}^{\aleph} \max_{k \in \mathscr{K}_i} \beta_k < \infty.$$

In particular, for $\phi \in \mathscr{C}_0[\mathbb{X}]$, it holds that

$$\forall t \geq 0, \xi \in \mathbb{X} : \boldsymbol{P}^t(\xi, \phi) = [\exp(t\boldsymbol{A})\phi](\xi). \tag{5}$$

Remark 2. Because $\|\boldsymbol{A}\| < \infty$ the (bounded) linear operator $\boldsymbol{A}$ extends in a natural way to the class of bounded functions in $\mathbb{R}^{\mathbb{X}}$ (having the same norm), and so does $\exp(t\boldsymbol{A})$; hence, (5) extends to all bounded ϕ's. ◇

Furthermore, $\boldsymbol{Q} := \boldsymbol{I} + (1/\lambda)\boldsymbol{A}$ defines a Markov (transition) operator on $\mathbb{X}$ for any $\lambda \geq \|\boldsymbol{A}\|$ and

$$\exp(t\boldsymbol{A}) = \exp[\lambda t(\boldsymbol{Q} - \boldsymbol{I})] = \exp(-\lambda t) \sum_{n \geq 0} \frac{(\lambda t)^n}{n!} \boldsymbol{Q}^n. \tag{6}$$

This property is called *uniformization* and allows one to sample the process $\mathscr{X}$ via the so-called uniformized (Markov) λ*-chain* as follows: letting $\Xi = \{\Xi_n : n \geq 0\}$ denote the Markov chain with transition operator $\boldsymbol{Q}$, a random sample from X_t can be obtained as Ξ_{N_t}, where $\{N_t : t \geq 0\}$ denotes a Poisson process with rate λ; that is, X_t and Ξ_{N_t} coincide in distribution as readily follows from (6).

4. Stochastic Monotonicity of Q-Processes

In this section, we first introduce a (stochastic) monotonicity concept tailored to Q-processes and then deduce further properties of Q-processes satisfying such monotonicity assumptions. In addition, we show that this type of monotonicity is quite common for Q-processes by identifying a pair of regularity conditions on the underlying queue policies and service allocations that guarantee monotonicity of a Q-process. Importantly, these conditions are met by virtually all usual queue policies and service allocations used in applications; see Appendix A.

If $\mathscr{K}$ is an arbitrary set of classes, we define the canonical partial ordering $\subseteq$ on $\overline{\mathbb{Q}}[\mathscr{K}]$ as follows: $p \subseteq q$ if the digits of p can be identified among the digits of q *in the same order*. Formally, $\emptyset \subseteq p$ for any p, and if $p = (k_1, \ldots, k_m)$ and $q = (l_1, \ldots, l_n)$ with $1 \leq m \leq n$, then $p \subseteq q$ iff there exists some increasing sequence $\nu_1 < \ldots < \nu_m$ satisfying $k_\imath = l_{\nu_\imath}$ for $\imath = 1, \ldots, m$. In addition, p and q are called *consecutive* if $n = m + 1$; in this case, q can be obtained by inserting an extra digit in the sequence p in some arbitrary position. Finally, we note that $D_k(p) \subseteq p \subsetneq I_l(p)$ for $k, l = 1, \ldots, d$ and that p and $I_l(p)$ are consecutive sequences for any queue policy.

Furthermore, we extend $\subseteq$ to $\mathbb{X}$, defined in (3), as follows: if $\xi = [p_1, \ldots, p_\aleph] \in \mathbb{X}$ and $\zeta = [q_1, \ldots, q_\aleph] \in \mathbb{X}$, then $\xi \subseteq \zeta$ iff $p_i \subseteq q_i$ for $i = 1, \ldots, \aleph$. Also, we say that $(\xi, \zeta) \in \mathbb{X} \times \mathbb{X}$ is a pair of *consecutive configurations* if there exists some i_0 such that $p_i = q_i$ for $i \neq i_0$, whereas for $i = i_0$ the sequences p_i and q_i are consecutive. Note that for any such pair of consecutive configurations there exists exactly one $b \in \mathscr{K}_{i_0}$ such that p_{i_0} and q_{i_0} differ by exactly one b-digit; we denote by $\Delta_b \subseteq \mathbb{X} \times \mathbb{X}$ the set of all (pairs of) consecutive configurations on $\mathbb{X}$ differing by a b-digit so that the family $\{\Delta_b : b = 1, \ldots, d\}$ forms a partition of the set of all (pairs of) consecutive configurations. Note that

$f_{(k,0)}(\xi) \subseteq \xi \subsetneq f_{(0,l)}(\xi)$ and $(\xi, f_{(0,l)}(\xi)) \in \Delta_l$ for $\xi \in \mathbb{X}$ and $k, l = 1, \ldots, d$. In addition, for any $\xi \subsetneq \zeta$, there exists a sequence of consecutive configurations $\xi_0, \xi_1, \ldots, \xi_n$; such that $\xi_0 = \xi$ and $\xi_n = \zeta$.

Finally, the mapping $\phi : \mathbb{X} \longrightarrow \mathbb{R}$ is called *increasing* if $\xi \subseteq \zeta$ entails $\phi(\xi) \leq \phi(\zeta)$. Note that it suffices to verify that the latter property only holds for consecutive configurations (ξ, ζ). In particular, if ϕ is increasing, then it holds that $(\Phi[f_{(0,k)}] - \boldsymbol{I})\ \phi \geq \mathbf{0}$ (point-wise) for any $k = 1, \ldots, d$.

In what follows, we denote by $\mathscr{I}[\mathbb{X}]$ the class of increasing functions on $\mathbb{X}$ and let $\mathscr{F} \subseteq \mathscr{I}[\mathbb{X}]$. Our next definition introduces the concept of $\mathscr{F}$-monotonicity.

Definition 3. The Q-process $\mathscr{X}$, having generator $\boldsymbol{A}$, is called $\mathscr{F}$-monotone if there exists some $\lambda \geq \|\boldsymbol{A}\|$ such that the transition operator,

$$\mathbf{Q}^n = [\boldsymbol{I} + (1/\lambda)\boldsymbol{A}]^n,$$

maps $\mathscr{F}$ onto $\mathscr{I}[\mathbb{X}]$ for any $n \geq 0$; that is, $\boldsymbol{A}$ is $\mathscr{F}$-monotone if $\xi \subseteq \zeta$ and $\phi \in \mathscr{F}$ entails $[\mathbf{Q}^n\phi](\xi) \leq [\mathbf{Q}^n\phi](\zeta)$ for any $n \geq 0$. ⋄

If $\mathscr{F} = \mathscr{I}[\mathbb{X}]$ in Definition 3, then we call $\mathscr{X}$ *strongly* monotone.

