Fluid Approximation Approach and Induced Vector Fields

Serguei Foss   Takis Konstantopoulos

Stochastic Stability, Lecture 2

Short Instructional Courses, September 2006

Heriot-Watt University
Edinburgh

We consider two stability methods for Markov chains based on drift analysis.

1 Fluid approximation approach

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, multi-class networks. For statements and proofs of the functional approximation theorems used here, the reader may consult the texts of Chen and Yao [3], Whitt [10] and references therein.

1.1 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.[1] 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. 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

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

individual constituents of N behave. So, we assume that[2]

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

where $c_1 + c_2 + c_3 = 1$. Then

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

uniformly on compact[3] sets of t, a.s., i.e.,

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

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 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 ψ, 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) we have:

$$\frac{\mathbf{E}Q_{[N\lambda t]}}{N} \to 0, \qquad \frac{\mathbf{E}\chi_{[N\lambda t]}}{N} \to 0, \qquad \text{as } N \to \infty, \qquad \text{for } t \ge t_0.$$

[2]Hence, strictly speaking, we should denote the process by 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.

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

In particular there is N_0, so that $\mathbf{E}Q_{[2N\lambda t_0]} + \mathbf{E}\chi_{[2N\lambda t_0]} \le N/2$ for all $N > N_0$. Also, the same uniform integrability condition, allows us to find a constant C such that $\mathbf{E}Q_{[2N\lambda t_0]} + \mathbf{E}\chi_{[2N\lambda t_0]} \le C$ for all $N \le N_0$. 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

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

The last two inequalities can then be written as $\mathbf{E}_x(V(X_{g(x)}) - V(X_0)) \le -h(x)$, $x \in \mathcal{X}$. It is easy to see that the function V, g, h satisfy conditions (L0)-(L4) from the previous lecture. Thus the main Theorem 2 of the previous lecture shows that the set $\{x \in \mathcal{X} : \ V(x) = q + \chi \le N_0\}$ is positive recurrent.

1.2 Fluid limit stability criterion for multiclass queueing networks

We now pass on to multiclass queueing networks. Rybko and Stolyar [9] first applied the method to a two-station, two-class network. Dai [4] generalised the method and his paper established and popularised it. Meanwhile, it became clear that the natural stability conditions[4] may not be sufficient for stability and several examples were devised to exemplify this phenomena; see, e.g., the paper by Bramson [2] 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 $i \ne j$. There is a single arrival stream[5], denoted by $A(t)$, which is the counting process of a renewal process, viz.,

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

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 σ_k used as follows: when a customers from class k is served, then its service time is an independent copy of σ_k. We let $\mu_k^{-1} = \mathbf{E}\sigma_k$. 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 ℓ with probability $p_{k,\ell}$, and leaves the network with probability $p_{k,\infty} - 1 - \sum_{ell} p_{k,\ell}$.

Examples. 1. Jackson-type (or generalised Jackson) network: there is one-to-one correspondence between stations and customer classes.
2. Kelly network. There are several deterministic routes, say, $(j_{1,1}, \ldots, j_{i,r_1}), \ldots, (j_{m,1}, \ldots, j_{m,r_m})$

[4] 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.

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

where $j_{i,r}$ are stations numbers. Introduce $K = \sum_{q=1}^{m} r_q$ customers classes numbered $1, \ldots, K$ and let

$$p_{k,k+1} = 1 \quad \text{for} \quad \neq r_1, r_1 + r_2, \ldots$$

and

$$p_{k,\infty} = 1 \quad \text{for} \quad k = r_1, r_1 + r_2, \ldots.$$

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 $\mathbf{P}(\alpha_k(n) = \ell) = p_{k\ell}$, 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., [3].

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 service. We refer to $X^j(t) = (Q^j(t), \zeta^j(t), \chi^j(t))$ as the state[6] 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.

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

We now let

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

and consider the system parametrised 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.

Example. For the single-server queue, the fluid model is a collection of fluid limits indexed, say by c_1 and c_2.

