Continuity of the M/G/c Queue

Lothar Breuer
Institute of Mathematics and Statistics
University of Kent, Canterbury, UK

November 14, 2006

Abstract

Consider an M/G/c queue with homogeneous servers and service time distribution F. It is shown that an approximation of the service time distribution F by stochastically smaller distributions, say F_n, leads to an approximation of the stationary distribution π of the original M/G/c queue by the stationary distributions π_n of the M/G/c queues with service time distributions F_n. Here all approximations are in weak convergence. The argument is based on a representation of M/G/c queues in terms of piecewise deterministic Markov processes as well as some coupling methods.

1 Introduction

Multi–server queues are typical elements of stochastic models in application areas like computer or communication networks. However, numerically feasible methods of their analysis are still to be developed. Some qualitative results, such as stability conditions or the form of the stationary distribution, can be found in Kiefer and Wolfowitz [12], Breuer [7], and Asmussen [1], chapter XII.

On the other hand, multi–server queues with Markovian service time distributions can be readily analyzed by standard methods (see e.g. Neuts [16] for the PH/M/c and the GI/PH/c queues). Steady–state waiting time distributions for this case have been computed in Asmussen and Moller [3]. Lucantoni and Ramaswami [15] provide efficient algorithms to compute the stationary distribution of the GI/PH/c queue.

The class PH of phase–type service time distributions appearing here is almost as tractable as the class of exponential distributions, yet sufficiently versatile as to be dense in the class of all distributions on the time axis (see Schassberger [19], section I.6). Statistical procedures for fitting PH distributions are given in Asmussen [4]. The versatility of the class PH gives rise to the wide–spread believe that an approximation of general multi–server queues should be possible by approximating the general service time distributions by PH distributions.

The purpose of the present paper is to validate this conjecture by a proof which is applicable to a large class of multi–server queues. Although the motivation to study continuity of multi–server queues stems mainly from the tractability of the special case having phase–type service time distributions, the proofs in this paper do not require any phase–type assumption. However, the arguments do require that the approximating service time distributions are stochastically smaller than the original one.

Previous results on approximating multi–server queues are summarized in Stoyan [20] and Kimura [13]. Rachev [18], chapter 12, contains results on the stability of single–server queues. Asmussen and Johansen [2] have proven continuity of the mean stationary waiting time for the GI/G/c queue. The present paper is concerned with continuity of the stationary distributions.

In a rough outline, the argument proceeds as follows: First the general multi–server queue is modelled as a piecewise deterministic Markov process (PDMP), using auxiliary variables for the remaining service times at each server. This leads to a Markov chain at jump epochs, for which the transition probability kernel can be derived explicitly. Its form shows that an approximation of the original service time distribution by stochastically smaller distributions yields an approximation of the transition kernel. This implies an approximation of the stationary distribution of the Markov chain embedded at jump epochs. As the transformation from the embedded stationary distribution to the stationary distribution of the queueing process is continuous and does not involve the service time distribution, the above approximation suffices to establish the result.

The paper is organized as follows: Section 2 contains a short presentation of basic notations and results for PDMPs. A representation of the M/G/c queueing process by means of PDMPs is given in section 3. Finally in section 4 it is shown how an approximation of the service time distribution implies an approximation of the stationary distribution of the queueing process.

2 Piecewise deterministic Markov processes

Piecewise deterministic Markov processes (PDMPs) are a powerful generalization of Markov jump processes. They have initially been analyzed by means of martingale theory aiming at more general models for optimal control theory. In the 1990s, Costa and Dufour [8, 11] achieved to find methods for deriving the stationary distribution of a PDMP, using either the embedded Markov chain after jump times or the special structure of its resolvent. For an extensive presentation of and a bibliography for PDMPs see Davis [10].

PDMPs are a generalization from Markov jump processes with respect to three main features. The state space now is not constrained to a countable set anymore, but will be allowed to be continuous. Second, between jumps the process is not restricted to remain constant, but may change deterministically. On the one hand this clearly is a great enhancement of modelling power, but on the other hand the fact that the moves between jumps are deterministic keeps the stochastic complexity of the process essentially on the level of a Markov jump process. Finally, the possibility of movements between jumps gives rise to a new kind of jump, namely jumps which occur immediately upon reaching a certain state. For queueing systems this usually will be the case whenever a server becomes idle and receives a new user immediately. This new kind of jump will be called intrinsic jump, since it is induced exclusively by the state of the system. The other kind of jump, as induced by Markovian arrivals, will be called an extrinsic jump.