Remark 3. $\mathscr{F}$-monotonicity of a Q-process requires that the transition operator $\mathbf{Q}^n$ maps $\mathscr{F}$ onto $\mathscr{I}[\mathbb{X}]$ for $n \geq 0$. In particular, if the statement in Definition 3 holds true for some $\lambda \geq \|\boldsymbol{A}\|$, then it holds for any $\lambda' \geq \lambda$ because

$$\boldsymbol{I} + \frac{1}{\lambda'}\boldsymbol{A} = \frac{(\lambda' - \lambda)}{\lambda'}\boldsymbol{I} + \frac{\lambda}{\lambda'}\boldsymbol{Q}.$$

Furthermore, we note that $\mathscr{F}$-monotonicity of a Q-process $\mathscr{X} = \{X_t : t \geq 0\}$ entails that $\boldsymbol{P}^t(\xi, \phi) \leq \boldsymbol{P}^t(\zeta, \phi)$ for any $t \geq 0$, for any $\xi \subseteq \zeta$, and $\phi \in \mathscr{F}$; this readily follows from (5) and (6). ⋄

$\mathscr{F}$-monotonicity is essentially different (in fact, weaker) from similar concepts introduced in Massey (1987), in which stochastic monotonicity amounts to $\boldsymbol{Q}$ leaving invariant some $\mathscr{F} \subseteq \mathscr{I}[\mathbb{X}]$. Although weaker than the standard stochastic monotonicity, $\mathscr{F}$-monotonicity still has rather powerful implications as shown by our next result.

Proposition 1. *Let $\mathscr{X}$ denote a Q-process that is $\mathscr{F}$-monotone for some given parameter $(\boldsymbol{\theta}, \boldsymbol{\beta}, \boldsymbol{R})$ for some $\mathscr{F} \subseteq \mathscr{I}[\mathbb{X}]$. Then*

I. *For every $\phi \in \mathscr{F}$, the mapping $t \longmapsto \boldsymbol{P}^t_{\boldsymbol{\theta}}(\emptyset, \phi)$ is nondecreasing.*

II. *If $\boldsymbol{\theta} \leq \boldsymbol{\vartheta}$, then $\boldsymbol{P}^t_{\boldsymbol{\theta}}(\xi, \phi) \leq \boldsymbol{P}^t_{\boldsymbol{\vartheta}}(\xi, \phi)$ for every $\xi \in \mathbb{X}, \phi \in \mathscr{F}$. The statement also holds true with $\leq$ replaced by $\geq$.*

In particular, if $\mathscr{X}$ is $\mathscr{F}$-monotone for any $\boldsymbol{\theta}$, the mapping $\boldsymbol{\theta} \longmapsto P^t_{\theta}(\xi, \phi)$ is nondecreasing for every $\xi \in \mathbb{X}, \phi \in \mathscr{F}$. ⋄

4.1. $\mathscr{F}$-Monotonicity for Elementary Q-Processes (Jackson Networks)

For elementary Q-processes, all service disciplines are equivalent (resulting in the same process; see Appendix A) so that the state-space $\mathbb{X}$ can be reduced to $\mathbb{N}^d$ and $\subseteq$ corresponds to the usual component-wise ordering $\leq$.

We claim that $\boldsymbol{Q} = \boldsymbol{I} + (1/\lambda)\boldsymbol{A}$ maps $\mathscr{I}[\mathbb{N}^d]$ onto itself for any $\lambda \geq \|\boldsymbol{A}\|$ for any parameter $(\boldsymbol{\theta}, \boldsymbol{\beta}, \boldsymbol{R})$. In fact, it suffices to prove that

$$\Pi[h_{(k,l)}](\Phi[f_{(k,l)}] - \boldsymbol{I}),$$

does that for each $(k, l) \in \mathscr{T}$. The last claim follows by Massey (1987, theorem 5.4) by noting that $\boldsymbol{x} \leq \boldsymbol{z}$ entails either

- $h_{(k,l)}(\boldsymbol{x}) = h_{(k,l)}(\boldsymbol{z})$, in which case it holds that $f_{(k,l)}(\boldsymbol{x}) \leq f_{(k,l)}(\boldsymbol{z})$, or
- $0 = h_{(k,l)}(\boldsymbol{x}) < h_{(k,l)}(\boldsymbol{z})$, in which case $k \neq 0$ and $x_k = 0$; hence,

$$\forall l = 0, 1, \ldots, d : \boldsymbol{x} \leq f_{(k,0)}(\boldsymbol{z}) \leq f_{(k,l)}(\boldsymbol{z}).$$

One concludes that elementary Q-processes are strongly monotone, meaning that $\boldsymbol{x} \leq \boldsymbol{z}$ entails $\boldsymbol{P}^t(\boldsymbol{x}, \phi) \leq \boldsymbol{P}^t(\boldsymbol{z}, \phi)$ for any $t \geq 0$ and $\phi \in \mathscr{I}[\mathbb{N}^d]$.

4.2. $\mathscr{F}$-Monotonicity for Nonelementary Q-Processes

In this subsection, we discuss $\mathscr{F}$-monotonicity in the nonelementary framework, that is, for Q-processes corresponding to McQNs in which at least one server is multiclass. It turns out that strong monotonicity does not carry over beyond the elementary framework, the main reason being that the same transition probabilities from a given pair of ordered states are *not* stochastically ordered (with respect to any sensible stochastic ordering that is compatible with $\subseteq$ on $\mathbb{X}$); hence, the results in Massey (1987) do not apply in this framework.

We establish, instead, a weaker form of monotonicity. More specifically, for some arbitrary set of classes $\mathscr{K}$, let us denote by $\ell : \mathbb{Q}[\mathscr{K}] \longrightarrow \mathbb{N}$ the mapping assigning to the sequence $p := (k_1, \ldots, k_n)$ its length $\ell(p) := n$. We extend the length mapping, as follows: first, we extend it to $\overline{\mathbb{Q}}[\mathscr{K}]$ by setting $\ell(\emptyset) := 0$ and finally on $\mathbb{X}$ in (3) via

$$\ell(\xi) := \sum_{i=1}^{\aleph} \ell(p_i).$$

Furthermore, let us consider the space

$$\mathscr{G} := \{\phi = G \circ \ell : G \in \mathbb{R}^{\mathbb{N}} \text{ is nondecreasing}\} \subseteq \mathscr{I}[\mathbb{X}], \tag{7}$$

that is, the class of functions that are nondecreasing in the total job population.

Our next result establishes sufficient conditions for $\mathscr{G}$-monotonicity.

Theorem 1. *Let $\mathscr{X}$ denote a Q-process (cf. Definition 2) satisfying that*

i. *for any $\xi \subseteq \zeta$ and $k = 1, \ldots, d$, it holds $f_{(0,k)}(\xi) \subseteq f_{(0,k)}(\zeta)$.*

ii. *for any pair of consecutive configurations $(\xi, \zeta) \in \Delta_b$ with $b = 1, \ldots, d$, it holds that $W_k(\zeta) \le W_k(\xi)$ for $k \ne b$; in particular, $W_b(\xi) \le W_b(\zeta)$.*

Then $\mathscr{X}$ is $\mathscr{G}$-monotone with $\mathscr{G}$ defined in (7) for any parameter $(\boldsymbol{\theta}, \boldsymbol{\beta}, \boldsymbol{R})$. ⋄

Conditions (i) and (ii) in Theorem 1 impose some monotonicity assumptions on the insertion operators, respectively, on the service allocations associated with the Q-process and can be verified for each server individually. Condition (i) requires that new arrivals do not influence the order of the jobs (already) in the queue (i.e., the insertion operators are order-preserving), whereas condition (ii) requires that the service capacity allocated to any class may not decrease by increasing the number of its representatives. Such conditions are fulfilled by a wide range of service disciplines used in queueing applications, for example, first-come, first-served (FCFS), static buffer priority or standard processor sharing disciplines; see Appendix A for details. As such, Theorem 1 covers virtually all McQNs of practical interest.

Note that, if $b \in \mathscr{K}_{i_0}$, then $W_k(\xi) = W_k(\zeta)$ for any $k \notin \mathscr{K}_{i_0}$ because the two configurations are identical at all servers $i \ne i_0$; therefore, the inequalities in condition (ii) are only relevant for $k \in \mathscr{K}_{i_0}$.

