Journal of the Operations Research
Society of Japan
2004, Vol. 47, No. 4, 275-303

# AN OVERVIEW OF SOME STOCHASTIC STABILITY METHODS*

Serguei Foss[†] Takis Konstantopoulos[‡]
*Heriott-Watt University*

(Received December 30, 2003)

*Abstract* This paper presents an overview of stochastic stability methods, mostly motivated by (but not limited to) stochastic network applications. We work with stochastic recursive sequences, and, in particular, Markov chains in a general Polish state space. We discuss, and frequently compare, methods based on (i) Lyapunov functions, (ii) fluid limits, (iii) explicit coupling (renovating events and Harris chains), and (iv) monotonicity. We also discuss existence of stationary solutions and instability methods. The paper focuses on methods and uses examples only to exemplify the theory. Proofs are given insofar as they contain some new, unpublished, elements, or are necessary for the logical reading of this exposition.

**Keywords**: Applied probability, stability, ergodicity, coupling, stationary processes, Markov chains, Harris chains, stochastic networks, stochastic recursive sequences, Lyapunov functions

## 1. Introduction

Our goal in this paper is to summarize some important methods used in deciding the stability of a stochastic system. We are principally concerned with those methods that are applicable to stochastic networks. However, many of them can be useful in stochastic dynamical systems, beyond the realm of queueing theory. These systems include, but are not limited to, iterated random functions, Markov chains, simulation algorithms, and stochastic nonlinear systems.

The area of stability, even in classical dynamical systems, has largely been driven by phenomena discovered in specific models (e.g., specific differential equations). Attempts to unify the observations have created stability "theories" which, in their turn, have found applications in systems other than the ones originally designed for. We observe the same trend in the area of stability in stochastic systems of applied probability. Queueing theory (stochastic networks) in particular, has led to important techniques which only now appear to be finding a solid place in the mathematics literature.

In writing this paper, we had to make several choices. Choosing to present methods rather than models was our first choice (compromise). Our second choice was to limit the presentation mostly to discrete-time systems, mainly to avoid technical complications, but also because our target applications can often be described (at least as regards their

*The writing of this paper benefited greatly from the wonderful research atmosphere provided by the Mathematisches Forschungsinstitut Oberwolfach, during an Applied Probability Workshop in December 2003 (organized by V. Schmidt, A. Hordijk and F. Baccelli), which gave the authors the opportunity to make substantial progress.

[†]This author's work was supported by EPSRC research grant GR/T27099/01.

[‡]This author's work was supported in part by NSF grant ANI-9903495 (University of Texas at Austin) and a Carathéodory research award (University of Patras, Greece).

question of stability) by *stochastic recursive sequences* (abbreviated SRS henceforth). An exception to this are the fluid approximation methods which can be stated in continuous time but proved by a discrete-time embedding. Also, we were forced to make injustice to many excellent research papers written in the field, because, due to space limitations, we could not include a complete or extensive bibliographical survey. So, our third choice was to deal with this problem by making the greatest possible injustice: we basically limit to research books and monographs, especially recent ones, and quote only the very few papers that contain parts of the proofs exposed here, and which we cannot fully complete due to space limitations. In a sense, we are referring to many research papers *through* the books and monographs cited here.

Some comments about the level/method of exposition: first of all, we give emphasis to the most general (in our opinion) methods and stress recent results. In particular, we survey, sometimes quite briefly, some research results of ours. Second, we present proofs only if they contain new (unpublished) elements in them or are necessary for the readability of the paper. Third, we present results that can be easily used and read without a full background in the area. Fourth, our discussion between theorems also contains material that should be considered as an integral part of our paper; however, we decided not to call them, formally, theorems, because they may contain ideas that need a lot of space for their rigorous proof/exposition.

Just as in classical deterministic dynamical systems (or, perhaps, worse yet) there are many notions of stability and instability. So the reader who will search for *our* definition of stability in this paper will not find it. This is on purpose: we use stability as idea, rather than as precise mathematical definition. For instance, at the level of Markov chains, stability may mean weak convergence to a stationary distribution, starting from a specific state, or starting from any state. It could also mean convergence over subsequences. But it could also mean "strong" convergence, due to coupling. It is true that, in concrete situations, the above may coincide, but there are important classes of systems for which the notions are and should be distinct. On the other hand, many stochastic systems of interest lack the Markovian property.[1] We give emphasis to SRS, i.e., to systems recursively defined by $X_{n+1} = f(X_n, \xi_n)$, where $\{\xi_n\}$ is a stationary (and often ergodic) sequence and $f$ a certain deterministic function. If the $\{\xi_n\}$ are i.i.d., then $\{X_n\}$ is Markov. On the other hand, any Markov chain in a Polish space (or, more generally, in a measurable space with a countably-generated $\sigma$-field) can be represented as a SRS; see Kifer [24]. To ask whether the SRS is stable, in its strongest form, is to ask whether its solution, starting, say, from a specific state converges to a stationary one; namely, there is a stationary process $\{\tilde{X}_n\}$ such that the entire future $(X_n, X_{n+1}, \ldots)$ after an index $n$, converges, in total variation, to $(\tilde{X}_0, \tilde{X}_1, \ldots)$; this stationary process satisfies the SRS. We may call this stationary process, the stationary version of $\{X_n\}$ but not always "the stationary solution of the SRS". A stationary solution of the SRS is a stationary process which satisfies the SRS. Clearly, such a stationary process may not be unique. If it is, then it can be called "the stationary solution of the SRS". As will be seen in Section 4 this distinction is oftentimes cast as the distinction between Harris theory and renovation theory: the former deals with the totality of the solutions of a Markovian SRS; the latter deals with a solution of an SRS.

Examples of SRS include: (i) The famous single-server queue, as represented by the Lindley recursion $W_{n+1} = (W_n + \sigma_n - \tau_n)^+$, where $\{\sigma_n\}$, $\{\tau_n\}$ are sequences of service, interarrival times, respectively; they should be assumed to be jointly stationary and ergodic;

[1] But even if they possess it, using it in order to prove stability may be a hindrance rather than an aid.

see [5]. (ii) A multiclass queueing network, whose recursion, of a rather daunting form, will not be written here; but see Section 3. (iii) A linear system $X_{n+1} = A_n X_n + B_n$, in $\mathbb{R}^d$, with $\{A_n, B_n\}$ being a stationary-ergodic sequence of matrices of appropriate dimensions. (iv) Iterates of a deterministic function: $X_{n+1} = f(X_n)$; even one-dimensional such functions are of interest to researchers in nonlinear systems (Lasota and Mackey [25]; Boyarski and Góra [10]); indeed, the existence of a non-trivial probability measure that is invariant under $f$ and stability of the above recursion describe the "statistical properties" of the deterministic system. (v) Iterated random functions–another word for a Markovian SRS when emphasis is given to choosing a function $f_\theta$, depending on a parameter $\theta$, by randomly selecting $\theta$ according to a probability measure $\mu$ living on a parameter space $\Theta$. This is the statistician's description of a Markovian SRS; see Diaconis and Freedman [16] for a survey of this area. (vi) SRS with *nonstationary* $\{\xi_n\}$ are of interest, e.g., in adaptive estimation and control. Stochastic approximations are typical applications. In such cases, a.s. convergence is also a notion of stability that makes sense; see, e.g., Duflo [17]; we shall not be concerned with these models in the present paper.

The first set of stability methods deals with Lyapunov functions (Section 2). In it, we present the proof of the most complete, to-date theorem, that deals with state-dependent drift criteria for a general state-space Markov chain. Section 3 deals with fluid approximation methods. Starting with a paper by Rybko and Stolyar [35], generalized in a paper of Dai [15], they have nowadays become a de facto "intuitive" set of criteria for stability. We show how their proof can be derived from the Lyapunov function methods with state-dependent drift and exemplify the technique in a multiclass queue and in Jackson-type networks. Section 4 deals with explicit coupling methods. These include renovation theory and Harris chains. Both subjects, as well as their relationships, are discussed. Section 5 deals entirely with the stationarity problem, i.e., that of existence of a stationary solution to a SRS. In fact, it focuses on a technique that is not based on regeneration. Section 6 studies monotone recursions, like the ones that were presented in what is thought of as the grandfather of stochastic stability for queues, i.e., the paper by Loynes [28]. Section 7 gives some recent results on instability methods based on Lyapunov functions. Finally, a few remarks on other methods are also presented. The reader should note that related questions, such as convergence rates and continuity, while natural consequences of some of the methods presented here, are not in fact studied in this paper.

## 2. Lyapunov Function Methods

Let $\{X_n\}$ be a time-homogeneous[2] Markov chain in some Polish space $\mathcal{X}$. Let $V : \mathcal{X} \to \mathbb{R}_+$ be a measurable function that is to be interpreted as a "norm", "Lyapunov" or "energy" function. We adopt the the standard notation $P_x(A) = P(A|X_0 = x)$, and $E_x Y$ stands for expectation with respect to $P_x$. The *drift* of $V$ under the Markov chain $X$ in $n$ time units is the function $x \mapsto E_x[V(X_n) - V(X_0)]$. Suppose that $g : \mathcal{X} \to \mathbb{N}$ is another measurable function that is to be interpreted as a state-dependent time. The drift of $V$ in $g(x)$ steps is the function

$$x \mapsto E_x[V(X_{g(x)}) - V(X_0)].$$

Finally, let $h : \mathcal{X} \to \mathbb{R}$ be a third measurable function such that $-h$ will provide an estimate on the size of the drift, in $g(x)$ steps. In order for the theorem below to be of any use at all, we must assume that $\sup_x V(x) = \infty$, and we shall do so everywhere in this paper. Assume that:

[2] Time-inhomogenous Markov chains can also be treated in a similar way, but more conditions are needed.

**(L1)** $h$ is bounded below: $\inf_{x\in\mathcal{X}} h(x) > -\infty$.

**(L2)** $h$ is eventually positive[3]: $\underline{\lim}_{V(x)\to\infty} h(x) > 0$.

**(L3)** $g$ is locally bounded above: $\sup_{V(x)\le N} g(x) < \infty, \quad$ for all $N > 0$.

**(L4)** $g$ is eventually bounded by $h$: $\overline{\lim}_{V(x)\to\infty} g(x)/h(x) < \infty$.

For a measurable set $B \subseteq \mathcal{X}$ define $\tau_B = \inf\{n > 0 : \ X_n \in B\}$ to be the first return time[4] to $B$. The set $B$ is called *recurrent* if $P_x(\tau_B < \infty) = 1$ for all $x \in B$. It is called *positive recurrent* if $\sup_{x\in B} E_x\tau_B < \infty$. It is this last property that is determined by a suitably designed Lyapunov function. This is the content of Theorem 1 below. That this property can be translated into a stability statement is the subject of later sections.

**Theorem 1.** *Suppose that the drift of $V$ in $g(x)$ steps satisfies the* "drift condition"

$$E_x[V(X_{g(x)}) - V(X_0)] \le -h(x),$$

where $V, g, h$ satisfy (L1)–(L4). Let

$$\tau \equiv \tau_N = \inf\{n > 0 : \ V(X_n) \le N\}.$$

Then there exists $N_0 > 0$, such that for all $N > N_0$ and any $x \in \mathcal{X}$, we have $E_x\tau < \infty$. Also, $\sup_{V(x)\le N} E_x\tau < \infty$.

*Proof.* We follow an idea that is essentially due to Tweedie [38]. From the drift condition, we obviously have that $V(x) - h(x) \ge 0$ for all $x$. We choose $N_0$ such that $\inf_{V(x)>N_0} h(x) > 0$ and $\sup_{V(x)>N_0} g(x)/h(x) < \infty$. Then, for $N \ge N_0$, $h(x)$ strictly positive, and we set

$$d = \sup_{V(x)>N} g(x)/h(x).$$

Then $0 < d < \infty$ as follows from (L2) and (L4). We also let

$$-H = \inf_{x\in\mathcal{X}} h(x),$$

and $H < \infty$, from (L1). We define an increasing sequence $t_n$ of stopping times recursively by

$$t_0 = 0, \quad t_n = t_{n-1} + g(X_{t_{n-1}}), \quad n \ge 1.$$

By the strong Markov property, the variables

$$Y_n = X_{t_n}$$

form a (possibly time-inhomogeneous) Markov chain with, as easily proved by induction on $n$, $E_xV(Y_{n+1}) \le E_xV(Y_n) + H$, and so $E_xV(Y_n) < \infty$ for all $n$ and $x$. Define the stopping time

$$\gamma = \inf\{n \ge 1 : \ V(Y_n) \le N\} \le \infty.$$

Observe that

$$\tau \le t_\gamma, \ \text{a.s.}$$

[3] The slightly unconventional notation used here is defined by $\underline{\lim}_{V(x)\to\infty} h(x) = \sup_{K>0} \inf_{x:V(x)>K} h(x)$. Also, $\overline{\lim}_{V(x)\to\infty} h(x) = -\underline{\lim}_{V(x)\to\infty} -h(x)$. The notation is used regardless of whether $V$ is a norm or not.

[4] This $\tau_B$ is a random variable. Were we working in continuous time, this would not, in general, be true, unless the paths of $X$ and the set $B$ were sufficiently "nice" (an instance of what technical complexities may arise in a continuous-time setup).