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[7] $X_n := X(T_n)$. 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 1. *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* $\{X_n\}$*.*

Remark. There is a number of papers where the instability conditions are analysed. One of the most recent is [8] where the large deviations techniques is used..

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

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

stable.

(ii) It is easy to see that the fluid model is stable in the sense of Definition 2 if and only 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, $\mathbf{E}\overline{Q}(t_0) \leq 1 - \varepsilon$.*

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") – see, e.g., [6]. The statement of Theorem 1 stays valid if one replaces the word "stable" by "weakly stable".

Proof of Theorem 1. 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,

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

Also,

$$\lim_{n\to\infty} \frac{1}{n} \sum_j \chi_n^j = 0, \quad \text{a.s.} \tag{2}$$

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

$$\frac{\chi_n^j}{n} \leq \frac{1}{n} \max_{k\in c(j)} \max_{1\leq i\leq D_{k,n}+1} \sigma_k(i) \leq \sum_{k\in c(j)} \frac{D_{k,n}+1}{n} \, \frac{\max_{1\leq i\leq D_{k,n}+1}\sigma_k(i)}{D_{k,n}+1}. \tag{3}$$

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

$$\overline{\lim_{n\to\infty}} \frac{D_{k,n}+1}{n} < \infty, \text{ a.s.} \tag{4}$$

In case the latter quantity is positive, we have that the last fraction of (3) tends to zero. In case the latter quantity is zero then $\chi^j(n)/n \to 0$, because R_k is a.s. finite. 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 (3), 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 (1) and (2) and the uniform integrability we have

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

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

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

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

$$\mathbf{E}_x[V(X_{g(x)}) - V(X_0)] \leq -\frac{1}{2}V(x), \qquad \text{if } V(x) > N_0.$$

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

$$\mathbf{E}_x[V(X_{g(x)}) - V(X_0)] \leq C, \qquad \text{if } V(x) \leq N_0,$$

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

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

□

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

1.3 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 λ. 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 μ_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. 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, and at which $\sum_k \overline{W}_k(t) > 0$,

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

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 1 applies, and so we have positive recurrence.

1.4 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 "generalised 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 σ_j with mean μ_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 describe the (traditional) stability conditions in terms of an auxiliary Markov chain which we call $\{Y_n\}$ and which takes values in $\{0, 1, \ldots, J, J+1\}$, it has transition probabilities $p_{j,i}$, $j \in \{0, 1, \ldots, J\}$, $i \in \{1, \ldots, J\}$, and $p_{j,J+1} = 1 - \sum_{i=1}^{J} p_{j,i}$, $j \in \{1, \ldots, J\}$, $p_{J+1,J+1} = 1$, 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 $\overline{Q}_j(t)' = \lambda_j + \sum_{i=1}^{J} \mu_i p_{i,j} - \mu_j$

and so $|\overline{Q}_j(t)|' = \lambda - \sum_{j=1}^J \sum_{i=1}^J \mu_i p_{i,j} - \sum_{j=1}^J \mu_j = \lambda - \sum_{i=1}^J \mu_i p_{i,J+1} =: -\varepsilon(\mathcal{J})$. 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 $\sum_{i=1}^J \pi(i) p_{i,J+1} = \sum_{i=1}^J \sum_n \mathbf{P}(Y_n = i, Y_{n+1} = J+1) = \sum_n \mathbf{P}(Y_n \neq J+1, Y_{n+1} = J+1) = 1$.

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 $p^{\mathcal{I}}_{i,j}$ that correspond to the Markov chain $\{Y^{\mathcal{I}}_m\}$ being 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 $|\overline{Q}(t)| \le -\varepsilon$. Hence the fluid model is stable.

We considered multiclass networks with single-server stations.

Exercise 1. Consider a two-server FCFS queue with i.i.d. inter-arrival and i.i.d. service times queue, and introduce a fluid model for it. Then find stability conditions.

Exercise 2. More generally, study a multi-server queue.

Exercise 3. Find stability conditions for a tandem of two 2-server queues.