Let $\mathcal{X} = (X_t : t \in \mathbb{R}_+)$ denote a continuous–time Markov process with a Polish state space E. Denote by $\mathcal{E}$ the σ–algebra generated by the Borel subsets of E. The process $\mathcal{X}$ shall be determined by the following characteristic representation:

- A flow $\phi : E \times \mathbb{R}^+ \to E$ on E.
- A closed set $\Delta \in \mathcal{E}$ containing the states that induce intrinsic jumps.
- A function $\lambda : E \to \mathbb{R}_+$ satisfying $\sup_{x \in E} \lambda(x) < \lambda_{\max} < \infty$. The value $\lambda(x)$ indicates the intensity of an extrinsic jump occurring if the process $\mathcal{X}$ is in state x .
- A stochastic transition measure $Q : E \times \mathcal{E}^0 \to [0, 1]$ with $\mathcal{E}^0 := \mathcal{E} \cap (E \setminus \Delta)$, describing the behaviour upon (extrinsic and intrinsic) jumps.

First define for all $x \in E^0 := E \setminus \Delta$ the deterministic variable

$$t_*(x) := \inf\{t \in \mathbb{R}^+ : \phi(x, t) \in \Delta\}$$

as the time until the set Δ is reached from a state $x \in E$. Then define the random variable $T(x)$ of the first (intrinsic or extrinsic) jump time after starting in state x. This is distributed as

$$\mathbb{P}(T(x) > t) = \begin{cases} e^{-\int_0^t \lambda(\phi(x,u))du} & t < t_*(x) \\ 0 & t \geq t_*(x) \end{cases}$$

for all $t \in \mathbb{R}_+$. We need to assume that there are only finitely many jumps of $\mathcal{X}$ in any finite interval. In the queueing application presented in section 3, this will be trivial to verify.

The PDMP $\mathcal{X}$ evolves in the following way: Starting in any state $x \in E \setminus \Delta$, it changes deterministically according to the flow Φ until it enters Δ, inducing an intrinsic jump, or an extrinsic jump occurs. Upon a jump, the state of $\mathcal{X}$ changes immediately according to the transition measure Q, leading to a state $y \in E \setminus \Delta$. Then the process starts a new cycle, behaving as described until the next jump.

Given the specification of a PDMP $\mathcal{X}$, one way to determine its stationary distribution is described in Costa [8]. Let $Z_0 = X_0$ denote the initial state and Z_n the state of $\mathcal{X}$ after the nth jump. Then $\mathcal{Z} = (Z_n : n \in \mathbb{N}_0)$ is called the Markov chain associated to $\mathcal{X}$. If $\mathcal{Z}$ has a stationary distribution π satisfying

$$\int_E \int_0^{t*(x)} e^{-\Lambda(t,x)} \, dt \, d\pi(x) < \infty$$

where $\Lambda(t,x) := \int_0^t \lambda(\phi(x,u)) \, du$, then a stationary distribution for $\mathcal{X}$ can be constructed as follows: Define the set $M := \{(x,t) \in E \times \mathbb{R}_+ : t < t_*(x)\}$ and denote the Borel σ–algebra on M by $\mathcal{M}$. For any set $A \in \mathcal{E}$ and measurable functions $t_1, t_2 : E \to [0,\infty]$ with $t_1(x) < t_2(x) \leq t^*(x)$ for all $x \in E$ define

$$B_A^{t_1,t_2} := \{(x,t) \in M : t_1(x) \leq t < t_2(x), x \in A\}$$

$$\nu_\pi(B_A^{t_1,t_2}) := \frac{\int_A \int_{t_1(x)}^{t_2(x)} e^{-\Lambda(t,x)} \, dt \, d\pi(x)}{\int_E \int_0^{t*(x)} e^{-\Lambda(t,x)} \, dt \, d\pi(x)} \tag{1}$$

By this definition ν_π can be uniquely extended to a measure on $\mathcal{M}$. Using the measurable restriction of the flow function $\phi : M \to E$ to the set M, we obtain a measure $\nu_\pi \phi^{-1}$. By theorem 2 of Costa [8] this is the stationary distribution of $\mathcal{X}$.

3 The M/G/c queue as a PDMP

Consider an M/G/c queue with the following characteristics. The Poisson input shall have rate λ. The service time distribution shall be denoted by F, being equal

for each of the c servers.

This queue can be described as a piecewise–deterministic Markov process in the following way. Define a state space $E := \mathbb{N}_0 \times \mathbb{R}_+^c$, where for $(n, x) = (n, x_1, \ldots, x_c) \in E$ the first component n represents the number of users waiting in the queue and the components x_i represent the remaining service time at the ith server. If the ith server is idle, then $x_i = 0$.