Let $\mathscr{F}_n$ be the sigma field generated by $Y_0, \ldots, Y_n$. Note that $\gamma$ is a predictable stopping time in that $\mathbf{1}(\gamma \geq i) \in \mathscr{F}_{i-1}$ for all $i$. We define the "cumulative energy" between 0 and $\gamma \wedge n$ by

$$\mathscr{E}_n = \sum_{i=0}^{\gamma\wedge n} V(Y_i) = \sum_{i=0}^{n} V(Y_i)\mathbf{1}(\gamma \geq i),$$

and estimate the change $E_x(\mathscr{E}_n - \mathscr{E}_0)$ (which is finite) in a "martingale fashion":[5]

$$\begin{aligned} E_x(\mathscr{E}_n - \mathscr{E}_0) &= E_x \sum_{i=1}^{n} E_x(V(Y_i)\mathbf{1}(\gamma \geq i) \mid \mathscr{F}_{i-1}) \\ &= E_x \sum_{i=1}^{n} \mathbf{1}(\gamma \geq i) E_x(V(Y_i) \mid \mathscr{F}_{i-1}) \\ &\leq E_x \sum_{i=1}^{n} \mathbf{1}(\gamma \geq i) E_x(V(Y_{i-1}) - h(Y_{i-1}) \mid \mathscr{F}_{i-1}) \\ &\leq E_x \sum_{i=1}^{n+1} \mathbf{1}(\gamma \geq i-1) E_x(V(Y_{i-1}) - h(Y_{i-1}) \mid \mathscr{F}_{i-1}) \\ &= E_x\mathscr{E}_n - E_x \sum_{i=0}^{n} h(Y_i)\mathbf{1}(\gamma \geq i), \end{aligned}$$

where we used that $V(x) - h(x) \geq 0$ and, for the last inequality, we also used $\mathbf{1}(\gamma \geq i) \leq \mathbf{1}(\gamma \geq i-1)$ and replaced $n$ by $n+1$. From this we obtain

$$E_x \sum_{i=0}^{n} h(Y_i)\mathbf{1}(\gamma \geq i) \leq E_x V(X_0) = V(x). \tag{1}$$

Assume $V(x) > N$. Then $V(Y_i) > N$ for $i < \gamma$, by the definition of $\gamma$, and so

$$h(Y_i) \geq d^{-1} g(Y_i) > 0, \quad \text{for } i < \gamma, \tag{2}$$

by the definition of $d$. Also,

$$h(Y_i) \geq -H, \quad \text{for all } i, \tag{3}$$

by the definition of $H$. Using (2) and (3) in (1) we obtain:

$$\begin{aligned} V(x) &\geq E_x \sum_{i=0}^{n} h(Y_i)\mathbf{1}(\gamma > i) + E_x \sum_{i=0}^{n} h(Y_i)\mathbf{1}(\gamma = i) \\ &\geq d^{-1} E_x \sum_{i=0}^{(\gamma-1)\wedge n} g(Y_i) - HP_x(\gamma \leq n), \end{aligned}$$

Recall that $g(Y_0) + \cdots + g(Y_k) = t_{k+1}$, and so the above gives:

$$V(x) \geq d^{-1} E_x t_{\gamma\wedge(n+1)} - HP_x(\gamma \leq n).$$

Now take limits as $n \to \infty$ (both relevant sequences are increasing in $n$) and obtain that

$$E_x t_\gamma \leq \frac{V(x) + H}{d}.$$

[5] albeit we do not make use of explicit martingale theorems

It remains to see what happens if $V(x) \le N$. By conditioning on $Y_1$, we have

$$\begin{aligned} E_x\tau &\le g(x) + E_x(E_{Y_1}\tau \mathbf{1}(V(Y_1 > N)) \\ &\le g(x) + E_x(d^{-1}(V(Y_1) + H)\mathbf{1}(V(Y_1) > N)) \\ &\le g(x) + d^{-1}H + d^{-1}(V(x) + H). \end{aligned}$$

Hence,

$$\sup_{V(x)\le N} E_x\tau \le \sup_{V(x)\le N} g(x) + d^{-1}(2H + N),$$

where the latter is a finite constant, by assumption (L3). □

**Discussion:** The theorem we just proved shows something quite strong about the set $B_N = \{x \in \mathcal{X} : \ V(x) \le N\}$. Namely, this set is *positive recurrent.* It is worth seeing that the theorem is a generalization of many more standard methods. When $g(x) = 1$ and $h(x) = \varepsilon - C_1\mathbf{1}(V(x) \le C_2)$, we have the classical Foster-Lyapunov criterion [22]. (This, in the particular case when $\mathcal{X} = \mathbb{Z}$ is often called Pakes' lemma [32].) Equivalently, the Foster-Lyapunov criterion seeks a function $V$ such that $E_x(V(X_1) - V(X_0)) \le -\varepsilon < 0$, when $V(x) > C_2$, and $\sup_{V(x)\le C_2} E_xV(X_1) < \infty$. When $g(x) = \lceil V(x)\rceil$ (where $\lceil t\rceil = \inf\{n \in \mathbb{N} : \ t \le n\}$, $t > 0$), and $h(x) = \varepsilon V(x) - C_1\mathbf{1}(V(x) \le C_2)$, we have Dai's criterion [15] which is the same as the "fluid limits" criterion. (See Section 3 for a further development of this.) Finally, when $h(x) = g(x) - C_1\mathbf{1}(V(x) \le C_2)$ we have the Meyn-Tweedie criterion [30]. The reader may also consult the monograph by Fayolle, Malyshev and Menshikov [18].

Observe that we never required the set $\{x : V(x) \le N\}$ to be "small" (in the sense that it is a finite or a compact set). Of course, if we are interested in stability, we must arrange that we prove that compact sets are positive recurrent, but the theorem above can be used to ensure that non-compact sets are positive recurrent. There are applications where one might, e.g., want to show that a half-space is positive recurrent.

**Remark:** condition (L4) is not only a technical condition. Its indispensability can be seen in the following simple example: Consider $\mathcal{X} = \mathbb{N}$, and transition probabilities

$$p_{1,1} = 1, \quad p_{k,k+1} \equiv p_k, \quad p_{k,1} = 1 - p_k \equiv q_k, \quad k = 2, 3, \ldots,$$

where $0 < p_k < 1$ for all $k \ge 2$ and $p_k \to 1$, as $k \to \infty$. Thus, jumps are either of size $+1$ or $-k$, till the first time state 1 is hit. Assume, for instance, $q_k = 1/k$, $k \ge 2$. Then, starting with $X_0 = 2$, we have $P(\tau = n) = 1/(n + 1)n$, $n = 1, 2, \ldots$. So $E\tau = \infty$. Therefore the Markov chain cannot be positive recurrent. Take now

$$V(k) = \log(1 \vee \log k), \quad g(k) = k^2.$$

We can estimate the drift and find

$$E_k[V(X_{g(k)}) - V(k)] \le -h(k), \tag{4}$$

where $h(k) = c_1V(k) - c_2$, and $c_1, c_2$ are positive constants. It is easily seen that (L1)-(L3) hold, but (L4) fails. This makes Theorem 1 inapplicable in spite of the negative drift (4). Physically, the time horizon $g(k)$ over which the drift was computed is far too large compared to the estimate $h(k)$ for the size of the drift itself.

## 3. Fluid Approximation Methods

In this section, we give essentially an application of Lyapunov methods to the so-called stability via fluid limits, a technique which became popular in the 90's. Roughly speaking, fluid approximation refers to a functional law of large numbers which can be formulated for large classes of Markovian and non-Markovian systems. Instead of trying to formulate the technique very generally, we focus on a quite important class of stochastic models, namely, multiclass networks. For statements and proofs of the functional approximation theorems used here, the reader may consult the texts of Chen and Yao [13], Whitt [39] and references therein.

### 3.1. Exemplifying the technique in a simple case

To exemplify the technique we start with a GI/GI/1 queue with general non-idling, work-conserving, non-preemptive service discipline[6], with arrival rate $\lambda$ and service rate $\mu$. Let $Q(t)$, $\chi(t)$, $\psi(t)$ be, respectively, the number of customers in the system, remaining service time of customer at the server (if any), and remaining interarrival time, at time $t$. The three quantities, together, form a Markov process (in continuous time). We will scale the whole process by

$$N = Q(0) + \chi(0) + \psi(0).$$

Although it is tempting, based on a functional law of large numbers (FLLN), to assert that $Q(Nt)/N$ has a limit, as $N \to \infty$, this is not quite right, unless we specify how the individual constituents of $N$ behave. So, we assume that[7]

$$Q(0) \sim c_1 N, \quad \chi(0) \sim c_2 N, \quad \psi(0) \sim c_3 N, \qquad \text{as } N \to \infty,$$

where $c_1, c_2, c_3 > 0$, with $c_1 + c_2 + c_3 = 1$. Then

$$\frac{Q(Nt)}{N} \to \overline{Q}(t), \qquad \text{as } N \to \infty,$$

uniformly on compact[8] sets of $t$, a.s., i.e.,

$$\lim_{N\to\infty} P(\sup_{0\le t\le T} |Q(kt)/k - \overline{Q}(t)| > \varepsilon, \text{ for some } k > N) = 0, \quad \text{for all } T, \varepsilon > 0.$$

The function $\overline{Q}$ is defined by:

$$\overline{Q}(t) = \begin{cases} c_1, & t < c_3 \\ c_1 + \lambda(t - c_3), & c_3 \le t < c_2 \;, \\ (c_1 + \lambda(c_2 - c_3) + (\lambda - \mu)(t - c_2))^+, & t \ge c_2 \end{cases} \qquad \text{if } c_3 \le c_2,$$

$$\overline{Q}(t) = \begin{cases} c_1, & t < c_2 \\ (c_1 - \mu(t - c_2))^+, & c_2 \le t < c_3 \;, \\ ((c_1 - \mu(c_3 - c_2))^+ + (\lambda - \mu)(t - c_3))^+, & t \ge c_3 \end{cases} \qquad \text{if } c_2 < c_3.$$

It is clear that $\overline{Q}(t)$ is the difference between two piecewise linear, increasing, functions. We shall not prove this statement here, because it is more than what we need: indeed, as will

[6] This means that when a customer arrives at the server with $\sigma$ units of work, then the server works with the customer without interruption, and it takes precisely $\sigma$ time units for the customer to leave.

[7] Hence, strictly speaking, we should introduce an extra index $N$ to denote this dependence, i.e., write $Q^{(N)}(t)$ in lieu of $Q(t)$, but, to save space, we shall not do so.

[8] We abbreviate this as "u.o.c."; it is the convergence also known as compact convergence.

be seen later, the full functional law of large numbers tells a more detailed story; all we need is the fact that there is a $t_0 > 0$ that does not depend on the $c_i$, so that $\overline{Q}(t) = 0$ for all $t > t_0$, provided we assume that $\lambda < \mu$. This can be checked directly from the formula for $\overline{Q}$. (On the other hand, if $\lambda > \mu$, then $\overline{Q}(t) \to \infty$, as $t \to \infty$.) To translate this FLLN into a Lyapunov function criterion, we use an embedding technique: we sample the process at the $n$-th arrival epoch $T_n$. (We take $T_0 = 0$.) It is clear that now we can omit the state component $\psi$, because

$$X_n := (Q_n, \chi_n) := (Q(T_n), \chi(T_n))$$

is a Markov chain with state space $\mathcal{X} = \mathbb{Z}_+ \times \mathbb{R}_+$. Using another FLLN for the random walk $T_n$, namely,

$$\frac{T_{[N\lambda t]}}{N} \to t, \qquad \text{as } N \to \infty, \quad \text{u.o.c.}, \quad \text{a.s.},$$

we obtain, using the usual method via the continuity of the composition mapping,

$$\frac{Q_{[N\lambda t]}}{N} \to (1 + (\lambda - \mu)t)^+, \qquad \text{as } N \to \infty, \quad \text{u.o.c.}, \quad \text{a.s.}.$$

Under the stability condition $\lambda < \mu$ and a uniform integrability (which shall be proved below–see the proof of Theorem 2) of $Q_{[N\lambda t]}/N, \chi_{[N\lambda t]}/N$, $N \in \mathbb{N}$, we have:

$$\frac{EQ_{[N\lambda t]}}{N} \to 0, \quad \frac{E\chi_{[N\lambda t]}}{N} \to 0, \qquad \text{as } N \to \infty, \qquad \text{for } t \geq t_0.$$

In particular there is $N_0$, so that $EQ_{[2N\lambda t_0]} + E\chi_{[2N\lambda t_0]} \leq N/2$ for all $N > N_0$. Also, the same uniform integrability condition, allows us to find a constant $C$ such that $EQ_{[2N\lambda t_0]} + E\chi_{[2N\lambda t_0]} \leq C$ for all $N \leq N_0$. To translate this into the language of a Lyapunov criterion, let $x = (q, \chi)$ denote a generic element of $\mathcal{X}$, and consider the functions

$$V(q,\chi) = q + \chi, \quad g(q,\chi) = 2(q+\chi)\lambda t_0, \quad h(q,\chi) = (1/2)(q+\chi) - C\mathbf{1}(q + \chi \leq N_0).$$

The last two inequalities can then be written as $E_x(V(X_{g(x)}) - V(X_0)) \leq -h(x)$, $x \in \mathcal{X}$. It is easy to see that the function $V, g, h$ satisfy (L1)-(L4). Thus Theorem 1 shows that the set $\{x \in \mathcal{X} : \; V(x) = q + \chi \leq N_0\}$ is positive recurrent.

### 3.2. Fluid limit stability criterion for multiclass queueing networks

We now pass on to multiclass queueing networks. Rybko and Stolyar [35] first applied the method to a two-station, two-class network. Dai [15] generalized the method and his paper established and popularized it.

Meanwhile, it became clear that the natural stability conditions [9] may not be sufficient for stability and several examples were devised to exemplify this phenomena; see, e.g., the paper by Bramson [11] which gives an example of a multiclass network which is unstable under the natural stability conditions, albeit operating under the "simplest" possible discipline (FIFO).

To describe a multiclass queueing network, we let $\{1, \ldots, K\}$ be a set of customer classes and $\{1, \ldots, J\}$ a set of stations. Each station $j$ is a single-server service facility that serves customers from the set of classes $c(j)$ according to a non-idling, work-conserving, non-preemptive, but otherwise general, service discipline. It is assumed that $c(j) \cap c(i) = \emptyset$ if

[9] By the term "natural stability conditions" in a work-conserving, non-idling, queueing network we refer to the condition that says that the rate at which work is brought into a node is less than the processing rate.