5. Stability of $\mathscr{F}$-Monotone Q-Processes

In this section, we discuss the stability of Q-processes and its relation to $\mathscr{F}$-monotonicity; here stability is understood in a Markovian sense. In particular, we consider a Q-process $\mathscr{X}$ with parameter $(\boldsymbol{\theta}, \boldsymbol{\beta}, \boldsymbol{R})$ and regard it as a parametric Markov process, regulated by the arrival-rate vector $\boldsymbol{\theta} \in \Theta$; hence, $\boldsymbol{\beta}$ and $\boldsymbol{R}$ are kept fixed. For such a process, we show that stability is a monotone property with respect to $\boldsymbol{\theta}$.

To start, we note that a Q-process is irreducible iff $(\boldsymbol{I} - \boldsymbol{R}')^{-1}\boldsymbol{\theta} > \boldsymbol{0}$, and the irreducibility support, denoted by $\mathbb{X}_0$, depends on the underlying queue policies but not on $\boldsymbol{\theta}$. In what follows, we let $\Theta \subseteq \{\boldsymbol{\theta} \ge \boldsymbol{0} : (\boldsymbol{I} - \boldsymbol{R}')^{-1}\boldsymbol{\theta} > \boldsymbol{0}\}$; hence, assuming that $\mathscr{X}$ is irreducible (on some $\mathbb{X}_0 \subseteq \mathbb{X}$) for any $\boldsymbol{\theta} \in \Theta$.

Remark 4. The empty configuration $\emptyset$ always belongs to the irreducibility support $\mathbb{X}_0$ because $\emptyset$ is attainable regardless of the queue policies. ⋄

By the standard theory, any irreducible continuous-time Markov chain is either transient or (null or positive) recurrent (Meyn and Tweedie 1993a, theorem 8.2.5); in addition, positive recurrence is equivalent to stability/ergodicity (Meyn and Tweedie 1993a, theorem 13.3.3) and guarantees the existence of an (essentially) unique equilibrium distribution supported on $\mathbb{X}_0$. We define the *stability region* associated with a Q-process $\mathscr{X}$ by

$$\Theta_s := \{\boldsymbol{\theta} \in \Theta : \mathscr{X} \text{ is positive recurrent under } \mathbb{P}_{\boldsymbol{\theta}}\};$$

here $\mathbb{P}_{\boldsymbol{\theta}}$ denotes the probability law of the Q-process. An alternative way to characterize stability is as follows: let $T_\xi := \inf\{t > 0 : X_t = \xi, X_{t-} \ne \xi\}$; for $\xi \in \mathbb{X}$, denote the (first) hitting time of the state ξ (after the process visited at least one different state). Then the process $\mathscr{X}$ is stable if and only if $\mathbb{E}^{\xi}_{\boldsymbol{\theta}}[T_\xi] < \infty$ for any $\xi \in \mathbb{X}_0$. Moreover, it suffices that the expected return time is finite (only) for some particular ξ; see, for example, Meyn and Tweedie (1993b).

For any bounded function $\phi : \mathbb{X} \longrightarrow \mathbb{R}$ and measure μ on $\mathbb{X}$, we set

$$\langle \mu, \phi \rangle := \int \phi(\xi)\mu(d\xi), \quad L_{\boldsymbol{\theta}}[\phi] := \mathbb{E}^{\emptyset}_{\boldsymbol{\theta}}\left[\int_0^{T_\emptyset} \phi(X_t)dt\right]. \tag{8}$$

Because T_ξ is always larger than the holding time of $\mathscr{X}$ in the initial state X_0, we have $L_{\boldsymbol{\theta}}[\mathbf{1}] = \mathbb{E}^{\emptyset}_{\boldsymbol{\theta}}[T_\emptyset] \geq 1/\|\boldsymbol{\theta}\|$; hence, stability amounts to $L_{\boldsymbol{\theta}}[\mathbf{1}] < \infty$. Furthermore, for $\boldsymbol{\theta} \in \Theta_{\mathrm{s}}$ and any bounded ϕ, it holds that $L_{\boldsymbol{\theta}}[\phi] < \infty$, and $L_{\boldsymbol{\theta}}[\phi]$ is continuous (even differentiable) in $\boldsymbol{\theta}$. In addition, denoting by $\pi_{\boldsymbol{\theta}}$ the limiting (equilibrium) distribution of $\mathscr{X}$ under $\mathbb{P}_{\boldsymbol{\theta}}$, the regenerative ratio formula $\langle \pi_{\boldsymbol{\theta}}, \phi \rangle = L_{\boldsymbol{\theta}}[\phi]/L_{\boldsymbol{\theta}}[\mathbf{1}]$ holds; see Asmussen and Glynn (2007). Hence, $\boldsymbol{\theta} \longmapsto \langle \pi_{\boldsymbol{\theta}}, \phi \rangle$ is continuous on Θ_{s} for any bounded $\phi \in \mathbb{R}^{\mathbb{X}}$. Finally, note that the stability region Θ_{s} is open in Θ since

$$\Theta_{\mathrm{s}} = \bigcup_{r \geq 0} \{\boldsymbol{\theta} \in \Theta : L_{\boldsymbol{\theta}}[\mathbf{1}] < r\}.$$

The main result of this section provides a characterization of the stability region Θ_{s} of an $\mathscr{F}$-monotone Q-process.

Theorem 2. *Let $\mathscr{X} = \{X_t : t \geq 0\}$ be an $\mathscr{F}$-monotone Q-process for all $\boldsymbol{\theta} \in \Theta$ and assume that there exists $\phi : \mathbb{X} \longrightarrow [0,1]$, vanishing at infinity, such that $\phi(\emptyset) \neq 0$ and $-\phi \in \mathscr{F}$. Define $\varphi_t(\boldsymbol{\theta}) := \boldsymbol{P}^t_{\boldsymbol{\theta}}(\emptyset, \phi) = \mathbb{E}^{\emptyset}_{\boldsymbol{\theta}}[\phi(X_t)]$ for $t \geq 0$ and $\boldsymbol{\theta} \in \Theta$.*

Then the family of functions $\varphi_t : \Theta \longrightarrow [0,1]$ satisfies that

I. *$(t, \boldsymbol{\theta}) \longmapsto \varphi_t(\boldsymbol{\theta})$ is nonincreasing in both t and $\boldsymbol{\theta}$ (component-wise);*

II. *the limit $\varphi := \lim_{t \to \infty} \varphi_t$ is continuous and nonincreasing, satisfying*

$$\Theta_{\mathrm{s}} = \{\boldsymbol{\theta} \in \Theta : \varphi(\boldsymbol{\theta}) > 0\}. \tag{9}$$

In particular, $\Theta_{\mathrm{s}} \subseteq \Theta$ is open, star-shaped, and $\varphi(\boldsymbol{\theta}) = \langle \pi_{\boldsymbol{\theta}}, \phi \rangle$ for $\boldsymbol{\theta} \in \Theta_{\mathrm{s}}$.