A flow function ϕ on E shall be defined by

$$\phi_t(n, x) := (n, (x_1 - t)^+, \ldots, (x_c - t)^+) \tag{2}$$

for all $(n, x) = (n, x_1, \ldots, x_c) \in E$ and $t \in \mathbb{R}_+$, with $(s - t)^+ := \max(0, s - t)$ for all $s, t \in \mathbb{R}$.

We define further for all $x = (x_1, \ldots, x_c) \in \mathbb{R}_+^c$ the value

$$t_*(x) := \begin{cases} \min\{x_i : 1 \le i \le c, x_i > 0\} & \text{for} \quad x \neq 0 \\ \infty & \text{for} \quad x = 0 \end{cases} \tag{3}$$

This denotes the time until the next server will become idle.

Differing from Davis [9] and Costa, Dufour [11], we will introduce two transition measures Q_1 and Q_2 for the jumps that can occur. This reflects the queueing process more transparently. Q_1 is the transition measure for arrivals, and thus we define for all $(n, x) = (n, x_1, \ldots, x_c) \in E$ and $A = A_1 \times \ldots \times A_c$

$$Q_1((n, x), \{m\} \times A) := \begin{cases} \delta_{m,n+1} \cdot 1_A(x) & \text{for} \quad \prod_{i=1}^{c} x_i > 0 \\ \delta_{m,n} \cdot \prod_{j \neq i} 1_{A_j}(x_j) \cdot F(A_i) & \text{for} \quad i = \min\{l : x_l = 0\} \end{cases} \tag{4}$$

Note that the latter case in the definition of Q_1 is possible only for $n = m = 0$.

The second transition measure Q_2 refers to the case of a server becoming idle. If there are any waiting users in the queue, it immediately will commence to serve a new user. Thus we have

$$Q_2((n, x), \{m\} \times A) := \begin{cases} \delta_{m,n-1} \cdot \prod_{j \neq i} 1_{A_j}(x_j) \cdot F(A_i) & \text{for} \quad n \ge 1, x_i = 0 \\ \delta_{m,n} \cdot 1_A(x) & \text{for} \quad n = 0 \end{cases} \tag{5}$$

Note that for the case $n \ge 1$, only one server can be idle at a time. Since the queue has Poisson single arrival input, the probability that two servers finish their work at the same time instant is zero. Furthermore, if one server had become idle before the other server and there had been any waiting users in the queue, it would

have commenced serving one of them. Also note that the case $n = 0$ does not correspond to a jump in the given formulation of the PDMP model. However, we could reformulate the state space, say $E = (\mathbb{N}_0 \times \mathbb{R}_+^c) \cup \bigcup_{k=1}^{c-1} \mathbb{R}_+^k \cup \{\mathbf{0}\}$, such that it suits exactly the specification in Davis [10]. Then we would observe jumps for the case $n = 0$, too. In order to simplify the presentation of the model, we chose to accept this slight inaccuracy.

In our queueing application, we have $\Lambda(x,t) = \lambda \cdot t$. The transition kernel of the embedded Markov chain $\mathcal{Z} = (Z_n : n \in \mathbb{N}_0)$ at jump times is given as

$$P((n,x),\{m\} \times A) = \int_0^{t_*(x)} e^{-\lambda s} \lambda Q_1((n, x - s \cdot \mathbf{1}), \{m\} \times A)\, ds + e^{-\lambda t_*(x)} Q_2((n, x - t_*(x) \cdot \mathbf{1}), \{m\} \times A)$$

with $\mathbf{1}$ denoting the c–dimensional column vector with all entries equal to one. It can be arranged by its first component in an $\mathbb{N}_0 \times \mathbb{N}_0$–matrix with kernel entries denoted as in

$$P(x,A) = \begin{pmatrix} P_{00}(x,A) & P_{01}(x,A) & 0 & 0 & 0 & \dots \\ P_{10}(x,A) & 0 & P_{01}(x,A) & 0 & 0 & \dots \\ 0 & P_{10}(x,A) & 0 & P_{01}(x,A) & 0 & \dots \\ 0 & 0 & P_{10}(x,A) & 0 & P_{01}(x,A) & \ddots \\ \vdots & \vdots & \ddots & \ddots & \ddots & \ddots \end{pmatrix}$$

The blockwise skip–free (or QBD) structure is due to the fact that there are only single arrivals and service completions at a jump time of $\mathcal{X}$. The blockwise Toeplitz structure follows from the product form of the jump kernels Q_1 and Q_2 in (4) and (5).