$i \neq j$. There is a single arrival stream[10], denoted by $A(t)$, which is the counting process of a renewal process, viz.,

$$A(t) = \mathbf{1}(\psi(0) \le t) + \sum_{n\ge1} \mathbf{1}(\psi(0) + T_n \le t),$$

where $T_n = \xi_1+\cdots+\xi_n$, $n \in \mathbb{N}$, and the $\{\xi_n\}$ are i.i.d. positive r.v.'s with $E\xi_1 = \lambda^{-1} \in (0,\infty)$. The interpretation is that $\psi(0)$ is the time required for customer 1 to enter the system, while $T_n$ is the arrival time of customer $n \in \mathbb{N}$. (Artificially, we may assume that there is a customer 0 at time 0.) To each customer class $k$ there corresponds a random variable $\sigma_k$ used as follows: when a customers from class $k$ is served, then its service time is an independent copy of $\sigma_k$. We let $\mu_k^{-1} = E\sigma_k$. Routing at the arrival point is done according to probabilities $p_k$, so that an arriving customer becomes of class $k$ with probability $p_k$. Routing in the network is done so that a customer finishing service from class $k$ joins class $\ell$ with probability $p_{k,\ell}$. Let $A_k(t)$ be the cumulative arrival process of class $k$ customers from the outside world. Let $D_k(t)$ be the cumulative departure process from class $k$. The process $D_k(t)$ counts the total number of departures from class $k$, both those that are recycled within the network and those who leave it. Of course, it is the specific service policies that will determine $D_k(t)$ for all $k$. If we introduce i.i.d. routing variables $\{\alpha_k(n), n \in \mathbb{N}\}$ so that $P(\alpha_k(n) = \ell) = p_{k\ell}$, then we may write the class-$k$ dynamics as:

$$Q_k(t) = Q_k(0) + A_k(t) + \sum_{\ell=1}^{K} \sum_{n=1}^{D_\ell(t)} \mathbf{1}(\alpha_\ell(n) = k) - D_k(t).$$

In addition, a number of other equations are satisfied by the system: Let $W^j(t)$ be the workload in station $j$. Let $C_{jk} = \mathbf{1}(k \in c(j))$. And let $V(n) = \sum_{m=1}^{n} \sigma_k(n)$ be the sum of the service times brought by the first $n$ class-$k$ customers. Then the total work brought by those customers up to time $t$ is $V_k(Q_k(0)+A_k(t))$, and part of it, namely $\sum_k C_{jk}V_k(Q_k(0)+A_k(t))$ is gone to station $j$. Hence the work present in station $j$ at time $t$ is

$$W^j(t) = \sum_k C_{jk}V_k(Q_k(0) + A_k(t)) - t + Y^j(t),$$

where $Y^j(t)$ is the idleness process, viz.,

$$\int W^j(t)dY^j(t) = 0.$$

The totality of the equations above can be thought of as having inputs (or "primitives") the $\{A_k(t)\}$, $\{\sigma_k(n)\}$ and $\{\alpha_k(n)\}$, and are to be "solved" for $\{Q_k(t)\}$ and $\{W^j(t)\}$. However, they are not enough: more equations are needed to describe how the server spends his service effort to various customers, i.e, we need policy-specific equations; see, e.g., [13].

Let $Q^j(t) = \sum_{k\in c(j)} Q_k(t)$. Let $\zeta_m^j(t)$ be the class of the $m$-th customer in the queue of station $j$ at time $t$, so that $\zeta^j(t) := (\zeta_1^j(t), \zeta_2^j(t), \ldots, \zeta_{Q^j(t)}^j(t))$ is an array detailing the classes of all the $Q^j(t)$ customers present in the queue of station $j$ at time $t$, where the leftmost one refers to the customer receiving service (if any) and the rest to the customers that are waiting in line. Let also $\chi^j(t)$ be the remaining service time of the customer receiving

[10] But do note that several authors consider many independent arrival streams

service. We refer to $X^j(t) = (Q^j(t), \zeta^j(t), \chi^j(t))$ as the state[11] of station $j$. Finally, let $\psi(t)$ be such that $t + \psi(t)$ is the time of the first exogenous customer arrival after $t$. Then the most detailed information that will result in a Markov process in continuous time is $X(t) := (X^1(t), \ldots, X^J(t); \psi(t))$. To be pedantic, we note that the state space of $X(t)$ is $\mathcal{X} = (\mathbb{Z}_+ \times K^* \times \mathbb{R}_+)^J \times \mathbb{R}_+$, where $K^* = \cup_{n=0}^{\infty}\{1, \ldots, K\}^n$, with $\{1, \ldots, K\}^0 = \{\emptyset\}$, i.e., $\mathcal{X}$ is a horribly looking creature–a Polish space nevertheless.

We now let

$$N = \sum_{j=1}^{J}(Q^j(0) + \chi^j(0)) + \psi(0),$$

and consider the system parametrized by this parameter $N$. While it is clear that $A(Nt)/N$ has a limit as $N \to \infty$, it is not clear at all that so do $D_k(Nt)/N$. The latter depends on the service policies, and, even if a limit exists, it may exist only along a certain subsequence. This was seen even in the very simple case of a single server queue.

To precise about the notion of limit point used in the following definition, we say that $\overline{X}(\cdot)$ is a limit point of $X_N(\cdot)$ if there exists a deterministic subsequence $\{N_\ell\}$, such that, $X_{N_\ell} \to \overline{X}$, as $\ell \to \infty$, u.o.c., a.s.

**Definition 1 (fluid limit and fluid model).** *A* fluid limit *is any limit point of the sequence of functions $\{D(Nt)/N, t \geq 0\}$. The* fluid model *is the set of these limit points.*

If $\overline{D}(t) = (\overline{D}_1(t), \ldots, \overline{D}_K(t))$ is a fluid limit, then we can define

$$\overline{Q}_k(t) = \overline{Q}_k(0) + \overline{A}_k(t) + \sum_{\ell=1}^{K}\overline{D}_\ell(t)p_{\ell,k} - \overline{D}_k(t), \quad k = 1, \ldots, K.$$

The interpretation is easy: Since $D(Nt)/t \to \overline{D}(t)$, along, possibly, a subsequence, then, along the same subsequence, $Q(Nt)/N \to \overline{Q}(t)$. This follows from the FLLN for the arrival process and for the switching process.

**Definition 2 (stability of fluid model).** *We say that the fluid model is* stable*, if there exists a deterministic $t_0 > 0$, such that, for all fluid limits, $\overline{Q}(t) = 0$ for $t \geq t_0$, a.s.*

To formulate a theorem, we consider the state process at the arrival epochs. So we let[12] $X_n := X(T_n)$. Then the last state component (the remaining arrival time) becomes redundant and will be omitted. Thus, $X_n = (X_n^1, \ldots, X_n^J)$, with $X_n^j = (Q_n^j, \zeta_n^j, \chi_n^j)$. Define the function

$$V : \big((q^j, \zeta^j, \chi^j), j = 1, \ldots, J\big) \mapsto \sum_{j=1}^{J}(q^j + \chi^j).$$

**Theorem 2.** *If the fluid model is stable, then there exists $N_0$ such that the set $B_{N_0} := \{x : V(x) \leq N_0\}$ is positive recurrent for $\{X_n\}$.*

**Remarks:**

(i) The definition of stability of a fluid model is quite a strong one. Nevertheless, if it holds – and it does in many important examples – then the original multiclass network is stable.

(ii) It is easy to see that the fluid model is stable in the sense of Definition 2 if and only

[11] Note that the first component is, strictly speaking, redundant as it can be read from the length of the array $\zeta^j(t)$.

[12] We tacitly follow this notational convention: replacing some $Y(t)$ by $Y_n$ refers to sampling at time $t = T_n$.

if there exist a deterministic time $t_0 > 0$ and a number $\varepsilon \in (0,1)$ such that, for all fluid limits, $\overline{Q}(t_0) \leq 1-\varepsilon$, a.s.

(iii) If all fluid limits are deterministic (non-random) – like in the examples below – then the conditions for stability of the fluid model either coincide with or are close to the conditions for positive recurrence of the underlying Markov chain $\{X_n\}$. However, if the fluid limits remain random, stability in the sense of Definition 2 is too restrictive, and the following weaker notion of stability may be of use:

**Definition 3 (weaker notion of stability of fluid model).** *The fluid model is (weakly) stable if there exist $t_0 > 0$ and $\varepsilon \in (0,1)$ such that, for all fluid limits, $E\overline{Q}(t_0) \leq 1-\varepsilon$.*

There exist examples of stable stochastic networks whose fluid limits are a.s. not stable in the sense of Definition 2, but stable in the sense of Definition 3 ("weakly stable"). The statement of Theorem 2 stays valid if one replaces the word "stable" by "weakly stable".
*Proof of Theorem 2.* Let

$$g(x) := 2\lambda t_0 V(x), \quad h(x) := \frac{1}{2}V(x) - C\mathbf{1}(V(x) \leq N_0),$$

where $V$ is as defined above, and $C$, $N_0$ are positive constants that will be chosen suitably later. It is clear that (L1)–(L4) hold. It remains to show that the drift criterion holds. Let $\overline{Q}$ be a fluid limit. Thus, $Q_k(Nt)/N \to \overline{Q}_k(t)$, along a subsequence. Hence, along the same subsequence, $Q_{k,[N\lambda t]}/N = Q_k(T_{[N\lambda t]})/N \to \overline{Q}_k(t)$. All limits will be taken along the subsequence referred to above and this shall not be denoted explicitly from now on. We assume that $\overline{Q}(t) = 0$ for $t \geq t_0$. So,

$$\overline{\lim_{N\to\infty}} \frac{1}{N}\sum_k Q_{k,[2\lambda t_0 N]} \leq 1/2, \quad \text{a.s.} \tag{5}$$

Also,

$$\lim_{n\to\infty} \frac{1}{n}\sum_j \chi_n^j = 0, \quad \text{a.s.} \tag{6}$$

To see the latter, observe that, for all $j$,

$$\frac{\chi_n^j}{n} \leq \frac{1}{n}\max_{k\in c(j)} \max_{1\leq i\leq D_{k,n}+1} \sigma_k(i) \leq \sum_{k\in c(j)} \frac{D_{k,n}+1}{n}\,\frac{\max_{1\leq i\leq D_{k,n}+1}\sigma_k(i)}{D_{k,n}+1}. \tag{7}$$

Note that

$$\frac{1}{m}\max_{1\leq i\leq m}\sigma_k(i) \to 0, \quad \text{as } m\to\infty, \quad \text{a.s.,}$$

and so

$$R_k := \sup_m \frac{1}{m}\max_{1\leq i\leq m}\sigma_k(i) < \infty, \quad \text{a.s.}$$

The assumption that the arrival rate is finite, implies that

$$\overline{\lim_{n\to\infty}} \frac{D_{k,n}+1}{n} < \infty, \text{ a.s.} \tag{8}$$

In case the quantity of (8) is zero then $\chi^j(n)/n \to 0$, because $R_k$ is finite. Otherwise, if it is positive, then, along any subsequence for which the limit of $D_{k,n}$ is infinity, we have that the limit of the last fraction of (7) is zero and so, again, $\chi^j(n)/n \to 0$. We next claim that

that the families $\{Q_{k,[2\lambda t_0 N]}/N\}$, $\{\chi^j_{[2\lambda t_0 N]}/N\}$ are uniformly integrable. Indeed, the first one is uniformly bounded by a constant:

$$\frac{1}{N} Q_{k,[2\lambda t_0 N]} \leq \frac{1}{N}(Q_{k,0} + A(T_{[2\lambda t_0 N]})) \leq 1 + [2\lambda t_0 N]/N \leq 1 + 4\lambda t_0.$$

To see that the second family is uniformly integrable, observe that, as in (7), and if we further loosen the inequality by replacing the maximum by a sum,

$$\frac{1}{N}\chi^j_{[2\lambda t_0 N]} \leq \sum_{k \in c(j)} \frac{1}{N} \sum_{i=1}^{D_{k,[2\lambda t_0 N]}+1} \sigma_k(i),$$

where the right-hand-side can be seen to be uniformly integrable by an argument similar to the one above. From (5) and (6) and the uniform integrability we have

$$\overline{\lim_{n\to\infty}} \frac{1}{N}\left(\sum_k EQ_{k,[2\lambda t_0 N]} + \sum_j E\chi^j_{[2\lambda t_0 N]}\right) \leq 1/2,$$

and so there is $N_0$, such that, for all $N > N_0$,

$$E\left(\sum_k Q_{k,[2\lambda t_0 N]} + \sum_j \chi^j_{[2\lambda t_0 N]} - N\right) \leq -N/2,$$

which, using the functions introduced earlier, and the usual Markovian notation, is written as

$$E_x[V(X_{g(x)}) - V(X_0)] \leq -\frac{1}{2}V(x), \quad \text{if } V(x) > N_0,$$

where the subscript $x$ denotes the starting state, for which we had set $N = V(x)$. In addition,

$$E_x[V(X_{g(x)}) - V(X_0)] \leq C, \quad \text{if } V(x) \leq N_0,$$

for some constant $C < \infty$. Thus, with $h(x) = V(x)/2 - C\mathbf{1}(V(x) \leq N_0)$, the last two displays combine into

$$E_x[V(X_{g(x)}) - V(X_0)] \leq -h(x).$$

□

In the sequel, we present two special, but important cases, where this assumption can be verified, under usual stability conditions.