III. *The mapping $\varphi = \langle \pi_{\cdot}, \phi \rangle : \Theta_{\mathrm{s}} \longrightarrow (0,1]$ is strictly decreasing.* ⋄

Theorem 2 provides a characterization of the stability region of an $\mathscr{F}$-monotone Q-process as the support of some continuous, nonincreasing functional φ. In particular, this shows that Θ_{s} is an open, star-shaped domain having the origin as a vantage point; that is, $\boldsymbol{\theta} \in \Theta_{\mathrm{s}}$ entails $c\boldsymbol{\theta} \in \Theta_{\mathrm{s}}$ for any $c \in (0,1]$. This property, also known in the literature as *star-convexity* (a weaker form of convexity) or *monotonicity* (Dai et al. 1999), turns out to be crucial property when setting up a simulation-based computation of the stability region (Leahu and Mandjes 2017).

Remark 5. Note that, in Theorem 2, φ is defined by means of some function ϕ, whereas the stability region is a fixed set (depending on the process itself but not on ϕ). The representation in (9) is valid for any functional φ (hence, function ϕ) satisfying the conditions of the theorem. In particular, $\mathscr{G}$-monotone Q-processes (e.g., satisfying the assumptions of Theorem 1) satisfy the conditions of Theorem 2; any $\phi = G \circ \ell$ with G nonincreasing and vanishing at infinity, for example, $G(x) = \exp(-\alpha x)$, $G(x) = (1+x)^{-\alpha}$, or $G(x) = \mathbf{1}\{x \leq \alpha\}$ $(\alpha > 0)$ can be used. ⋄

6. Concluding Remarks and Discussion

In this paper, we introduced a general Markovian model (a Q-process) for modeling the dynamics of a Markovian McQN over time. In addition, we introduced a new concept of stochastic ($\mathscr{F}$-)monotonicity. Then we proved that this property holds for a wide class of Q-processes, corresponding to virtually all McQN models that are relevant in applications. Indeed, it turns out that any McQN in which any server employs either a first-come, first-served static buffer priority or processor-sharing service discipline exhibits such monotonicity properties. Furthermore, we proved that this type of monotonicity is strong enough to ensure that stability is a monotonic property with respect to the external arrival rates.

The results of this paper facilitate the numerical methods developed in Leahu and Mandjes (2017) for determining the stability region of an McQN. To be more precise, let

$$\Theta = \{\boldsymbol{\theta} = r \cdot \vec{v} : r > 0\},$$

where $\vec{v} \geq \mathbf{0}$ denotes a d-dimensional vector, satisfying $\|\vec{v}\| = 1$; that is, Θ is a one-dimensional manifold (positive direction) in $\mathbb{R}^d$, endowed with the natural ordering. By Theorem 2, $\Theta_{\mathrm{s}} = (0, \boldsymbol{\theta}_*)$, where

$$\boldsymbol{\theta}_* = \sup\{\boldsymbol{\theta} \in \Theta : \varphi(\boldsymbol{\theta}) > 0\} = \min\{\boldsymbol{\theta} \in \Theta : \varphi(\boldsymbol{\theta}) = 0\};$$

the value $\boldsymbol{\theta}_*$ is called the *stability threshold along direction $\vec{v}$*. Furthermore, the mapping $\varphi : (\mathbf{0}, \boldsymbol{\theta}_*) \longrightarrow [0,1]$, $\varphi(\boldsymbol{\theta}) = \langle \pi_{\boldsymbol{\theta}}, \phi \rangle$, is strictly decreasing; hence, for any $\varepsilon \in (0,1)$ there exists some unique $\boldsymbol{\theta}_\varepsilon \in (\mathbf{0}, \boldsymbol{\theta}_*)$ satisfying $\varphi(\boldsymbol{\theta}_\varepsilon) = \varepsilon$; the value $\boldsymbol{\theta}_\varepsilon$ is called the *ε-congestion threshold along direction $\vec{v}$*. Finally, it is immediate that $\boldsymbol{\theta}_\varepsilon \uparrow \boldsymbol{\theta}_*$ as $\varepsilon \downarrow 0$.

Numerical (simulation-based) methods for evaluating congestion (and stability) thresholds were developed in Leahu and Mandjes (2017) under some minimal monotonicity assumptions, under which stochastic approximation schemes of Robbins–Monro type were applied to some specific McQN examples. The results in this paper formally validate the numerical results in Leahu and Mandjes (2017).

On the other hand, our analysis shows that concepts such as "stochastic monotonicity" carry over from Jackson networks to the more general multiclass framework, albeit in a weaker form. Here it is stressed that this weaker form still retains the most interesting monotonicity features. From a practical standpoint, the results in this paper cover, by and large, all relevant cases, but at the theoretical level, there are still some questions left. For instance, an interesting question would be whether there exist Q-processes (respectively, McQNs) that are *not* $\mathcal{G}$-monotone; one might expect that instances of nonmonotone Q-processes could possibly be obtained by violating the conditions of Theorem 1. Finally, we note that such monotonicity properties do not extend to non-Markovian McQN models (which involve nonexponential distributions) as shown in Whitt (1993).

Acknowledgments

The authors thank A. M. Oprescu, researcher at University of Amsterdam, for assisting with the implementation of the numerical experiments that led to establishing the results in this paper.

Appendix A. On Queue Policies and Service Allocations

In this appendix, we illustrate the formalism introduced in Section 3 by providing explicit details on how the most common service disciplines used in queueing applications fit into our modeling paradigm. In addition, we identify usual queue policies and service allocations satisfying the assumptions of Theorem 1 (hence, giving rise to monotone Q-processes) and also show how the model complexity can be reduced in some special cases.

Recall that $\mathbb{Q}[\mathcal{K}]$ denotes the space of finite sequences over the set $\mathcal{K}$ defined by (2). An insertion operator I_k on $\mathbb{Q}[\mathcal{K}]$ is a mapping such that for any p there exists a decomposition $p = (p', p'')$ such that $I_k(p) = (p', k, p'')$.

Perhaps the most natural queue policy (family of insertion operators) is FCFS. In our modeling paradigm, this is obtained as follows: for any $p \in \mathbb{Q}[\mathcal{K}]$, the insertion operator I_k inserts a k-digit at the end of the sequence, that is, $I_k(p) = (p, k)$, corresponding to the trivial decomposition $p' = p$ and $p'' = \emptyset$.

An important class of queue policies widely used in applications is the class of priority-based policies. To formalize that, we call a *priority ranking* on $\mathcal{K}$ a partition $\{\mathcal{C}_1, \ldots, \mathcal{C}_\nu\}$ of $\mathcal{K}$; this induces a natural (partial) ordering on $\mathcal{K}$, as follows: $k \prec l$ iff there exist $1 \le \imath < \jmath \le \nu$ such that $k \in \mathcal{C}_\imath$ and $l \in \mathcal{C}_\jmath$. The interpretation is that each partition block $\mathcal{C}_\imath$ represents a *caste* (subset of unranked classes) with representatives of higher castes (corresponding to smaller indexes) being allowed to overtake (in the queue) representatives of lower castes (larger indexes). When every $\mathcal{C}_\imath$ is a singleton, the priority ranking is called *total* as $\prec$ becomes a total ordering on $\mathcal{K}$; on the other hand, for $\nu = 1$ the ordering is trivial.

We call a *priority policy* a queue policy $\{I_k : k \in \mathcal{K}\}$ such that for any $p \in \mathbb{Q}[\mathcal{K}]$ there exists some priority ranking (set of castes) so that $I_k(p) = (p', k, p'')$, where p'' is the maximal tail sequence consisting of (consecutive) digits belonging to lower castes than k; that is, k will overtake all digits with lower priority ranking. Within each caste, a FCFS policy applies. When there is only one caste, it reduces to FCFS. Furthermore, if the priority ranking does not depend on p, we call it *static*; otherwise, we call it *dynamic*. Finally, if p'' may include the first digit of the sequence, then the policy is called *preemptive*; otherwise, we call it *nonpreemptive*.