Exercise 4. Study a tandem of two 2-server queues with feedback: upon service completion at station 2, a customer returns to station 1 with probability p and leaves the network otherwise.

2 Inducing (second) vector field

In this section, we consider only a particular class of models: Markov chains in the positive quadrant $\mathcal{ZR}^2$. An analysis of more general models may be found, e.g., in [1, 5, 11]. We follow here [1], Chapter 7.

Let $\{X_n\}$ be a Markov chain in $\mathcal{ZR}^2$ with initial state X_0. For $(x,y) \in \mathcal{R}^2$, let a random vector $\xi_{x,y}$ have a distribution

$$\mathbf{P}(\xi_{x,y} \in \cdot) = \mathbf{P}(X_1 - X_0 \in \cdot \mid X_0 = (x,y))$$

and let

$$a_{x,y} = \mathbf{E}\xi_{x,y} \equiv (a^{(1)}_{x,y}, a^{(2)}_{x,y})$$

be a 1-step mean drift vector from point (x,y).

Assume that random variables $\{\xi_{x,y}\}$ are uniformly integrable and that a Markov chain is *asymptotically homogeneous* in the following sense: first,

$$\xi_{x,y} \to \xi \quad \text{weakly as} \quad x, y \to \infty,$$

then $a_{x,y} \to a = \mathbf{E}\xi$. Also,

$$\xi_{x,y} \to \xi_{x,\infty} \quad \text{weakly as} \quad y \to \infty, \quad \forall x,$$

then $a_{x,y} \to a_{x,\infty} = \mathbf{E}\xi_{x\infty}$; and

$$\xi_{x,y} \to \xi_{\infty,y} \quad \text{weakly as} \quad x \to \infty, \forall y,$$

then $a_{x,y} \to a_{\infty,y} = \mathbf{E}\xi_{\infty,y}$. Note also that $a_{x,\infty} \to a$ as $x \to \infty$ and $a_{\infty,y} \to a$ as $y \to \infty$.

Consider a homogeneous Markov chain $V_n^{(1)}$ on $\mathbb{R}$ with distributions of increments

$$\mathbf{P}_v(V_1^{(1)} - V_0^{(1)} \in \cdot) = \mathbf{P}(\xi^{(1)}_{\infty,v} \in \cdot)$$

and a homogeneous Markov chain $V_n^{(2)}$ on $\mathbb{R}$ with distributions of increments

$$\mathbf{P}_y(V_1^{(2)} - V_0^{(2)} \in \cdot) = \mathbf{P}(\xi^{(2)}_{\infty,v} \in \cdot)$$

We also need an extra
Assumption. For $i = 1, 2$, if $a^{(i)} < 0$, then a Markov chain $\{V_n^{(i)}\}$ converges to a stationary distribution $\pi^{(i)}$. In this case, let

$$c^{(i)} = \int_0^\infty \pi^{(i)}(dv) a^{(3-i)}(...)$$

Here (...) means (v, ∞) if $i = 1$ and (∞, v) if $i = 2$.

Theorem 2. *Assume that* $a^{(1)} \neq 0$ *and* $a^{(2)} \neq 0$*. Assume further that* $\min(a^{(1)}, a^{(2)}) < 0$ *and, for* $i = 1, 2$*, if* $a^{(i)} < 0$*, then* $c^{(i)} < 0$*. Then a Markov chain* X_n *is positive recurrent.*

Proof is omitted. We provide some intuition instead....

Example. Consider a tandem of two queues with state-dependent feedback. Assume that all driving random variables are mutually independent and have exponential distributions:
– an exogenous input is a Poisson process with parameter λ (then interarrival times are i.i.d. Exp(λ);
– service time at station $i = 1, 2$ have exponential distribution with parameter μ_i.

In addition, after a service completion at station 2, a customer returns to station 1 with probability p_{n_1,n_2} and leaves the network otherwise. Here n_i is a number of customers at station i prior the completion of service..