### 3.3. Multiclass queue

In this system, a special case of a multiclass queueing network, there is only one station, and $K$ classes of customers. There is a single arrival stream $A$ with rate $\lambda$. Upon arrival, a customer becomes of class $k$ with probability $p_k$. Let $A_k$ be the arrival process of class-$k$ customers. Class $k$ customers have mean service time $\mu_k^{-1}$. Let $Q_k(t)$ be the number of customers of class $k$ in the system at time $t$, and let $\chi(t)$ be the remaining service time (and hence time till departure because service discipline is non-preemptive) of the customer in service at time $t$. We scale according to $N = \sum_k Q_k(0) + \chi(0)$. We do not consider the initial time till the next arrival, because we will apply the embedding method of the previous section. The traffic intensity is $\rho := \sum_k \lambda_k/\mu_k = \lambda \sum_k p_k/\mu_k$. Take any subsequence such that

$$Q_k(0)/N \to \overline{Q}_k(0), \quad \chi(0)/N \to \overline{\chi}(0), \text{ a.s.,}$$
$$A_k(Nt)/N \to \overline{A}_k(t) = \lambda_k t, \quad D_k(Nt)/N \to \overline{D}_k(t), \text{ u.o.c., a.s.}$$

That the first holds is a consequence of a FLLN. That the second holds is a consequence of Helly's extraction principle (c.f. Chung [14, pg. 83]). Then $Q(Nt)/N \to \overline{Q}(t)$, u.o.c., a.s., and so any fluid limit satisfies

$$\overline{Q}_k(t) = \overline{Q}_k(0) + \overline{A}_k(t) - \overline{D}_k(t), \quad k = 1, \ldots, K$$
$$\sum_k \overline{Q}_k(0) + \overline{\chi}(0) = 1.$$

In addition, we have the following structural property for any fluid limit: define

$$\overline{I}(t) := t - \sum_k \mu_k^{-1} \overline{D}_k(t), \quad \overline{W}_k(t) := \mu_k^{-1} \overline{Q}_k(t).$$

Then $\overline{I}$ is an increasing function, such that

$$\int_0^\infty \sum_k \overline{W}_k(t) d\overline{I}(t) = 0.$$

Hence, for any $t$ at which the derivative exists (which exists a.e., owing to Lipschitz continuity–see also the discussion in Section 3.4), and at which $\sum_k \overline{W}_k(t) > 0$,

$$\frac{d}{dt} \sum_k \overline{W}_k(t) = \frac{d}{dt}\Big( \sum_k \mu_k^{-1} \big(\overline{Q}_k(0) + \overline{A}_k(t)\big) - t \Big) - \frac{d}{dt}\overline{I}(t) = -(1-\rho).$$

Hence, if the stability condition $\rho < 1$ holds, then the above is strictly bounded below zero, and so, an easy argument shows that there is $t_0 > 0$, so that $\sum_k \overline{W}_k(t) = 0, \quad$ for all $t \geq t_0$. N.B. This $t_0$ is given by the formula $t_0 = C/(1-\rho)$ where $C = \max\{\sum_k \mu_k^{-1} q_k + \chi : \ q_k \geq 0, \ k = 1, \ldots, K, \chi \geq 0, \sum_k q_k + \chi = 1\}$. Thus, the fluid model is stable, Theorem 2 applies, and so we have positive recurrence.

### 3.4. Jackson-type network

Here we consider another special case, where there is a customer class per station. Traditionally, when service times are exponential, we are dealing with a classical Jackson network. This justifies our terminology "Jackson-type", albeit, in the literature, the term "generalized Jackson" is also encountered.

Let $\mathcal{J} := \{1, \ldots, J\}$ be the set of stations (= set of classes). There is a single arrival stream $A(t) = \mathbf{1}(\psi(0) \leq t) + \sum_{n \geq 1} \mathbf{1}(\psi(0) + T_n \leq t)$, $t \geq 0$, where $T_n = \xi_1 + \cdots + \xi_n$, $n \in \mathbb{N}$, and the $\{\xi_n\}$ are i.i.d. positive r.v.'s with $E\xi_1 = \lambda^{-1} \in (0, \infty)$. Upon arrival, a customer is routed to station $j$ with probability $p_{0,j}$, where $\sum_{j=1}^J p_{0,j} = 1$. To each station $j$ there corresponds a random variable $\sigma_j$ with mean $\mu_j$, i.i.d. copies of which are handed out as service times of customers in this station. We assume that the service discipline is non-idling, work-conserving, and non-preemptive. $\{X(t) = [(Q^j(t), \zeta^j(t), \chi^j(t), j \in \mathcal{J}); \psi(t)], t \geq 0\}$, as above.

The internal routing probabilities are denoted by $p_{j,i}$, $j, i \in \mathcal{J}$: upon completion of service at station $j$, a customer is routed to station $i$ with probability $p_{j,i}$ or exits the network with probability $1-\sum_{i=1}^{J} p_{j,i}$. We assume that the spectral radius of the matrix $[p_{j,i}]_{j,i\in\mathcal{J}}$ is strictly less than 1. We describe the (traditional) stability conditions in terms of an auxiliary Markov chain which we call $\{Y_n\}$ and which takes values in $\{0,1,\dots,J,J+1\}$, it has transition probabilities $p_{j,i}$, $j\in\{0,1,\dots,J\}$, $i\in\{1,\dots,J\}$, and $p_{j,J+1}=1-\sum_{i=1}^{J}p_{j,i}$, $j\in\{1,\dots,J\}$, $p_{J+1,J+1}=1$, i.e. $J+1$ is an absorbing state. We start with $Y_0=0$ and denote by $\pi(j)$ the mean number of visits to state $j\in\mathcal{J}$:

$$\pi(j)=E\sum_n \mathbf{1}(Y_n=j)=\sum_n P(Y_n=j).$$

Firstly we assume (and this is no loss of generality) that $\pi(j)>0$ for all $j\in\mathcal{J}$. Secondly, we assume that

$$\max_{j\in\mathcal{J}}\pi(j)\mu_j^{-1}<\lambda^{-1}.$$

Now scale according to $N=\sum_{j=1}^{J}[Q_j(0)+\chi_j(0)]$. Again, due to our embedding technique, we assume at the outset that $\psi(0)=0$. By applying the FLLN it is seen that any fluid limit satisfies

$$\overline{Q}_j(t)=\overline{Q}_j(0)+\overline{A}_j(t)+\sum_{i=1}^{J}\overline{D}_i(t)p_{i,j}-\overline{D}_j(t),\quad j\in\mathcal{J}$$
$$\sum_j[\overline{Q}_j(0)+\overline{\chi}_j(0)]=1,$$
$$\overline{A}_j(t)=\lambda_j t=\lambda p_{0,j}t,\quad \overline{D}_j(t)=\mu_j(t-\overline{I}_j(t)),$$

where $\overline{I}_j$ is an increasing function, representing cumulative idleness at station $j$, such that

$$\sum_{j=1}^{J}\int_0^\infty \overline{Q}_j(t)d\overline{I}_j(t)=0.$$

We next show that the fluid model is stable, i.e., that there exists a $t_0>0$ such that $\overline{Q}(t)=0$ for all $t\geq t_0$.

We base this on the following facts: If a function $g:\mathbb{R}\to\mathbb{R}^n$ is Lipschitz then it is a.e. differentiable. A point of differentiability of $g$ (in the sense that the derivative of all its coordinates exists) will be called "regular". Suppose then that $g$ is Lipschitz with $\sum_{i=1}^{n}g_i(0)=:|g(0)|>0$ and $\varepsilon>0$ such that ($t$ regular and $|g(t)|>0$) imply $|g(t)|'\leq-\varepsilon$; then $|g(t)|=0$ for all $t\geq|g(0)|/\varepsilon$. Finally, if $h:\mathbb{R}\to\mathbb{R}$ is a non-negative Lipschitz function and $t$ a regular point at which $h(t)=0$ then necessarily $h'(t)=0$.

We apply these to the Lipschitz function $\overline{Q}$. It is sufficient to show that for any $\mathcal{I}\subseteq\mathcal{J}$ there exists $\varepsilon=\varepsilon(\mathcal{I})>0$ such that, for any regular $t$ with $\min_{i\in\mathcal{I}}\overline{Q}_i(t)>0$ and $\max_{i\in\mathcal{J}-\mathcal{I}}\overline{Q}_i(t)=0$, we have $|\overline{Q}(t)|'\leq-\varepsilon$. Suppose first that $\mathcal{I}=\mathcal{J}$. That is, suppose $\overline{Q}_j(t)>0$ for all $j\in\mathcal{J}$, and $t$ a regular point. Then $\overline{Q}_j(t)'=\lambda_j+\sum_{i=1}^{J}\mu_ip_{i,j}-\mu_j$ and so $|\overline{Q}_j(t)|'=\lambda-\sum_{j=1}^{J}\sum_{i=1}^{J}\mu_ip_{i,j}-\sum_{j=1}^{J}\mu_j=\lambda-\sum_{i=1}^{J}\mu_ip_{i,J+1}=:-\varepsilon(\mathcal{J})$. But $\mu_i>\pi(i)\lambda$ and so $\varepsilon(\mathcal{J})>\lambda(1-\sum_{i=1}^{J}\pi(i)p_{i,J+1})=0$, where the last equality follows from $\sum_{i=1}^{J}\pi(i)p_{i,J+1}=\sum_{i=1}^{J}\sum_n P(Y_n=i,Y_{n+1}=J+1)=\sum_n P(Y_n\neq J+1,Y_{n+1}=J+1)=1$.

Next consider $\mathcal{I}\subset\mathcal{J}$. Consider an auxiliary Jackson-type network that is derived from the original one by $\sigma_j=0$ for all $j\in\mathcal{J}-\mathcal{I}$. It is then clear that this network has routing probabilities $p_{i,j}^{\mathcal{I}}$ that correspond to the Markov chain $\{Y_m^{\mathcal{I}}\}$, defined as a subsequence of

$\{Y_n\}$ at those epochs $n$ for which $Y_n \in \mathcal{I} \cup \{J+1\}$. Let $\pi^{\mathcal{I}}(i)$ the mean number of visits to state $i \in \mathcal{I}$ by this embedded chain. Clearly, $\pi^{\mathcal{I}}(i) = \pi(i)$, for all $i \in \mathcal{I}$. So the stability condition $\max_{i \in \mathcal{I}} \pi(i)\mu_i < \lambda^{-1}$ is a trivial consequence of the stability condition for the original network. Also, the fluid model for the auxiliary network is easily derived from that of the original one. Assume then $t$ is a regular point with $\min_{i \in \mathcal{I}} \overline{Q}_i(t) > 0$ and $\max_{i \in \mathcal{J}-\mathcal{I}} \overline{Q}_i(t) = 0$. Then $|Q_j(t)|' = 0$ for all $j \in \mathcal{J} - \mathcal{I}$. By interpreting this as a statement about the fluid model of the auxiliary network, in other words that all queues of the fluid model of the auxiliary network are positive at time $t$, we have, precisely as in the previous paragraph, that $\overline{Q}_j(t)' = \lambda p^{\mathcal{I}}_{0,j} + \sum_{i \in I} \mu_i p^{\mathcal{I}}_{i,j} - \mu_j$, for all $j \in \mathcal{I}$, and so $|\overline{Q}(t)|' = \lambda - \sum_{i \in \mathcal{I}} \mu_i p^{\mathcal{I}}_{i,J+1} =: -\varepsilon(\mathcal{I})$. As before, $\varepsilon(\mathcal{I}) > \lambda(1 - \sum_{i \in \mathcal{I}} \pi(i) p^{\mathcal{I}}_{i,J+1}) = 0$.

We have thus proved that, with $\varepsilon := \min_{\mathcal{I} \subseteq \mathcal{J}} \varepsilon(\mathcal{I})$, for any regular point $t$, if $|\overline{Q}(t)|' > 0$, then $|\overline{Q}(t)| \leq -\varepsilon$. Hence the fluid model is stable.

**Remark:** We also refer to the recent monographs by Chen and Yao [13, Ch. 8] and Robert [34, Ch. 9] regarding the fluid limit technique.

## 4. Methods Based on Explicit Coupling

When we pass from the Markovian to the non-Markovian world, criteria for stability change in nature. This should come as no surprise, for a SRS, albeit a dynamical system on the space of paths, loses its semigroup property on the space of probability measures. In the first part of this section, we discuss the method of renovating events, due to Borovkov, which finds applications in several non-Markovian systems. The method is intimately related to the so-called Harris theory, which is presented in the second part of the section.

### 4.1. Coupling and renovating events

We will focus on the SRS

$$X_{n+1} = f(X_n, \xi_n),$$

where $\{\xi_n, \in \mathbb{Z}\}$ is a stationary-ergodic sequence of random variables taking values in some measurable space $(\mathcal{Y}, \mathscr{B}_{\mathcal{Y}})$. The variables $X_n$ take values in another measurable space $(\mathcal{X}, \mathscr{B}_{\mathcal{X}})$ and $f : \mathcal{X} \times \mathcal{Y} \to \mathcal{X}$ is measurable. We refer everything to a probability space $(\Omega, \mathscr{F}, P)$ with a $P$-preserving ergodic flow $\theta$ such that $\xi_n \circ \theta = \xi_{n+1}$. (This can always be done by, say, using the space of paths of $\xi_\cdot$ as $\Omega$-see [5].) The process $\{X_n\}$ may be defined from some index $n$ onwards, or for all $n \in \mathbb{Z}$. Our goal is to find conditions for which a stationary-ergodic $\{X_n\}$ exists and is unique.