On the other hand, recall that a service allocation is a state-dependent probability vector on $\mathcal{K}$, specifying what fraction of service capacity is assigned to any class. An important class of service allocations is the class of *head-of-the-queue* (HQ) allocations; that is, if we define $\kappa : \mathbb{Q}[\mathcal{K}] \longrightarrow \mathcal{K}$ as $\kappa(k_1, \ldots, k_n) := k_1$, then the HQ allocation is defined by $W_k(p) = \mathbf{1}\{\kappa(p) = k\}$. HQ allocations correspond to service disciplines in which only one job may receive service at a time. A service allocation that is not of HQ type is called a *bulk service* (BS) allocation. Among the non-HQ (BS) policies, the most important in applications are the so-called *processor sharing* (PS) service allocations, which only depend on p by means of its *composition vector* $\{p\}$, that is, the vector whose kth component $\{p\}_k$ denotes the number of k-digits in the sequence p; in particular, the ordering of the sequence is irrelevant. for a PS allocation W, there exists some $w : \mathbb{N}^{\mathcal{K}} \setminus \{\mathbf{0}\} \longrightarrow \mathcal{P}[\mathcal{K}]$ such that $W(p) = w(\boldsymbol{x})$, where $\boldsymbol{x} = \{p\}$. Usual choices are

- *Egalitarian* allocation, specified by the mapping

$$w_k(\boldsymbol{x}) = \frac{\mathbf{1}\{k \in \varsigma[\boldsymbol{x}]\}}{\#\varsigma[\boldsymbol{x}]},$$

that is, the server capacity is uniformly distributed among classes.

- *Proportional* allocation, specified by the mapping

$$w_k(\boldsymbol{x}) = \frac{x_k}{\|\boldsymbol{x}\|},$$

that is, the server capacity is distributed proportionally with the number of representatives in each class.

- *Preferential* allocation (assumes a total priority order on $\mathcal{K}$) given by

$$w_k(\boldsymbol{x}) = \mathbf{1}\{\kappa(\boldsymbol{x}) = k\},$$

where $\kappa(\boldsymbol{x})$ denotes the highest-ranked class in $\varsigma(\boldsymbol{x})$.

The usual FCFS and static buffer priority service disciplines are recovered in our model via the respective queue policies in combination with an HQ service allocation. One can easily verify that FCFS and static priority policies (of both preemptive and nonpreemptive types) satisfy condition (i) in Theorem 1, whereas dynamic priority policies do not, in general. Furthermore, processor-sharing disciplines are obtained via the corresponding PS allocation in combination with any queue policy (which is irrelevant). The PS allocations listed satisfy condition (ii) in Theorem 1. One concludes that, indeed, Markovian models associated with virtually all McQNs of interest in applications are $\mathscr{G}$-monotone; compare with Theorem 1.

In some situations of interest, it is possible to reduce the complexity of the full space $\mathbb{Q}[\mathscr{K}]$ (and of the full space $\mathbb{X}$, accordingly) by identifying equivalent configurations; this is formalized as follows: the space $(\mathbb{Q}[\mathscr{K}]; I_k, D_k, W_k : k \in \mathscr{K})$ of multiclass configurations over the set $\mathscr{K}$ is *reducible* if there exists an equivalence relation $\sim$ on $\mathbb{Q}[\mathscr{K}]$ such that $p \sim q$ entails $\{p\} = \{q\}$ and $W_k(p) = W_k(q)$, $I_k(p) \sim I_k(q)$ and $D_k(p) \sim D_k(q)$, for all $k \in \mathscr{K}$. One can further extend $\sim$ to the augmented space $\overline{\mathbb{Q}}[\mathscr{K}]$ by identifying the empty configuration with itself. The mappings I_k, D_k, W_k are then well defined on the quotient space $\overline{\mathbb{Q}}[\mathscr{K}]/\sim$, which is called a *reduced space* of multiclass configurations.

Instances of reducible spaces of multiclass configurations are given here:

1. If $\#\mathscr{K} = 1$ (single class), then $p \sim q$ iff $\{p\} = \{q\}$ gives $\overline{\mathbb{Q}}[\mathscr{K}]/\sim = \mathbb{N}$. Queue policies and service allocations are irrelevant.
2. For a static (total, nonpreemptive) priority policy with HQ allocation, let $p \sim q$ iff $\kappa(p) = \kappa(q)$ and $\{p\} = \{q\}$, which gives $\overline{\mathbb{Q}}[\mathscr{K}]/\sim = \mathscr{K} \times \mathbb{N}^{\mathscr{K}}$.
3. For a PS allocation, $p \sim q$ iff $\{p\} = \{q\}$; in this case, $\overline{\mathbb{Q}}[\mathscr{K}]/\sim = \mathbb{N}^{\mathscr{K}}$. The same factorization holds for static (total, preemptive) priority policies with HQ allocations, which can be recovered via a PS preferential allocation.

Appendix B. Proofs of the Results

In this appendix, we provide the proofs of the results presented in this paper.

Proof of Proposition 1. Let $\boldsymbol{A}_{\boldsymbol{\theta}}$ for $\boldsymbol{\theta} \geq \boldsymbol{0}$, denote the generator of the Q-process $\mathscr{X}$, defined in (4) with arrival-rate vector $\boldsymbol{\theta}$, that is, with parameter $(\boldsymbol{\theta}, \boldsymbol{\beta}, \boldsymbol{R})$.

I. By Kolmogorov's equation, the mapping of interest is differentiable with respect to t and, compare with (5), it holds that

$$\frac{d}{dt}\boldsymbol{P}^t_{\boldsymbol{\theta}}(\emptyset, \phi) = \frac{d}{dt}[\exp(t\boldsymbol{A}_{\boldsymbol{\theta}})\phi](\emptyset) = [\boldsymbol{A}_{\boldsymbol{\theta}} \exp(t\boldsymbol{A}_{\boldsymbol{\theta}})\phi](\emptyset). \tag{B.1}$$

Because $W_k(\emptyset) = 0$, it follows that

$$h_{(k,l)}(\emptyset) = \beta_k W_k(\emptyset) R_{kl} = 0$$

for any (k, l) with $k = 1, \ldots, d$, $l = 0, 1, \ldots, d$; hence, the r.h.s. in (B.1) equals

$$\sum_{k=1}^{d} \theta_k \big[(\Phi[f_{(0,k)}] - \boldsymbol{I}) \exp(t\boldsymbol{A}_{\boldsymbol{\theta}})\phi\big](\emptyset),$$

which is nonnegative for $t \geq 0$ by assumption; this concludes the first part.