After doing embedding (or uniformisation), we get a discrete time Markov chains. For this Markov chain, one of three events may happen: either a new customer arrives to station 1 (with prob $\lambda/(\lambda+\mu_1+\mu_2)$) or a service is completed at station 1 (w.p. $\mu_1/(\lambda+\mu_1+\mu_2)$, this will be an artificial service if station 1 is empty) or a service is completed at station 2

(again it may be an artificial service, and if not, then a customer returns to station 1 with probability $p(\cdot,\cdot)$). Thus, only moves to some neighbouring states are possible. Given that a Markov chain is at state (i,j),
(a) if $i>0, j>0$, then

$$P((i,j),(i+1,j)) = \frac{\lambda}{\lambda+\mu_1+\mu_2}, \quad P((i,j),(i-1,j+1)) = \frac{\mu_1}{\lambda+\mu_1+\mu_2},$$
$$P((i,j),(i+1,j-1)) = \frac{\mu_2 p(i,j)}{\lambda+\mu_1+\mu_2}, \quad P((i,j),(i,j-1)) = \frac{\mu_2(1-p(i,j))}{\lambda+\mu_1+\mu_2}$$

(b) if $i>j=0$, then

$$P((i,0),(i+1,0)) = \frac{\lambda}{\lambda+\mu_1+\mu_2}, \quad P((i,0),(i-1,1)) = \frac{\mu_1}{\lambda+\mu_1+\mu_2},$$
$$P((i,0),(i,0)) = \frac{\mu_2}{\lambda+\mu_1+\mu_2},$$

(c) if $j>i=0$, then

$$P((0,j),(1,j)) = \frac{\lambda}{\lambda+\mu_1+\mu_2}, \quad P((0,j),(0,j)) = \frac{\mu_1}{\lambda+\mu_1+\mu_2},$$
$$P((0,j),(1,j-1)) = \frac{\mu_2 p(i,j)}{\lambda+\mu_1+\mu_2}, \quad P((0,j),(0,j-1)) = \frac{\mu_2(1-p(i,j))}{\lambda+\mu_1+\mu_2}$$

(d) finally, if $i=j=0$, then

$$P((0,0),(1,0)) = 1 - P((0,0),(0,0)) = \frac{\lambda}{\lambda+\mu_1+\mu_2}.$$

This is *asymptotically* and, moreover, *partially homogeneous* Markov chain.

We will consider only a particular case when probabilities $p(n_1,n_2)$ depend on n_2 only. Assume that there exists a limit $p = \lim_{n_2\to\infty} p(n_2)$.

Exercise 5. Find stability conditions in terms of λ, μ_1, and μ_2.

Exercise 6. In the example of tandem queue, consider first the case when probabilities $p(n_1, n_2)$ depend on n_1 only. Assuming an existence of limit $p = \lim_{n_2 \to \infty} p(n_2)$, find stability conditions. Consider then ageneral case when probabilities $p(n_1, n_2)$ may depend on both n_1 and n_2.

References

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

[2] Bramson, M. (1993) Instability of FIFO queueing networks with quick service times. *Ann. Appl. Proba.* **4**, 693-718.

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

[4] Dai, J.G. (1995) On positive Harris recurrence of multiclass queueing networks: a unified approach via fluid limit models. *Ann. Appl. Proba.* **5**, 49-77.

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

[6] Foss, S. and Kovalevskii, A. (1999) A stability criterion via fluid limits and its application to a polling model. *Queueing Systems* **32**, 131-168.

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

[8] Pukhal'ski A.A. and Rybko A.N. (2000) Nonergodicity of queueing networks when their fluid models are unstable. *Problemy Peredachi Informatsii* **36**, 26-46.

[9] Rybko, A.N. and Stolyar, A.L. (1992) Ergodicity of stochastic processes describing the operations of open queueing networks. *Problemy Peredachi Informatsii* **28**, 3-26.

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

[11] Zachary, S. (1995) On two-dimensional Markov chains in the positive quadrant with partial spatial homogeneity. *Markov Processes and Related Fields.* **1**, 267-280