A key role in this are played by various notions of coupling. We shall need three of them. Below, $\{X_n\}$, $\{Y_n\}$ are sequences of random variables. A notation that is used is

$$X_n^{-m} := X_{m+n} \circ \theta^{-m}.$$

To understand the meaning of this notation, it is best to think in terms of an SRS: suppose we start with $X_0 = x_0$, some fixed value, and solve the SRS forward. Then $X_n$ is a certain function $X_n = f^{(n)}(x_0, \xi_0, \ldots, \xi_{n-1})$. Using the stationarity of $\{\xi_n\}$, we have $X_n^{-m} = f^{(m+n)}(x_0, \xi_{-m}, \ldots, \xi_0, \ldots, \xi_{n-1})$, i.e., we start the recursion at time $-m \leq 0$ and solve till we reach $n \geq 0$. The notation $X_n^{-m}$ reflects this, but need not refer to this specific situation always. We now give definitions for three notions of coupling.

**Definition 4.**

Simple coupling: *Let $\nu = \inf\{n \geq 0 : X_k = Y_k$ for all $k \geq n\}$. We say that $X$ couples with $Y$ if $\nu < \infty$ a.s. This $\nu$ is the minimal coupling time.*

Strong coupling (or strong forward coupling): *Let $\sigma(m) = \inf\{n \geq 0 : X_k^{-m} = Y_k$ for all $k \geq n\}$ and $\sigma = \sup_{m \geq 0} \sigma(m)$. We say that $X$ strongly couples with $Y$ if $\sigma < \infty$ a.s. This $\sigma$ is the minimal strong coupling time.*

Backward coupling: *Let $\tau(m) = \inf\{n \geq 0 : X_m^{-n} = X_m^{-(n+k)}$ for all $k \geq 0\}$ and $\tau = \sup_{m \geq 0} \tau(m)$. We say that $X$ backward-couples if $\tau < \infty$ a.s. This $\tau$ is the minimal backward coupling time.*

**Remarks:**
(i) The last definition is intrinsic to $X$: it requires no additional sequence.
(ii) If $\nu < \infty$, a.s., then any random time $\nu'$, with $\nu' \geq \nu$ is a coupling time. Likewise, any $\sigma' > \sigma$ is a strong coupling time, and any $\tau' > \tau$ is a backward coupling time.
(iii) Note also that the definitions can be simplified in case $X$ and $Y$ obey the same SRS, because we can remove the quantifier "for all" in each of the definitions above.
(iv) If $Y$ is stationary, then simple coupling of $X$ with $Y$ implies that $X$ converges in total variation in the sense that

$$\lim_{n \to \infty} \sup_{B \in \mathscr{B}_{\mathcal{X}}^{\infty}} |P((X_n, X_{n+1}, \ldots) \in B) - P((Y_n, Y_{n+1}, \ldots) \in B)| = 0.$$

(v) Strong coupling implies simple coupling.
(vi) Suppose $Y$ is stationary. Then strong coupling is equivalent to backward coupling: in fact, $\tau$, has the same distribution as $\sigma$, and $Y$ itself can be constructed from $X$, since backward coupling is, after all, an intrinsic criterion.
(vii) If there is backward coupling, then we set

$$\tilde{X}_0 = X_\tau \circ \theta^{-\tau} = X_0^{-\tau},$$
$$\tilde{X}_n = \tilde{X}_0 \circ \theta^n, \quad n \in \mathbb{Z}.$$

The latter is the stationary sequence with which $X$ strongly couples.

We now formulate a criterion for strong coupling. We say that $A$ is a *renovating event* for the SRS $X$ at time $n$ if, at time $n+1$, there is a "decoupling" of $X_{n+1}$ from its past, namely, if there is a deterministic function $g : \mathcal{Y} \to \mathbb{R}$ such that $X_{n+1} = g(\xi_n)$ a.s. on the event $A$. Roughly speaking, such a decoupling, if it happens frequently enough, ensures that coupling will take place, which in turn ensures existence of a unique stationary solution. In fact, this decoupling may not take place at time $n$ but at some future time $n + m + 1$, so it is necessary to bring this into the definition as well. More precisely:

**Definition 5.** *We say that $A$ is an $\langle n, m, g \rangle$ renovating event for the SRS $X$ if $X_{n+m+1} = g(\xi_n, \ldots, \xi_{n+m})$ a.s. on $A$.*

If a sequence $\{A_n\}$ of positive[13] stationary renovating events exists then we have strong coupling. More precisely:

**Theorem 3.** *Let $X$ be SRS with stationary-ergodic driver. If there is $n_0$ such that for each $n \geq n_0$ there is a $\langle n, m, g \rangle$ renovating event $A_n$, and $\{A_n\}$ is stationary-ergodic with $P(A_n) > 0$ then there is backward coupling.*

This theorem is not the most general one in the area. First, there is a theorem that ensures backward coupling when the sequence $\{A_n\}$ is only asymptotically stationary; for details, see Borovkov [8]. Second, there are processes that are not necessarily SRS. They

[13] i.e., $P(A_n) > 0$

could, for instance be functions of the process $X$. It was proved in [20] that a renovating events criterion does exist in this case as well (extended renovation theory).

Here is a sketch of the proof of Theorem 3, based on [8, 9, 20]. Define the time

$$\gamma = \inf\{n \geq 0 : \ \mathbf{1}_{A_{-n}} = 1\}.$$

Observe first that $\gamma$ is a.s. finite due to ergodicity. Observe next that the probability that the family of processes $\{X_{\cdot}^{-k}, k \geq n\}$ all couple after some deterministic time is not smaller than $P(\gamma \leq n)$. Finally, let $n \to \infty$ to conclude that backward coupling, as defined in Def. 4 does indeed take place.

A stationary process $\{\tilde{X}_n\}$ can then be constructed as follows:

$$\tilde{X}_n := X_n^{-\gamma} = X_{\gamma+n} \circ \theta^{-\gamma}, \quad n \geq 0. \tag{9}$$

It can then be checked that this $\tilde{X}$ satisfies the same SRS, and that there is strong coupling between $X$ and $\tilde{X}$.

An interesting example here is a system comprising of an infinite number of queues in tandem. It can be proved that there is stability in the sense that each finite collection of queues strongly couples with their stationary versions. In this example, and frequently more generally, the assumption of stationarity of $\{\xi_n, n \in \mathbb{Z}\}$ may be replaced by coupling stationarity, i.e., with the condition that the driving sequence strongly couples with a stationary one.

It is interesting to note that equation (9) can be the cornerstone for modern perfect simulation algorithms. See, e.g., [20].

## 4.2. Harris chains

Specializing to the Markovian case, consider a Markov chain in a Polish space $\mathcal{X}$. Let

$$P^n(x, \cdot) := P_x(X_n \in \cdot) = P(X_{m+n} \in \cdot \mid X_m = x).$$

We assume that there is a recurrent set $R$, i.e.,

$$\tau_R < \infty, \quad P_x - \text{a.s., for all } x \in \mathcal{X},$$

and an integer $\ell$ such that the family of probability measures $\{P_n^{\ell}(x, \cdot), x \in R\}$ have a common component $Q$:[14]

$$P^{\ell}(x, \cdot) \geq pQ(\cdot), \qquad \text{for all } x \in R,$$

where $Q$ is a probability measure, and $0 < p < 1$. We then say that the chain possesses the Harris property, or that it is *Harris recurrent*, or simply a *Harris chain.*[15] The set $R$

[14] A family $\{P_\alpha\}$ of probability measures possesses a common component, if there is a finite measure $\mu$ such that $\mu \leq \inf_\alpha P_a$. Then it is possible to define, on a common probability space, random variables $\{X_{\alpha,n}, \ n \in \mathbb{N}\}$ such that for each $n$, $X_{\alpha,n}$ has law $P_\alpha$, and such that $\tau := \inf\{n : \ X_{\alpha,n} = X_{\beta,n} \text{ for all } \alpha, \beta\} < \infty$, a.s. This is done by first writing $P_\alpha = pQ + (1-p)\overline{P}_\alpha$, where $p = ||\mu||$, $Q = \mu/||\mu||$, and $\overline{P}_\alpha$ another probability measure defined by the same relation, and then considering independent sequences of random variables: $\{\zeta_n, \ n \in \mathbb{N}\}$, $\{Y_n, \ n \in \mathbb{N}\}$, $\{\overline{Y}_{\alpha,n}, \ n \in \mathbb{N}\}$, such that the $\{\zeta_n, \ n \in \mathbb{N}\}$ are i.i.d. with $P(\zeta_n = 1) = 1 - P(\zeta_n = 0) = p$, the $\{Y_n, \ n \in \mathbb{N}\}$ are i.i.d. with $P(Y_n \in \cdot) = Q$, the $\{\overline{Y}_{\alpha,n}, \ n \in \mathbb{N}\}$ are, for each $\alpha$, i.i.d. with $P(\overline{Y}_{\alpha,n} \in \cdot) = \overline{P}$. Based on these, the $X_{\alpha,n}$ are explicitly defined by $X_{\alpha,n} = \zeta_n Y_n + (1 - \zeta_n)\overline{Y}_{\alpha,n}$. It is clear that $P(X_{\alpha,n} \in \cdot) = P_\alpha$ and that $\tau \leq \inf\{n : \ \zeta_n = 1\}$ which is obviously a.s. finite.

[15] Note that there is no universal "standard" definition of a Harris chain; different authors may use different definitions.

is often called a *regeneration set*. The discussion that follows justifies the terminology and shows that a Harris chain always possesses an invariant measure (which may possibly have infinite mass).

Suppose $X_0 = x \in R$. Write $X_\ell^x$ for the Markov chain at time $\ell$. As described in the footnote 14, we may realize the family of random variables $\{X_\ell^x, x \in R\}$ in a way that $P(X_\ell^x = X_\ell^y$ for all $x, y \in R) > 0$. This is done by generating a *single* random variable, say $Y$, with law $Q$, and by tossing a coin with probability of success $p$. If successful, we let $X_\ell^x = Y$, for all $x \in R$. If not, we distribute according to the remaining probability.

We now show that each Harris chain has an invariant measure. This is basically done by an inverse Palm construction. Start the chain according to the law $Q$. Let $T_{R,k}$, $k \in \mathbb{N}$ be the times at which the chain hits $R$. For each such $T_{R,k}$ consider a 0/1 r.v. $\zeta_k$ with $P(\zeta_k = 1) = p$ and a r.v. $Y_k$ with $P(Y_k \in \cdot) = Q$. Let $K = \inf\{k : \ \zeta_k = 1\}$. Consider the path $\{X_n, 0 \le n \le T_{R,K}\}$. Forcefully set $X_{T_{R,K}+\ell} = Y_K$. Realize the path $\{X_n, T_{R,K} < n < T_{R,K} + \ell\}$ conditionally on the values $X_{T_{R,K}}$ and $X_{T_{R,K}+\ell}$. The path $\mathscr{C}_0 := \{X_n, 0 \le n \le T_{R,K} + \ell\}$ is the first cycle of the chain. Considering the iterates of $T_{R,K}$, namely, $T_{R,K}^{(m+1)} = T_{R,K}^{(m)} + T_{R,K} \circ \theta^{T_{R,K}^{(m)}}$, $m \ge 0$, $T_{R,K}^{(0)} \equiv 0$, we obtain the successive cycles $\mathscr{C}_{m+1} = \mathscr{C}_m \circ \theta^{T_{R,K}^{(m)}}$. It is clear that $X_{T_{R,K}+\ell}$ has law $Q$ and that the sequence of cycles is stationary. (This is referred to as Palm stationarity.) Moreover, we have a regenerative structure: $\{T_{R,K}^{(m)}, 0 \le m \le n\}$ is independent of $\{\mathscr{C}_m, m \ge n+1\}$, for all $n$. The only problem is that the cycle durations may be random variables with infinite mean.

Now let $P_Q$ be the law of the chain when $X_0$ is chosen according to $Q$. Define the measure

$$\mu(\cdot) = E_Q \sum_{n=1}^{T_{R,K}} \mathbf{1}(X_n \in \cdot).$$

Strong Markov property ensures that $\mu$ is stationary, i.e.,

$$\mu(\cdot) = \int_{\mathcal{X}} P(x, \cdot) \mu(dx).$$

If, in addition, to the Harris property, we also have

$$E_Q(T_{R,K}) < \infty, \tag{10}$$

we then say that the chain is *positive Harris recurrent*. In such a case, $\pi(\cdot) = \mu(\cdot)/E_Q(T_{R,K})$ defines a stationary probability measure. Moreover, the assumption that $R$ is recurrent ensures that there is no other stationary probability measure.

A sufficient condition for positive Harris recurrence is that

$$\sup_{x \in R} E_x \tau_R < \infty. \tag{11}$$

This is a condition that does not depend on the (usually unknown) measure $Q$. To see the sufficiency, just use the fact that $T_{R,K}$ may be represented as a geometric sum of r.v.'s with uniformly bounded means, so that (11) implies (10). To check (11) the Lyapunov function methods of Section 2 are very useful. Let us also offer a further remark on the relation between (11) and (10): it can be shown that if (10) holds, then there is a $R' \subseteq R$ such that

$$\sup_{x \in R'} E_x \tau_{R'} < \infty.$$

We next give a brief description of the stability by means of coupling, achieved by a positive Harris recurrent chain. To avoid periodicity phenomena, we assume that the discrete random variable $T_{R,K}$ has aperiodic distribution under $P_Q$. (A sufficient condition for this is: $P_Q(T_{R,K} = 1) > 0$.) Then we can construct a successful coupling between the chain starting from an arbitrary $X_0 = x_0$ and its stationary version. Assume that $E_{x_0}\tau_R < \infty$. (This $x_0$ may or may not be an element of $R$.) Let $\{X_n\}$ be the resulting chain. Then one can show, by means of backwards coupling construction, that the process $\{X_n\}$ *strongly couples* (in the sense of the definition of the previous subsection). with the stationary version. (The paper [21] contains the proof of this assertion for the special case where $R$ is a singleton; however, the construction can be extended to the general case.)