II. Define for $\boldsymbol{0} \leq \boldsymbol{\theta} \leq \boldsymbol{\vartheta}$ and $0 \leq s \leq t$,

$$\boldsymbol{E}(s, t) := \exp[(t - s)\boldsymbol{A}_{\boldsymbol{\vartheta}}] \exp(s\boldsymbol{A}_{\boldsymbol{\theta}}).$$

Because $\boldsymbol{E}(0, t) = \exp(t\boldsymbol{A}_{\boldsymbol{\vartheta}})$ and $\boldsymbol{E}(t, t) = \exp(t\boldsymbol{A}_{\boldsymbol{\theta}})$, one obtains

$$\begin{aligned} \exp(t\boldsymbol{A}_{\boldsymbol{\vartheta}}) - \exp(t\boldsymbol{A}_{\boldsymbol{\theta}}) &= \int_0^t -\frac{d}{ds}\boldsymbol{E}(s, t) ds \\ &= \int_0^t \exp[(t - s)\boldsymbol{A}_{\boldsymbol{\vartheta}}](\boldsymbol{A}_{\boldsymbol{\vartheta}} - \boldsymbol{A}_{\boldsymbol{\theta}}) \exp(s\boldsymbol{A}_{\boldsymbol{\theta}}) ds, \end{aligned} \tag{B.2}$$

and the claim follows from the fact that (for $s \geq 0$)

$$(\boldsymbol{A}_{\boldsymbol{\vartheta}} - \boldsymbol{A}_{\boldsymbol{\theta}}) \exp(\boldsymbol{s}\boldsymbol{A}_{\boldsymbol{\theta}})\phi = \sum_{\boldsymbol{k}=\boldsymbol{1}}^{\boldsymbol{d}} (\vartheta_{\boldsymbol{k}} - \theta_{\boldsymbol{k}})(\Phi[\boldsymbol{f}_{(\boldsymbol{0},\boldsymbol{k})}] - \boldsymbol{I}) \exp(\boldsymbol{s}\boldsymbol{A}_{\boldsymbol{\theta}})\phi, \tag{B.3}$$

with the r.h.s. being nonnegative for any $\phi \in \mathscr{F}$ by assumption.

The case $\boldsymbol{\vartheta} \leq \boldsymbol{\theta}$ can be treated similarly. This concludes the proof. □

Proof of Theorem 1. We prove that $\boldsymbol{Q} = \boldsymbol{I} + (1/\lambda)\boldsymbol{A}$ with

$$\lambda = \sum_{k=1}^{d} (\theta_k + \beta_k),$$

satisfies Definition 3. To this end, let $\mathscr{E} := \{A_1, \ldots, A_d, B_1, \ldots, B_d\}$ and consider on $\mathscr{E}$ the probability distribution μ given by

$$\forall k = 1, \ldots, d : \mu(A_k) := \frac{\theta_k}{\lambda}; \quad \mu(B_k) := \frac{\beta_k}{\lambda}.$$

A random transition of the λ-chain (as defined in Section 3) from an arbitrary state $\xi \in \mathbb{X}$ can be constructed as follows: Generate a random variable J on $\mathscr{E}$, having distribution μ and

A. if $J = A_k$ (arrival event), then define $\Xi = f_{(0,k)}(\xi)$.

B. if $J = B_k$, then define $\Xi = f_{(k,l)}(\xi)$ with probability (w.p.) $W_k(\xi)R_{kl}$ $(l = 0, 1, \ldots, d)$ and let $\Xi = \xi$ w.p. $1 - W_k(\xi)$, independently of J.

For an arbitrary sample $\{\Xi_\nu : 0 \le \nu \le n\}$ of the λ-chain, generated according to (A)–(B) by means of a sequence of random variables $J_1, \ldots, J_n$, we consider the sequence of arrival events $\mathscr{A}_\nu[\Xi] = [k_1, \ldots, k_m]$ $(0 \le m \le \nu \le n)$ observed by the chain in the first ν steps; that is, one accounts for the arrival events within the sequence of J's and records the underlying classes. Note that the probability of observing a given sequence of arrival events equals

$$\Pr\{\mathscr{A}_\nu[\Xi] = [k_1, \ldots, k_m]\} = \binom{\nu}{m} \frac{\theta_{k_1} \cdot \ldots \cdot \theta_{k_m} \cdot \|\beta\|^{\nu - m}}{\lambda^\nu},$$

and does not depend on the initial state of the chain.

Returning to the proof of the theorem, we note that it suffices to construct a pair of λ-chains $\{(\Xi'_\nu, \Xi''_\nu) : 0 \le \nu \le n\}$ satisfying

$$(\Xi'_0, \Xi''_0) = (\xi, \zeta), \ell(\Xi'_n) \le \ell(\Xi''_n), \text{ a.s.}$$

Because the distribution of the sequence of arrival events does not depend on the initial state (hence, it is the same for both chains), it suffices to prove the statement conditioned on the event that the two chains share the same sequence of arrival events. More specifically, we prove that the following statement holds true for any $n \ge 0$:

$(\mathbf{H}_n)$: For any pair of initial configurations $\xi \subseteq \zeta$, there exists a pair of λ-chains $\{(\Xi'_\nu, \Xi''_\nu) : 0 \le \nu \le n\}$ satisfying

$$(\Xi'_0, \Xi''_0) = (\xi, \zeta), \mathscr{A}_n[\Xi'] = \mathscr{A}_n[\Xi''], \ell(\Xi'_n) \le \ell(\Xi''_n), \text{ a.s.}$$

Note that, in the statement, one may equivalently assume that either $\xi = \zeta$ or (ξ, ζ) is a pair of consecutive configurations. More concisely, one may assume that $(\xi, \zeta) \in \Delta_0 \cup \Delta_1 \cup \ldots \cup \Delta_d$, where, for convenience, we let $\Delta_0 := \{(\xi, \xi) : \xi \in \mathbb{X}\}$ denote the diagonal set of $\mathbb{X}$.

We prove this claim by induction. The key fact in this proof is that, given $\Xi'_0 \subseteq \Xi''_0$, there exists a coupling (Ξ'_1, Ξ''_1) satisfying $\Xi'_1 \subseteq \Xi''_1$ on the event $\{\Xi''_1 \neq \Xi''_0\}$; this is guaranteed by conditions (i) and (ii). Although the technical details of this proof are given, Figure B.1 displays the reasoning used for proving the induction step.

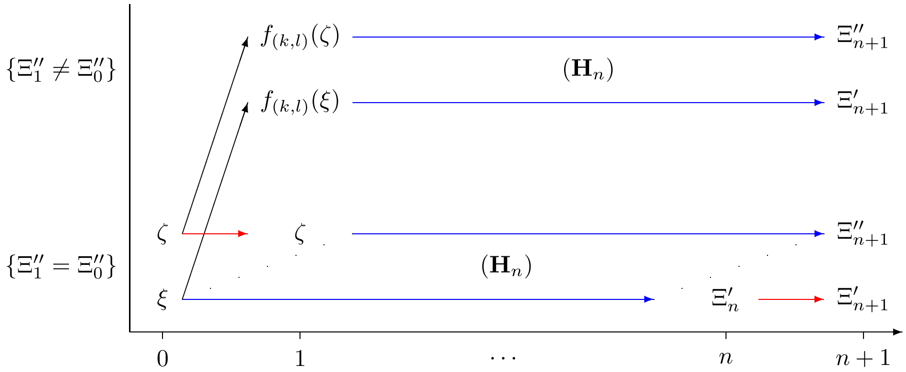
To begin with, note that the statement $(\mathbf{H}_0)$ is straightforward. Assume now that $(\mathbf{H}_n)$ holds true for some $n \ge 0$. Given $(\xi, \zeta) \in \Delta_b$ for $b = 0, 1, \ldots, d$, we construct a pair of λ-chains $\{(\Xi'_\nu, \Xi''_\nu) : 0 \le \nu \le n + 1\}$, satisfying

$$(\Xi'_0, \Xi''_0) = (\xi, \zeta), \mathscr{A}_{n+1}[\Xi'] = \mathscr{A}_{n+1}[\Xi''], \ell(\Xi'_{n+1}) \le \ell(\Xi''_{n+1}), \text{ a.s.} \tag{B.4}$$

For $b = 0$ (i.e., $\xi = \zeta$), the statement is trivial. Consider now the case $b \neq 0$. The key step in this construction is that, given $\Xi'_0 = \xi$ and $\Xi''_0 = \zeta$, there exists a coupling (Ξ'_1, Ξ''_1) satisfying (recall condition (ii)):

1. For $k = 1, \ldots, d$, we have $\Xi''_1 = f_{(0,k)}(\zeta)$ if and only if $\Xi'_1 = f_{(0,k)}(\xi)$.
2. For $k \neq b$, $\Xi''_1 = f_{(k,l)}(\zeta)$ entails $\Xi'_1 = f_{(k,l)}(\xi)$ for any $l = 0, 1, \ldots, d$.
3. $\Xi''_1 = f_{(b,l)}(\zeta)$ entails either $\Xi'_1 = f_{(b,l)}(\xi)$ or $\Xi'_1 = \xi$ for $l = 0, 1, \ldots, d$.