### 4.3. Relation between Harris theory and renovating events

Consider a Harris chain in a Polish space $\mathcal{X}$ and represent it as an SRS:

$$X_{n+1} = f(X_n, \xi_n),$$

where $\{\xi_n\}$ are i.i.d.r.v.'s, which, without loss of generality, can be taken to be uniformly distributed in the unit interval $[0, 1]$, and $f(x, \xi_0)$ has distribution $P(x, \cdot)$. Assuming that a regeneration set $R$ exists, we can represent the chain by another SRS. To simplify, suppose $\ell = 1$. Introduce random variables $\{(\xi_n, \zeta_n, Z_n), n \in \mathbb{Z}\}$, which are i.i.d., and such that $\{\xi_n\}$, $\{\zeta_n\}$, $\{Z_n\}$ are independent. The first sequence is as before. The second one is 0/1–valued with $P(\zeta_n = 1) = 1 - P(\zeta_n = 0) = p$. The third one has $P(Z_n \in \cdot) = Q(\cdot)$. Let $f(x, \xi_0)$ be distributed according to $P(x, \cdot)$, as before, and $\overline{f}(x, \xi_0)$ according to $\overline{P}(x, \cdot) := (1 - p)^{-1}(P(x, \cdot) - pQ(\cdot))$. Consider then the SRS

$$\begin{aligned} X_{n+1} &= h(X_n; \xi_n, \zeta_n, Z_n) \\ &= f(X_n, \xi_n)\mathbf{1}(X_n \notin R) + (\zeta_n Z_n + (1 - \zeta_n)\overline{f}(X_n, \xi_n))\mathbf{1}(X_n \in R). \end{aligned}$$

It is clear that $P(X_{n+1} \in \cdot \mid X_n = x) = P(x, \cdot)$, and so the second SRS defines the same chain in law. Observe that the event

$$A_n := \{X_n \in R, \zeta_n = 1\}$$

is $\langle n, 0, g\rangle$–renovating, with $g(Z_n) \equiv Z_n$, i.e., on $\{X_n \in R, \zeta_n = 1\}$, we have $X_{n+1} = Z_n$. The difficulty with this type of event is that the sequence $\{A_n\}$ is non-stationary, and so Theorem 3 does not immediately apply. Although there is an analog of it for non-stationary renovating events, we chose not to present it in this paper.

To describe a rigorous connection, consider first the case where $R = \mathcal{X}$, i.e., the whole state space is a regeneration set. Then the event

$$A_n := \{\zeta_n = 1\}$$

is $\langle n, 0, g\rangle$–renovating, and the sequence $\{A_n\}$ is stationary. Here Theorem 3 applies immediately.

Now, more generally, assume that $\{X_n\}$ is a Harris process for which

$$E_x\tau_R < \infty, \quad \text{for all } x \in \mathcal{X}.$$

Then $\{X_n\}$ has a stationary version $\{\tilde{X}_n\}$ and there is *strong coupling* between the two processes. It is known that, for any process which strongly couples with a stationary one, there exists a sequence of stationary positive renovating events (see Borovkov [8]). In this sense, there is an intimate connection between Harris ergodicity of a Markov chain and existence of a stationary sequence of renovating events.

## 5. Existence not Based on Regeneration or Renovation

While renovating events provide criteria for strong stability, it is frequently the case that such strong stability may not take place, and that we only have weak convergence, if any convergence at all. Going to the other extreme, we should therefore ask when stationary solutions exist, ignoring problems of uniqueness or convergence. For such questions, very little is known outside the Markovian world.

A setup developed in $[1, 2]$ adopts the following point of view: think of the SRS as a flow on an enriched probability space and ask whether there exists a measure invariant under this flow. If there is, then projecting this measure back to the original probability space will provide a stationary solution. We refer to this procedure as "weak stationary solution" and describe it in the sequel.

### 5.1. General method: weak stationary solutions

Consider an SRS $X_{n+1} = f(X_n, \xi_n)$ as before, and let $(\Omega, \mathscr{F}, P)$ be the underlying probability space, equipped with a measurable bijection $\theta : \Omega \to \Omega$ that preserves $P$ and which is also ergodic. In addition, assume that $\Omega$ is a Polish space (i.e., a complete, separable, metrizable topological space) with $\mathscr{F}$ being the $\sigma$-algebra the Borel sets of $\Omega$.

Define, for each $\omega \in \Omega$, the map $\varphi_0(\omega) : \mathcal{X} \to \mathcal{X}$ by

$$\varphi_0(\omega)(x) := f(x, \xi_0(\omega))$$

and assume that there is a Polish space $\mathcal{X}$, such that $P(\varphi_0 \in \mathcal{X}) = 1$. Letting $\varphi_n(\omega) := \varphi_0(\theta_n(\omega))$, our SRS reads $X_{n+1} = \varphi_n(X_n)$. Clearly, $\{\varphi_n\}$ is a stationary-ergodic sequence of random elements of $\mathcal{X}$. Next consider the enlarged probability space $\Omega \times \mathcal{X}$ and, on it, define the new flow (or "shift")

$$\Theta : \Omega \times \mathcal{X} \to \Omega \times \mathcal{X};$$
$$\Theta(\omega, x) := (\theta\omega, \varphi_0(\omega)(x)).$$

It is clear that $\Omega \times \mathcal{X}$ is a Polish space itself and that powers of $\Theta$ behave as:

$$\Theta^n(\omega, x) = (\theta^n\omega, \varphi_{n-1}(\omega) \cdots \varphi_0(\omega)(x)),$$

where successive dots in the formula mean composition of functions. We would like to equip the Polish space $\Omega \times \mathcal{X}$ with an appropriate probability measure $Q$. First, we define extended random variables $\overline{X}_n, \overline{\varphi}_n$, on $\Omega \times \mathcal{X}$, by:

$$\overline{X}_0(\omega, x) = x, \quad \overline{\varphi}_0(\omega, x) = \varphi_0(\omega),$$
$$\overline{X}_n(\omega, x) = X_0(\Theta^n(\omega, x)), \quad \overline{\varphi}_n(\omega, x) = \overline{\varphi}_0(\Theta^n(\omega, x)).$$

Then we observe that

$$\overline{X}_n(\omega, x) = \varphi_{n-1}(\omega) \cdots \varphi_0(\omega)(x), \quad \overline{\varphi}_n(\omega, x) = \varphi_n(\omega),$$

and so

$$\overline{X}_{n+1}(\omega, x) = \overline{\varphi}_n(\overline{X}_n(\omega, x)),$$

which, upon omitting the argument $(\omega, x)$, as in common probability usage, reads

$$\overline{X}_{n+1} = \overline{\varphi}_n(\overline{X}_n).$$

The point is this: if $Q$ is $\Theta$-invariant and has $\Omega$-marginal $P$ (i.e., if $Q$ is a lifting of $P$ from $\Omega$ to $\Omega \times \mathcal{X}$) then the last recursion is identical, in law, to the original one, because the $P$-stationarity of $\{\varphi_n\}$ on $\Omega$ is tantamount to the $Q$-stationarity of $\{\overline{\varphi}_n\}$ on $\Omega \times \mathcal{X}$ and because the $P$-law of $\{\varphi_n\}$ coincides with the $Q$-law of $\{\overline{\varphi}_n\}$. The existence of such a $Q$ is therefore what we are after:

**Definition 6.** *A probability measure $Q$ on $\Omega \times \mathcal{X}$ that is $\Theta$-stationary and has $\Omega$-marginal $P$ is called a* weak stationary solution *to the SRS.*

The following theorem basically says that, under some assumptions, tightness translates into existence of a weak stationary solution.

**Theorem 4.** *Let $Q_0$ be a probability measure on $\Omega \times \mathcal{X}$ with $\Omega$-marginal $P$. Suppose that the sequence of probability measures $\{Q_0 \circ \Theta^n\}$ is tight, and that $\Theta$ is continuous. Then there exists a weak stationary solution.*

Since continuity may be too strong, it should be weakened. Define

$$\overline{Q}_n = \frac{1}{n} \sum_{i=0}^{n-1} Q_i. \tag{12}$$

Then:

**Theorem 5.** *Suppose the continuity assumption for $\Theta$ in the preceding theorem is replaced by any of the following conditions:*

- *The set of discontinuities of $\Theta$ is $Q$-null, for some weak limit point $Q$ of $\{\overline{Q} \circ \Theta^n\}$.*
- *There exists a sequence $\{\overline{\Theta}_\ell\}$ of continuous maps on $\Omega \times \mathcal{X}$ and a sequence $\{U_\ell\}$ of open subsets of $\Omega \times \mathcal{X}$ such that $\overline{\Theta}_\ell = \overline{\Theta}$ outside $U_\ell$, for all $\ell$, and $\lim_{\ell \to \infty} \underline{\lim}_{n \to \infty} \overline{Q}_n(U_\ell) = 0$*

*Then the conclusion still holds: there is a weak stationary solution.*

To prove Theorems 4–5, we use the continuity condition or the weaker conditions of Theorem 5 in order to extract a subsequential limit $Q$ of the Cesàro averages (12). It can then be checked that any such subsequential limit satisfies $Q \circ \Theta^{-1} = Q$ and has $\Omega$-marginal $P$, and so, according to Definition 6, it qualifies as a weak stationary solution.

To check tightness on the enlarged probability space is not as bad as it sounds, because checking tightness on the original probability space is enough. Indeed, pick any $x_0 \in \mathcal{X}$ and let $Q_0^{x_0}$ be the distribution of $(\omega, x_0)$. If we can show tightness of the sequence $\{\varphi_{n-1} \cdots \varphi_n(x), n = 1, 2, \ldots\}$ on the original probability space, then we have also shown tightness of $\{Q_0^{x_0} \circ \Theta^{-n}, n = 1, 2, \ldots\}$.

### 5.2. Compact state space

A particular case of interest is that of a compact state space. It might, at first sight, appear that any SRS in a compact state space, and with stationary driver, admits a (weak) stationary solution. Here is a counterexample. Consider the deterministic SRS, in $[0, 1]$, defined by

$$X_{n+1} = \frac{1}{2} X_n \mathbf{1}(0 < X_n \leq 1) + \mathbf{1}(X_n = 0).$$

It is easy to see that there is no probability measure on $[0, 1]$ that remains invariant for this SRS. What is bad here is the discontinuity at 0.

To remedy the situation, consider $X_{n+1} = \varphi_n(X_n)$, where $\{\varphi_n\}$ are stationary-ergodic random continuous maps of $\mathcal{X}$ into itself, where $\mathcal{X}$ is a compact Polish space. We can realize this recursion on the probability space $(\Omega, \mathscr{F})$, where $\Omega = C(\mathcal{X}, \mathcal{X})^{\mathbb{Z}}$. Here, $C(\mathcal{X}, \mathcal{X})$

is the space of continuous maps from $\mathcal{X}$ into $\mathcal{X}$, equipped with the topology of uniform convergence. It can be proved that $C(\mathcal{X},\mathcal{X})$ is Polish. The $\sigma$-field on $C(\mathcal{X},\mathcal{X})$ is taken to be the Borel $\sigma$-field generated by the open sets of $C(\mathcal{X},\mathcal{X})$. So, $\Omega$ is then the set of doubly-infinite sequences of elements of $C(\mathcal{X},\mathcal{X})$, and $\mathscr{F}$ is the product $\sigma$-field. The space $\Omega$ is also Polish. The shift $\theta$ is the canonical one-step shift. It is easy to see that the extended shift $\Theta$ is also continuous, owing to the continuity of $\varphi_0$. Hence the previous Theorem holds, and so the SRS has a weak stationary solution.

### 5.3. An application

Here is a methodological application of this last fact. Suppose that $\{X_n\}$ is a Markov chain that can be represented as SRS $X_{n+1} = f(X_n, \xi_n) = \varphi_n(X_n)$ with $f(\cdot, \xi_n) \equiv \varphi_n(\cdot)$ a continuous function. Suppose that we can prove (e.g., using Lyapunov function methods) that there exists a compact set $R$ which is positive recurrent:

$$\sup_{x\in R} E_x \tau_R < \infty.$$

Suppose also that any of the conditions of Theorem 5 hold. Then there exists a stationary version for the entire chain. The reason is as follows: Consider the embedded chain $\{X_k^R\}$ at those times at which the original chain is in $R$. Then, owing to the discussion in the previous paragraph, this embedded chain possesses a stationary version. Indeed, if $T_k$ is the $k$-th time $n$ at which $X_n \in R$, we have

$$X_{k+1}^R = X_{T_{k+1}} = \varphi_{T_{k+1}-1}\varphi_{T_{k+1}-2}\cdots\varphi_{T_k}(X_{T_k}) \equiv \psi_k(X_k^R),$$

where $\psi_k := \varphi_{T_{k+1}-1}\varphi_{T_{k+1}-2}\cdots\varphi_{T_k}$. Composition of continuous functions is continuous and $X^R$ lives in a compact state space. Hence the last SRS possesses a weak stationary solution. This, together with the fact that $R$ is positive recurrent enable us to construct a stationary version for the original chain. This is a method for existence that is not based on regeneration.

## 6. Monotonicity Methods

The paper of Loynes [28] was the first to consider a system (single queue, and, later, queues in tandem) with stationary-ergodic driver. The classical recursion $X_{n+1} = (X_n + \xi_n)^+$ studied by Loynes is monotone. We next provide a little survey about the important area of monotone recursions.