Figure B.1. Graphic Representation of the Proof (Induction Step) of Theorem 1



Note. Pairs of parallel blue arrows represent the coupled sample paths presumed by the induction hypothesis $(\mathbf{H}_n)$, whereas the red arrows represent one-step transitions conditioned to nonarrival events.

Therefore, one infers that, on the event $\{\Xi''_1 \neq \Xi''_0\}$, we have $(\Xi'_1, \Xi''_1) \in \Delta_b$ (cases 1 and 2) and $(\Xi'_1, \Xi''_1) \in \Delta_b$ or $(\Xi'_1, \Xi''_1) \in \Delta_l$ (in case 3); this follows by condition (i). On the other hand, given that $\{\Xi''_1 = \Xi''_0\}$, it readily follows that $(\Xi'_0, \Xi''_1) = (\xi, \zeta) \in \Delta_b$. As such, we distinguish the following cases:

I. If $\Xi''_1 \neq \Xi''_0$ (cases 1–3), then $(\Xi'_1, \Xi''_1) \in \Delta_0 \cup \Delta_1 \cup \ldots \cup \Delta_d$. By $(\mathbf{H}_n)$, there exists a pair $\{(\Psi'_\nu, \Psi''_\nu) : 0 \le \nu \le n\}$ such that

$$(\Psi'_0, \Psi''_0) = (\Xi'_1, \Xi''_1),\ \mathscr{A}_n[\Psi'] = \mathscr{A}_n[\Psi''],\ell(\Psi'_n) \le \ell(\Psi''_n),\ \text{a.s.}$$

Defining further $\Xi'_\nu := \Psi'_{\nu-1}$ and $\Xi''_\nu := \Psi''_{\nu-1}$ $(\nu = 1, \ldots, n+1)$, one can easily verify the validity of (B.4).

II. If $\Xi''_1 = \Xi''_0$, then $(\Xi'_0, \Xi''_1) = (\xi, \zeta) \in \Delta_b$ and using again $(\mathbf{H}_n)$, one obtains a pair of λ-chains $\{(\Psi'_\nu, \Psi''_\nu) : 0 \le \nu \le n\}$ such that

$$(\Psi'_0, \Psi''_0) = (\Xi'_0, \Xi''_1),\ \mathscr{A}_n[\Psi'] = \mathscr{A}_n[\Psi''],\ell(\Psi'_n) \le \ell(\Psi''_n),\ \text{a.s.}$$

Defining further $\Xi'_\nu := \Psi'_\nu$ and $\Xi''_\nu := \Psi''_{\nu-1}$ $(\nu = 1, \ldots, n)$, we note that $\mathscr{A}_n[\Xi'] = \mathscr{A}_{n+1}[\Xi'']$ (because $\mathscr{A}_1[\Xi''] = \emptyset$) and $\ell(\Xi'_n) \le \ell(\Xi''_{n+1})$. On the other hand, the constraint $\mathscr{A}_{n+1}[\Xi'] = \mathscr{A}_{n+1}[\Xi''] = \mathscr{A}_n[\Xi']$ implies that the $(n+1)$st transition of Ξ' does not correspond to an arrival event, whence

$$\ell(\Xi'_{n+1}) \le \ell(\Xi'_n) \le \ell(\Xi''_{n+1});$$

this proves (B.4).

Therefore, we proved $(\mathbf{H}_{n+1})$, which concludes the proof of the theorem. □

Remark B.1. Regarding the proof of Theorem 1, a few remarks are in order:

- The coupling $\{(\Xi'_\nu, \Xi''_\nu) : 0 \le \nu \le n\}$ has the special feature that it depends on the time horizon n.
- The chain Ξ'' is "forced" to perform the same transitions as Ξ' in the same order. Because the extra initial job Ξ'' may lag behind Ξ' in that the same transitions will be observed later or is even "dropped" by Ξ''. Because the two chains are bound to share the same sequence of arrival events, this might result in fewer (network) departures.

Proof of Theorem 2.

I. It follows directly by Proposition 1.

II. Monotonicity of $\varphi_t(\boldsymbol{\theta})$ with respect to t shows that the limit $\varphi := \inf_t \varphi_t$ exists and preserves monotonicity with respect to $\boldsymbol{\theta}$. Furthermore, we claim that

$$\varphi(\boldsymbol{\theta}) = \begin{cases} \langle \pi_{\boldsymbol{\theta}}, \phi \rangle, & \boldsymbol{\theta} \in \Theta\mathrm{s}; \\ 0, & \boldsymbol{\theta} \notin \Theta\mathrm{s}, \end{cases} \tag{B.5}$$

Indeed, because stability entails ergodicity in this context, we have

$$\forall \boldsymbol{\theta} \in \Theta_\mathrm{s} : \varphi(\boldsymbol{\theta}) = \lim_{t\to\infty} \mathbb{E}^{\emptyset}_{\boldsymbol{\theta}}[\phi(\mathrm{X}_t)] = \langle \pi_{\boldsymbol{\theta}}, \phi \rangle;$$

the r.h.s. is strictly positive because $\phi(\emptyset) > 0$, and $\pi_{\boldsymbol{\theta}}$ is supported on $\mathbb{X}_0$, which contains $\emptyset$. On the other hand, let $\boldsymbol{\theta} \notin \Theta_\mathrm{s}$. Then $\boldsymbol{P}^t_{\boldsymbol{\theta}}(\emptyset, \Omega) \longrightarrow 0$ for every compact (finite) set $\Omega \subset \mathbb{X}$; in particular, because ϕ is vanishing at infinity, there exists exhausting compacts $\{\Omega_n\}_{n\in\mathbb{N}}$ such that $\sup\{\phi(\xi) : \xi \notin \Omega_n\} \longrightarrow 0$, whence

$$\forall n \in \mathbb{N}, t \ge 0 : \varphi_t(\boldsymbol{\theta}) = \boldsymbol{P}^t_{\boldsymbol{\theta}}(\emptyset, \phi) \le \boldsymbol{P}^t_{\boldsymbol{\theta}}(\emptyset, \Omega_n) + \sup\{\phi(\xi) : \xi \notin \Omega_n\};$$

letting $t \to \infty$ yields $\varphi(\boldsymbol{\theta}) \le \sup_{\xi\notin\Omega_n} \phi(\xi) \longrightarrow 0$, which proves (B.5), hence, (9).