### 6.1. General statements on monotone and homogeneous recursions

There are many applications, especially in queueing networks, where monotonicity in the dynamics can be exploited to prove existence and uniqueness of stationary solutions. Although the theory can be presented in the very general setup of a partially ordered state space (see Brandt *et al.* [12]) we will only focus on the case where the state is $\mathbb{R}^d$. Consider then the SRS

$$X_{n+1} = f(X_n, \xi_n) =: \varphi_n(X_n)$$

and assume that $\varphi_0 : \mathbb{R}_+^d \to \mathbb{R}_+^d$ is increasing and right-continuous, where the ordering is the standard component-wise ordering[16] on $\mathbb{R}^d$. Let $\theta$ be stationary and ergodic flow on $(\Omega, \mathscr{F}, P)$ and assume that $\varphi_n = \varphi_0 \circ \theta^n$, $n \in \mathbb{Z}$. In other words, $\{\varphi_n\}$ is a stationary-ergodic

[16] For $x = (x^i, i = 1,\ldots,d), y = (y^i, i = 1,\ldots,d)$, we say that $x \le y$ iff $x^i \le y^i$ for all $i$. A function $\varphi : \mathbb{R}^d \to \mathbb{R}^d$ is said to be increasing iff $x \le y \Rightarrow \varphi(x) \le \varphi(y)$.

sequence of random elements of the space of right-continuous increasing functions on $\mathbb{R}^d_+$. We first explain Loynes' method. Define

$$\Phi_n := \varphi_{n-1} \cdots \varphi_0.$$

Thus, $\Phi_n(Y)$ is the solution of the SRS at $n \geq 0$ when $X_0 = Y$, a.s. Since 0 is the least element of $(R^d_+, \leq)$, we have $\Phi_n(0) \leq \Phi_n(Y)$, a.s., for any $R^d_+$–valued r.v. $Y$. Next consider

$$\Phi_{m+n}(0) \circ \theta^{-m} = \varphi_{n-1} \cdots \varphi_{-m}(0), \quad n \geq -m,$$

and interpret $\Phi_{m+n}(0)$ as the solution of the SRS at time $n \geq -m$, starting with 0 at time $-m$. Clearly, $\Phi_{m+n}(0)$ increases as $m$ increases, because:

$$\begin{aligned}\Phi_{(m+1)+n}(0) \circ \theta^{-(m+1)} &= \varphi_{n-1} \cdots \varphi_{-m} \varphi_{-(m+1)}(0) \\ &= \varphi_{n-1} \cdots \varphi_{-m}(\varphi_{-(m+1)}(0)) \\ &\geq \varphi_{n-1} \cdots \varphi_{-m}(0) = \Phi_{m+n}(0) \circ \theta^{-m}.\end{aligned}$$

Finally define

$$\tilde{X}_n := \lim_{m \to \infty} \Phi_{m+n}(0) \circ \theta^{-m}, \quad n \in \mathbb{Z}.$$

The r.v. $\tilde{X}_n$ is either finite a.s., or is infinite a.s., by ergodicity. Assuming that the first case holds, we further have

$$\begin{aligned}\tilde{X}_{n+1} &= \lim_{m \to \infty} \Phi_{m+(n+1)}(0) \circ \theta^{-m} \\ &= \lim_{m \to \infty} \varphi_{m+n} \varphi_{m+n-1} \cdots \varphi_0(0) \circ \theta^{-m} \\ &= \lim_{m \to \infty} \varphi_n \varphi_{n-1} \cdots \varphi_{-m}(0) \\ &= \lim_{m \to \infty} \varphi_n(\varphi_{n-1} \cdots \varphi_{-m}(0)) \\ &= \lim_{m \to \infty} \varphi_n(\Phi_{m+n}(0) \circ \theta^{-m}) \\ &= \varphi_n(\tilde{X}_n).\end{aligned}$$

Provided then that we have a method for proving $P(\tilde{X}_0 < \infty) > 0$, Loynes' technique results in the construction of a stationary-ergodic solution $\{\tilde{X}_n\}$ of the SRS.

Without further assumptions and structure, not much can be said. Assume next that, in addition, $\varphi_0$ is homogeneous, i.e.,

$$\varphi_0(x + c\mathbf{1}) = \varphi_0(x) + c\mathbf{1},$$

for all $x \in \mathbb{R}^d_+$ and all $c \in \mathbb{R}$. Such is the case, e.g., with the usual Lindley function $\varphi_0 : \mathbb{R}_+ \to \mathbb{R}_+$, with $\varphi_0(x) = \max(x + \xi_0, 0)$. The homogeneity assumption is quite frequent in queueing theory. It is easy to see that

$$|\varphi_0(x) - \varphi_0(y)| \leq |x - y|,$$

where $|x| := \max(|x_1|, \ldots, |x_d|)$. Suppose then that $\{X_n\}$, $\{Y_n\}$ are two stationary solutions of the SRS. Then $|X_{n+1} - Y_{n+1}| = |\varphi_n(X_n) - \varphi_n(Y_n)| \leq |X_n - Y_n|$, for all $n$, a.s., and since $\{|X_n - Y_n|, n \in \mathbb{Z}\}$ is stationary and ergodic, this a.s. monotonicity may only hold if $|X_n - Y_n| = r$, for some constant $r \geq 0$. Thus, a necessary and sufficient condition for the two solutions to coincide is that

$$P(|\varphi_0(X_0) - \varphi_0(Y_0)| < |X_0 - Y_0|) > 0.$$

A classical example where this is the case is the $G/G/s$ queue, that is, the $s$-server queue with stationary-ergodic data. Let $\lambda$, $\mu$ be the arrival and service rates, respectively. Here, there is a minimal and a maximal stationary solution which, provided that $\lambda < s\mu$, coincide a.s. For details see Brandt et al. [12].

### 6.2. The Monotone-Homogeneous-Separable (MHS) framework

Consider a recursion of the form

$$W_{n+1} = f(W_n, \xi_n, \tau_n),$$

where $\xi_n$ are general marks, and $\tau_n \geq 0$. The interpretation is that $\tau_n$ is the interarrival time between the $n$-th and $n+1$-th customer, and $W_n$ is the state just before the arrival of the $n$-th customer. We consider arrival epochs $\{T_n\}$ such that $T_{n+1} - T_n = \tau_n$. We write $W_{m,n}$ for the solution of the recursion at index $n$ when we start with a specific state, say 0, at $m \leq n$. Finally we consider a functions of the form

$$X_{[m,n]} = f_{m+n-1}(W_{m,n}; T_m, \ldots, T_n; \xi_m, \ldots, \xi_n),$$

which will be thought of as epochs of last activity in the system. For instance, when we have an $s$-server queue, $X[m,n]$ represents the departure time of the last customer when the queue is fed only by customers with indices from $m$ to $n$. Correspondingly, we define the quantity

$$Z_{[m,n]} := X_{[m,n]} - T_n,$$

the time elapsed between the arrival of the last customer and the departure of the last customer. The framework is formulated in terms of the $X_{[m,n]}$, $Z_{[m,n]}$ and their dependence on the $\{T_n\}$. For $c \in \mathbb{R}$, let $\{T_n\} + c = \{T_n + c\}$. For $c > 0$, let $c\{T_n\} = \{cT_n\}$. Define $\{T_n\} \leq \{T'_n\}$ if $T_n \leq T'_n$ for all $n$. We require a set of four assumptions:

**(A1)** $Z_{[m,n]} \geq 0$

**(A2)** $\{T_n\} \leq \{T'_n\} \Rightarrow X_{[m,n]} \leq X'_{[m,n]}$.

The first assumption is natural. In the second one, $X'_{[m,n]}$ are the variables obtained by replacing each $T_n$ by $T'_n$; it says that delaying the arrival epochs results in delaying of the last activity epochs.

**(A3)** $\{T'_n\} = \{T_n\} + c \Rightarrow X'_{[m,n]} = X_{[m,n]} + c$.

This is a time-homogeneity assumption.

**(A4)** For $m \leq \ell < \ell + 1 \leq n$, $X_{[m,\ell]} \leq T_{\ell+1} \Rightarrow X_{[m,n]} = X_{[\ell+1,n]}$.

If the premise $X_{[m,\ell]} \leq T_{\ell+1}$ of the last assumption holds, we say that we have separability at index $\ell$. It means that the last activity due to customers with indices in $[m,\ell]$ happens prior to the arrival of the $\ell+1$-th customer, and so the last activity due to customers with indices in $[m,n]$ is not influenced by those customers with indices in $[m,\ell]$. Basic consequences of the above assumptions are summarized in:

**Lemma 1.** *(i) The response $Z_{[m,n]}$ depends on $T_m, \ldots, T_n$ only through the differences $\tau_m, \ldots, \tau_{n-1}$.*
*(ii) Let $a \leq b$ be integers. Let $T'_n = T_n + Z_{[a,b]}\mathbf{1}(n > b)$, $T''_n = T_n - Z_{[a,b]}\mathbf{1}(n \leq b)$. And let $X'_{[m,n]}$, $X''_{[m,n]}$ be the corresponding last activity epochs. Then both of them exhibit separability at index $b$.*
*(iii) The variables $X_{[m,n]}$, $Z_{[m,n]}$ increase when $m$ decreases.*
*(iv) For $a \leq b < b+1 \leq c$, $Z_{[a,c]} \leq Z_{[a,b]} + Z_{[b+1,c]}$.*

*Proof.* (i) Follows from the definition $Z_{[m,n]} = X_{[m,n]} - T_n$ and the homogeneity assumption (A3).
(ii) Obviously, $Z_{[a,b]} \leq \tau_b + Z_{[a,b]}$, and so $X_{[a,b]} - T_b \leq \tau_b + Z_{[a,b]}$, which implies $X_{[a,b]} \leq T_{b+1} + Z_{[a,b]}$. The right-hand side is $T'_{b+1}$, by definition. The left-hand side is equal to $X'_{[a,b]}$ because $T'_n = T_n$ for $n \leq b$. So $X'_{[a,b]} \leq T'_{b+1}$ and this is separability at index $b$. Similarly for the other variable.
(iii) Let $a = b = m$ in (ii). Since we have separability at index $m$, we conclude that $X''_{[m,n]} = X''_{[m+1,n]}$. But $T''_k = T_k$ for $k \in [m+1, n]$ and so $X''_{[m+1,n]} = X_{[m+1,n]}$. On the other hand, $\{T''_k\} \leq \{T_k\}$ and so, by (A2), $X''_{[m,n]} \leq X_{[m,n]}$. Thus $X_{[m,n]} \geq X_{[m+1,n]}$. And so $Z_{[m,n]} \geq Z_{[m+1,n]}$ also.
(iv) Apply (ii) again. Since $\{T_k\} \leq \{T'_k\}$, (A2) gives $X_{[a,c]} \leq X'_{[a,c]}$. By separability at index $b$, as proved in (ii), we have $X'_{[a,c]} = X'_{[b+1,c]}$. Because $T'_k = T_k + Z_{[a,b]}$ for all $k \in [b+1, c]$, we have, by (A3), $X'_{[b+1,c]} = X_{[b+1,c]} + Z_{[a,b]}$. Thus, $X_{[a,c]} \leq X_{[b+1,c]} + Z_{[a,b]}$. Subtracting $T_c$ from both sides gives the desired. □

Introduce next the usual stationary-ergodic assumptions. Namely, consider $(\Omega, \mathscr{F}, P)$ and a stationary-ergodic flow $\theta$. Let $\xi_n = \xi_0 \circ \theta^n$, $\tau_n = \tau_0 \circ \theta^n$, set $T_0 = 0$, and suppose $E\tau_0 = \lambda^{-1} \in (0, \infty)$, $EZ_{0,0} < \infty$. Stability of the original system can, in specific but important cases, be translated in a stability statement for $Z_{[m,n]}$. Hence we shall focus on it. Note that $Z_{[m,n]} \circ \theta^k = Z_{[m+k,n+k]}$ for all $k \in \mathbb{Z}$. For any $c \geq 0$, introduce the epochs $c\{T_n\} = \{cT_n\}$ and let $X_{[m,n]}(c)$, $Z_{[m,n]}(c)$ be the quantities of interest. The subadditive ergodic theorem gives that

$$\gamma(c) := \lim_{n\to\infty} \frac{1}{n} Z_{[-n,-1]}(c) = \lim_{n\to\infty} \frac{1}{n} EZ_{[-n,-1]}(c)$$

is a nonnegative, finite constant. The previous lemma implies that $\gamma(c) \geq \gamma(c')$ when $c > c'$. Similarly, $\lim n^{-1} X_{[1,n]}(c) = \gamma(c) + \lambda^{-1} c$, and the latter quantity increases as $c$ increases. Monotonicity implies that $Z_{[-n,-1]}(c)$ increases as $n$ increases, and let $\tilde{Z}(c)$ be the limit. Ergodicity implies that $P(\tilde{Z}(c) < \infty) \in \{0, 1\}$. Put $\tilde{Z} = \tilde{Z}(1)$. The stability theorem[17] is:

**Theorem 6.** *If $\lambda\gamma(0) < 1$ then $P(\tilde{Z} < \infty) = 1$. If $\lambda\gamma(0) > 1$ then $P(\tilde{Z} < \infty) = 0$.*

*Proof.* Assume first that $\lambda\gamma(0) > 1$. Fix $n \geq 1$. Define $T'_k = T_{-n}$ for all $k \in \mathbb{Z}$. Hence $X'_{[-n,0]}(1) \leq X_{[-n,0]}(1) = Z_{[-n,0]}(1)$, by (A2). On the other hand, by (A3), $X'_{[-n,0]}(1) = X_{[-n,0]}(0) + T_{-n} = Z_{[-n,0]}(0) + T_{-n}$. Thus, $n^{-1} Z_{[-n,0]}(1) \geq n^{-1} Z_{[-n,0]}(0) + n^{-1} T_{-n}$, and, taking limits as $n \to \infty$, we conclude $\liminf n^{-1} Z_{[-n,0]}(1) \geq \gamma(0) - \lambda^{-1} > 0$, a.s.

Assume next that $\lambda\gamma(0) < 1$. Let $\gamma_n(0) := EZ_{[-n+1,0]}(0)/n$. Since $\gamma(0) = \lim_{n\to\infty} \gamma_n(0) = \inf_n \gamma_n(0)$, we can find an integer $K$ such that $\lambda\gamma_K(0) < 1$. Consider next an auxiliary single server queue with service times $s_n := Z_{[-Kn+1,-K(n-1)]}(0)$ and interarrival times $t_n := \sum_{i=-Kn+1}^{-K(n-1)} \tau_i$. Notice that $\{(t_n, s_n), n \in \mathbb{Z}\}$ is stationary-ergodic and consider the waiting time $W_n$ of this auxiliary system: $W_{n+1} = (W_n + s_n - t_n)^+$. Since $Es_n = \gamma_K < \lambda^{-1} = Et_n$, the auxiliary queue is stable. Since the separability property holds, we have the following domination:

$$Z_{[-nK+1,0]}(1) \leq W_n \circ \theta^{-n} + s_0, \text{ a.s.},$$

where $W_n$ here is the waiting time of the $n$-th customer if the queue starts empty. By the Loynes' scheme, $W_n \circ \theta^{-n}$ converges (increases) to an a.s. finite random variable. Hence $\tilde{Z} = \lim_n Z_{[-nK+1,0]}(1)$ is also a.s. finite. □

[17] This is known as the "saturation rule"

**Remark:** The saturation rule (in an extended form) was first introduced in the paper by Baccelli and Foss [6]. A variant of the particular version presented here appears also in the book by Baccelli and Brémaud [5].

## 7. Instability

In this last section, we present some new criteria for instability. We focus on a Markov chain $\{X_n\}$ in a Polish space $\mathcal{X}$ and adopt the following strong notion of transience (which is not a standard one): a set $B \subseteq \mathcal{X}$ is called *transient* if $P_x(\tau_B = \infty) > 0$ for all $x \in \mathcal{X}$, where $\tau_B = \inf\{n \geq 1 : \ X_n \in B\}$ is the first return time to $B$. By instability, here, we mean that the members of a certain class of sets is transient. More precisely, let $L : \mathcal{X} \to \mathbb{R}_+$ be a "norm-like" function, i.e., suppose (at least) that $L$ is unbounded. We say that the chain is transient if each set of the form $B_N = \{x \in \mathcal{X} : \ L(x) \leq N\}$ is transient.

In the sequel, we will present criteria that decide whether $\lim_{n\to\infty} L(X_n) = \infty$, $P_x$-a.s. Clearly then, this will imply transience of each $B_N$.

Thinking of $L$ as a Lyapunov function, it is natural to seek criteria that are, in a sense, opposite to those of Theorem 1. One would expect that if the drift $E_x[L(X_1) - L(X_0)]$ is bounded from below by a positive constant, outside a set of the form $B_N$, then that would imply instability. However, this is not true and this has been a source of difficulty in formulating a general enough criterion thus far. To the best of our knowledge, the most general criterion is Theorem 2.2.7. of Fayolle et al. [18] which is, however, rather restrictive because (i) it is formulated for countable state Markov chains and (ii) it requires that a transition from a state $x$ to a state $y$, with $L(x) - L(y)$ larger than a certain constant, is not possible. However, it gives insight as to what problems one might encounter: one needs to regulate, not only the drift from below, but also its size when the drift is large.

The theorem below is a generalization of the one mentioned above. First, define

$$\sigma_N := \tau_{B_N^c} = \inf\{n \geq 1 : \ L(X_n) > N\}$$
$$\Delta := L(X_1) - L(X_0).$$

We then have:

**Theorem 7.** *Suppose there exist $N, M, \varepsilon > 0$ and a measurable $h : [0,\infty) \to \ [1,\infty)$ with the property that $h(t)/t$ be concave-increasing on $1 \leq t < \infty$, and $\int_1^\infty h(t)^{-1}dt < \infty$, such that*

**(I1)** $P_x(\sigma_N < \infty) = 1$ *for all* $x$.

**(I2)** $\inf_{x\in B_N^c} E_x[\Delta, \Delta \leq M] \geq \varepsilon$.

**(I3)** *The family $\{P_x(h(\Delta) \in \cdot), \ x \in B_N^c\}$ is uniformly integrable, i.e., $\lim_{K\to\infty} \sup_{x\in B_N^c} \int_K^\infty tP(h(\Delta) \in dt) = 0$.*

*Then $P_x(\lim_{n\to\infty} L(X_n) = \infty) = 1$, for all $x \in \mathcal{X}$.*

This theorem is proved in detail in [19]. We remark that there are extensions for non-homogeneous Markov chains. Condition (I1) says that the set $B_N^c$ is recurrent. Of course, if the chain itself forms one communicating class, then this condition is automatic. Condition (I2) is the positive drift condition. Condition (I3) is the condition that regulates the size of the drift. We also note that an analog of this theorem, with state-dependent drift can also be derived. (The theorem of Fayolle et al. does use state-dependent drift.)

To see that (I3) is essential, consider the following example: Let $\mathcal{X} := \mathbb{Z}_+$, and $\{X_n\}$ a Markov chain with transition probabilities

$$p_{i,i+1} = 1 - p_{i,0}, \quad i \geq 1,$$
$$p_{0,1} = p_{0,0} = 1/2.$$

Suppose that $0 < p_{i,0} < 1$ for all $i$, and $\sum_i p_{i,0} < \infty$. Then the chain forms a single communicating class. Also, with $\tau_0$ the first return to 0, we have

$$P_i(\tau_0 = \infty) = \prod_{j \geq i} (1 - p_{j,0}) > 0.$$

So the chain is transient. However note that the natural choice for $L$, namely $L(x) \equiv x$ trivially makes (I2) true.

We finally mention that, frequently, the choice $h(t) := t^{1+\delta}$, for some $\delta \in (0,1)$, may be a suitable one. In such a case, only conditions (I1) and (I2) are needed.

**Remark:** We presented here a criterion in terms of drifts of the original Markov chain. There exist recent results for instability of a stochastic network in terms of conditions for their fluid limits (see, e.g., Meyn [29], Puhalskii and Rybko [33], and Gamarnik and Hasenbein [23]).

## 8. Other Methods

We have not covered every possible method in this paper. In fact, there are other ones, some of which are more case-specific.

For instance, there are comparison methods. Frequently, it is the case that one can somehow dominate the system under study by a system whose stability is known or can easily be deciphered. Quite useful for this kind of method are the stochastic ordering concepts; see, e.g., Baccelli and Brémaud [5].

Another method is based on contractivity: Suppose that the Markov chain $\{X_n\}$ in a Polish space $\mathcal{X}$ with metric $\rho$ is represented by the SRS $X_{n+1} = f(X_n, \xi_n)$ which is contractive in the first argument in the following sense: There is $x_0 \in \mathcal{X}$ such that

$$\rho(f(x, \xi_0), f(x_0, \xi_0)) \leq \rho(x, x_0), \text{ a.s., for all } x \in \mathcal{X}.$$

Suppose also that the set $B_{N_0} = \{x : \ \rho(x, x_0) \leq N\}$ is positive recurrent and compact. Write $X_n^x$ for the chain started at $x$ (i.e., the solution of the SRS started at $x$). In addition to the above, assume that there is $m \in \mathbb{N}$, $\gamma, \delta \in (0,1)$, such that

$$P(\rho(X_m^x, X_m^{x_0}) \leq \gamma \rho(x, x_0)) > \delta, \text{ for all } x \in B_{N_0}.$$

Since contractivity implies continuity, we can use what is described in Sections 5.2, 5.3 to prove that there exists at least one stationary distribution. Then we can use the inequalities stipulated above to prove convergence toward this stationary distribution; see, e.g.. Borovkov [8].

We also mention that, in several applications, direct use of the subadditive ergodic theorem (see, e.g., Liggett [26, pg. 277]) is used to prove stability. It played a key role, for instance, in Section 6.2, where the "monotone-homogeneous-separable" framework was discussed.

Martingale arguments are used either explicitly or implicitly. For example, the instability considerations of Section 7 depend crucially on martingale arguments (c.f. [19]). Also, the proof of Theorem 1 makes implicit use of martingale arguments; special cases of it are frequently formulated in terms of supermartingales. (For standard basic martingale theory see, e.g., Chung [14] and Shiryayev [36].)

Finally, large deviation techniques (which we did not touch at all in this survey) have also been used in stability and instability studies of various stochastic systems; see, e.g., Puhalskii and Rybko [33], and Gamarnik and Hasenbein [23].

### Acknowledgements

The authors would like to thank Sean Meyn, Sasha Stolyar, Bert Zwart, and two anonymous referees for their helpful and constructive comments and suggestions.

### References

[1] V. Anantharam and T. Konstantopoulos: Stationary solutions of stochastic recursions describing discrete event systems. *Stochastic Processes and Applications*, **68** (1997), 181-194.

[2] V. Anantharam and T. Konstantopoulos: A correction and some additional remarks on: stationary solutions of stochastic recursions describing discrete event systems. *Stochastic Processes and Applications*, **80** (1999), 271-278.

[3] L. Arnold: *Random Dynamical Systems* (Springer-Verlag, Berlin, 1998).

[4] S. Asmussen: *Applied Probability and Queues* (Springer, New York, 2003).

[5] F. Baccelli and P. Brémaud: *Elements of Queueing Theory* (Springer, Berlin, 2003).

[6] F. Baccelli and S. Foss: On the saturation rule for the stability of queues. *Journal of Applied Probability*, **32** (1995), 494-507.

[7] A. A. Borovkov: *Asymptotic Methods in Queueing Theory* (Wiley, New York, 1984).

[8] A. A. Borovkov: *Ergodicity and Stability of Stochastic Processes* (Wiley, New York, 1998).

[9] A. A. Borovkov and S. G. Foss: Stochastically recursive sequences and their generalizations. *Siberian Advances in Mathematics*, **2** (1992), n.1, 16-81.

[10] A. Boyarsky and P. Góra: *Laws of Chaos: Invariant Measures and Dynamical Systems in One Dimension.* (Birkhäuser, Boston, 1997).

[11] M. Bramson: Instability of FIFO queueing networks with quick service times. *The Annals of Applied Probability* **4** (1993), 693-718.

[12] A. Brandt, P. Franken and B. Lisek: *Stationary Stochastic Models* (Wiley, New York, 1992).

[13] H. Chen and D. D. Yao: *Fundamentals of Queueing Networks* (Springer, New York, 2001).

[14] K. L. Chung: *A Course in Probability Theory* (Academic Press, New York, 1974).

[15] J. G. Dai: On positive Harris recurrence of multiclass queueing networks: a unified approach via fluid limit models. *The Annals of Applied Probability*, **5** (1995), 49-77.

[16] P. Diaconis and P. Freedman: Iterated random functions. *SIAM Review*, **41** (1999), 45-76.

[17] M. Duflo: *Random Iterative Models* (Springer, Berlin, 1997).

[18] G. Fayolle, V. Malyshev and M. Menshikov: *Topics in the Constructive Theory of Markov Chains* (1995).

[19] S. G. Foss and D. E. Denisov: On transience conditions for Markov chains. *Siberian Mathematical Journal*, **42** (2001), 425-433.

[20] S. Foss and T. Konstantopoulos: Extended renovation theory and limit theorems for stochastic ordered graphs. *Markov Processes and Relared Fields*, **9** (2003), 413-468.

[21] S. G. Foss, R. L. Tweedie and J. N. Corcoran: Simulating the invariant measures of Markov chains using backward coupling at regeneration times. *Probability in Engineering and Informational Sciences*, **12** (1998), 303-320.

[22] F. G. Foster: On the stochastic matrices associated with certain queueing processes. *The Annals of Mathematical Statistics*, **24** (1953), 355-360.

[23] D. Gamarnik and J. Hasenbein: Instability in stochastic and fluid queueing networks. *The Annals of Applied Probability*, (2003), To appear.

[24] Yu. Kifer: *Ergodic Theory of Random Transformations* (Birkhäuser, Boston, 1986).

[25] A. Lasota and M. C. Mackey: *Chaos, Fractals, and Noise: Stochastic Aspects of Dynamics* (Springer, New York, 1994).

[26] T. M. Liggett: *Interacting Particle Systems* (Springer, New York, 1985).

[27] T. Lindval: *Lectures on the Coupling Method* (Wiley, New York, 1992).

[28] R. M. Loynes: The stability of queues with non-independent interarrival and service times. *Mathematical Proceedings of the Cambridge Philosophical Society*, **58** (1962), 497-520.

[29] S. P. Meyn: Transience of multiclass queueing networks and their fluid models. *The Annals of Applied Probability*, **5** (1995), 946-957.

[30] S. P Meyn and R. L. Tweedie: *Markov Chains and Stochastic Stability* (Springer, New York, 1993).

[31] E. Nummelin: *General Irreducible Markov Chains and Nonnegative Operators* (Cambridge, U.P., Cambridge, 1984).

[32] A. G. Pakes: Some conditions for ergodicity and recurrence of Markov chains. *Operations Research*, **17** (1969), 1048-1061.

[33] A. A. Puhalskii and A. N. Rybko: Nonergodicity of queueing networks when their fluid models are unstable. *Problems of Information Transmission*, **36** (2000), 26-46.

[34] Ph. Robert: *Stochastic Networks and Queues* (Springer, Berlin, 2003).

[35] A. N. Rybko and A. L. Stolyar: Ergodicity of stochastic processes describing the operations of open queueing networks. *Problems of Information Transmission*, **28** (1992), 3-26.

[36] A. N. Shiryayev: *Probability* (Springer, New York, 1984).

[37] H. Thorisson: *Coupling, Stationarity, and Regeneration* (Springer, New York, 2000).

[38] R. L. Tweedie: Criteria for classifying general Markov chains. *Advances in Applied Probability* **8** (1976), 737-771.

[39] W. Whitt: *Stochastic-Process Limits* (Springer, New York, 2002).

Takis Konstantopoulos
Department of Actuarial Mathematics and Statistics
School of Mathematical and Computer Sciences
Heriot-Watt University
Edinburgh EH14 4AS, UK
E-mail: `takis@master.math.upatras.gr`