Finally, note that both expressions in the r.h.s. of (B.5) define continuous functions; hence, we only need to verify that φ is continuous at boundary points of Θ_s. To this end, let $\boldsymbol{\theta}_* \in \partial\Theta_\mathrm{s}$. Because Θ_s is open, we have $\partial\Theta_\mathrm{s} \subseteq \Theta_\mathrm{s}^\complement$; hence, $\varphi(\boldsymbol{\theta}_*) = 0$. On the other hand, $\varphi = \inf_t \varphi_t$ is upper semicontinuous, whence

$$0 \le \limsup_{\boldsymbol{\theta}\to\boldsymbol{\theta}_*} \varphi(\boldsymbol{\theta}) \le \varphi(\boldsymbol{\theta}_*) = 0,$$

that is, $\lim_{\boldsymbol{\theta}\to\boldsymbol{\theta}_*} \varphi(\boldsymbol{\theta}) = 0$. Therefore, φ is continuous at $\boldsymbol{\theta}_*$, which proves the claim.

III. Let now $\mathbf{0} \le \boldsymbol{\theta} < \boldsymbol{\vartheta}$ be such that $\boldsymbol{\vartheta} \in \Theta_\mathrm{s}$; in particular, we have $\boldsymbol{\theta} \in \Theta_\mathrm{s}$ (cf. II), and both $\pi_{\boldsymbol{\theta}}$ and $\pi_{\boldsymbol{\vartheta}}$ are supported on $\mathbb{X}_0$. First, one infers from (B.2), (B.3), and (6) that

$$\forall t \ge 0 : \left[\exp(t\boldsymbol{A}_{\boldsymbol{\vartheta}}) - \exp(t\boldsymbol{A}_{\boldsymbol{\theta}})\right]\boldsymbol{Q}_{\boldsymbol{\theta}}\phi \le \mathbf{0};$$

letting $t \to \infty$, yields $\langle \pi_{\boldsymbol{\vartheta}} - \pi_{\boldsymbol{\theta}}, \boldsymbol{Q}_{\boldsymbol{\theta}}\phi \rangle \le 0$. Furthermore, using the identity $\pi_{\boldsymbol{\theta}}\boldsymbol{A}_{\boldsymbol{\theta}} = \mathbf{0}$ (valid for all $\boldsymbol{\theta}$'s) one obtains

$$\begin{aligned} \langle \pi_{\boldsymbol{\vartheta}} - \pi_{\boldsymbol{\theta}}, \phi \rangle &= \langle \pi_{\boldsymbol{\vartheta}} - \pi_{\boldsymbol{\theta}}, (\boldsymbol{Q}_{\boldsymbol{\theta}} - (1/\lambda)\boldsymbol{A}_{\boldsymbol{\theta}})\phi \rangle \\ &\le -(1/\lambda) \cdot \langle \pi_{\boldsymbol{\vartheta}} - \pi_{\boldsymbol{\theta}}, \boldsymbol{A}_{\boldsymbol{\theta}}\phi \rangle = (1/\lambda) \cdot \langle \pi_{\boldsymbol{\vartheta}}, (\boldsymbol{A}_{\boldsymbol{\vartheta}} - \boldsymbol{A}_{\boldsymbol{\theta}})\phi \rangle \\ &= (1/\lambda) \sum_{k=1}^{d} (\vartheta_k - \theta_k) \cdot \langle \pi_{\boldsymbol{\vartheta}}, (\Phi[f_{(0,k)}] - \boldsymbol{I})\phi \rangle. \end{aligned}$$

Finally, we need to prove that the last expression in the display is strictly negative for $\boldsymbol{\theta} < \boldsymbol{\vartheta}$. To this end, let k be such that $\theta_k < \vartheta_k$. Because $\pi_{\boldsymbol{\vartheta}}$ is supported on $\mathbb{X}_0$, it is enough to show that $(\Phi[f_{(0,k)}] - \boldsymbol{I})\phi$ may not vanish everywhere on $\mathbb{X}_0$. Indeed, let $\xi_0 := \emptyset$ and $\xi_{n+1} := f_{(0,k)}(\xi_n)$ for $n \geq 0$; note that $\xi_n \in \mathbb{X}_0$ for all $n \geq 0$. Assuming that $(\Phi[f_{(0,k)}] - \boldsymbol{I})\phi$ vanishes on $\mathbb{X}_0$, it follows (by induction) that $\phi(\xi_n) = \phi(\emptyset) > 0$ (by assumption) for all $n \geq 0$; hence, ϕ is constant and strictly positive along the infinite sequence $\{\xi_n\}_{n\geq 0} \subseteq \mathbb{X}$. But because ϕ must vanish at infinity, this is a contradiction. This completes the proof of the theorem. □

Endnote

[1] $\phi : \mathbb{X} \longrightarrow \mathbb{R}$ is vanishing at infinity if there exists an increasing sequence of compacts $\{\Omega_n\}_{n\in\mathbb{N}}$ satisfying $\sup\{|\phi(\xi)| : \xi \notin \Omega_n\} \longrightarrow 0$ for $n \to \infty$.

References

Asmussen S, Glynn PW (2007) *Stochastic Simulation: Algorithms and Analysis* (Springer-Verlag, New York).

Bramson M (1994a) Instability of FIFO queueing networks. *Ann. Appl. Probab.* 4(2):414–431.

Bramson M (1994b) Instability of FIFO queueing networks with quick service times. *Ann. Appl. Probab.* 4(3):693–718.

Bramson M (2008) Stability of queueing networks. *Probab. Surveys* 5:169–345.

Dai JG (1995) On positive Harris recurrence of multiclass queueing networks: A unified approach via fluid limit models. *Ann. Appl. Probab.* 5(1): 49–77.

Dai JG, Hasenbein JJ, Vande Vate JH (1999) Stability of a three-station fluid network. *Queueing Systems* 33(4):293–325.

Dumas V (1997) A multiclass network with non-linear, non-convex, non-monotonic stability conditions. *Queueing Systems* 25(1–4):1–43.

Kelly FP (1975) Networks of queues with customers of different types. *J. Appl. Probab.* 12(3):542–554.

Kumar PR, Seidman TI (1990) Dynamic instabilities and stabilization methods in distributed real-time scheduling of manufacturing systems. *IEEE Trans. Automatic Control* 35(3):289–298.

Leahu H, Mandjes M (2017) A numerical approach to stability of multiclass queueing networks. *IEEE Trans. Automatic Control* 62(10):5478–5484.

Massey WA (1987) Stochastic orderings for Markov processes on partially ordered spaces. *Math. Oper. Res.* 12(2):350–367.

Meyn SP, Tweedie RL (1993a) *Markov Chains and Stochastic Stability* (Springer-Verlag, London).

Meyn SP, Tweedie RL (1993b) Stability of Markovian processes II: Continuous-time processes and sampled chains. *Adv. Appl. Probab.* 25(3): 487–517.

Rybko AN, Stolyar AL (1993) Ergodicity of stochastic processes describing the operation of open queueing networks. *Problemy Peredachi Informatsii* 28(3):3–26.

Seidman TI (1994) 'First come, first serve' can be unstable! *IEEE Trans. Automatic Control* 39(10):2166–2171.

Shanthikumar JG, Yao DD (1989) Stochastic monotonicity in general queueing networks. *J. Appl. Probab.* 26(2):413–417.

Vande Vate JH, Hasenbein JJ, Dai JG (2004) Stability and instability of a two-station queueing network. *Ann. Appl. Probab.* 14(1):326–377.

Whitt W (1993) Large fluctuations in a deterministic multiclass network of queues. *Management Sci.* 39(8):1020–1028.