For our purposes it is enough to notice that the construction of $\nu_\pi \phi^{-1}$ at the end of section 2 does not involve the service time distribution of the queue. This suffices to prove

Theorem 1 *Let $\mathcal{X}$ and $\mathcal{X}_n$, $n \in \mathbb{N}$, denote PDMPs with embedded Markov chains $\mathcal{Z}$ and $\mathcal{Z}_n$, $n \in \mathbb{N}$, respectively. Further let π, μ and π_n, μ_n denote the stationary distributions of $\mathcal{Z}$, $\mathcal{X}$ and $\mathcal{Z}_n$, $\mathcal{X}_n$ respectively. Then weak convergence $\pi_n \to \pi$ implies weak convergence $\mu_n \to \mu$.*

Proof: Since the flow function ϕ is continuous and identical for all processes $\mathcal{X}$ and $\mathcal{X}_n$, $n \in \mathbb{N}$, it suffices by theorem 5.1 of Billingsley [5] to show that weak convergence $\pi_n \to \pi$ implies weak convergence $\nu_{\pi_n} \to \nu_\pi$. By (1) and the special

form of $\Lambda(x,t)$ we obtain for continuous and bounded functions $f : E \to \mathbb{R}$ and $g : \mathbb{R}_+ \to \mathbb{R}$

$$\int f(x)g(t)d\nu_\pi(x,t) = C^{-1} \cdot \int_E f(x) \int_0^{t^*(x)} g(t)e^{-\lambda t}\, dt\, d\pi(x)$$

with $C = \int_E \int_0^{t*(x)} e^{-\lambda t}\ dt\ d\pi(x)$. As $t_*(x)$ is continuous in x, the integrand $f(x) \int_0^{t^*(x)} g(t)e^{-\lambda t}\ dt$ is continuous in x, too. Following §8.4 in Breiman [6], this completes the proof.
□

4 The Approximation

In this section it is shown that an approximation of the service time distribution F by stochastically smaller distributions, say F_n, leads to an approximation of the stationary distribution π of the original M/G/c queue by the stationary distributions π_n of the M/G/c queues with service time distributions F_n.

Let $(F_n : n \in \mathbb{N})$ denote a sequence of distribution functions that converge weakly to F. This means that bounded and continuous functions $f : \mathbb{R}_+ \to \mathbb{R}$ satisfy

$$\int f dF_n \to \int f dF \qquad \text{as} \qquad n \to \infty$$

By the same construction as in the previous section, the M/G/c queues with service time distribution F_n lead to transition kernels P_n of the Markov chains embedded at jump times. These have the same structure as P, and the same expressions for the subkernels $P_{n;ij}$, with $i, j \in \{0, 1\}$, except for a substitution of F by F_n. Note that P_{ij} and $P_{n;ij}$ differ only for $j = 0$.

Lemma 1 The weak convergence

$$P_n(x, .) \to P(x, .) \qquad \text{as} \qquad n \to \infty$$

holds uniformly for all $x \in E$.

Proof: It suffices to show for $i \in \{0, 1\}$ and $x \in \mathbb{R}_+^c$ that $P_{n;i,0}(x, .) \to P_{i,0}(x, .)$

weakly as $n \to \infty$. Again following Breiman [6], §8.4, it suffices to verify

$$\int f_1(y_1) \dots f_c(y_c) P_{n;i,0}(x; dy_1, \dots, dy_c) \\ \to \int f_1(y_1) \dots f_c(y_c) P_{i,0}(x; dy_1, \dots, dy_c)$$

as $n \to \infty$ for all bounded and continuous functions $f_k : \mathbb{R}_+ \to \mathbb{R}$, $1 \leq k \leq c$. However, given the product form of the kernels in (4) and (5), this is an immediate consequence of the assumption that $F_n \to F$ weakly. Furthermore, the convergence is uniform in $x \in \mathbb{R}_+^c$.
□

Lemma 2 The higher order iterates of P_n converge weakly to the ones of P, i.e. for all $k \in \mathbb{N}$ and bounded and continuous functions $f : E \to \mathbb{R}$ the limit

$$\int f(y) P_n^k(x, dy) \to \int f(y) P^k(x, dy) \quad \text{as} \quad n \to \infty$$

holds uniformly for $x \in E$.

Proof: For $k = 1$ this is the statement of lemma 1. The induction step from $k-1$ to k is seen as follows. First of all abbreviate for a kernel K and a function f the function $Kf(x) := \int f(y) K(x, dy)$. For any bounded and continuous function $f : E \to \mathbb{R}$ and $x \in E$ we can write

$$P_n^k f(x) - P^k f(x) = P_n^{k-1}(P_n f(x) - Pf(x)) + (P_n^{k-1} - P^{k-1}) Pf(x) \qquad (6)$$

By proposition 4.9 in Costa and Dufour [11], the kernel P is weak Feller, which means that the function Pf is again bounded and continuous. By induction hypothesis there is an $N_1 \in \mathbb{N}$ such that the last term of the sum above satisfies

$$|P_n^{k-1} Pf(x) - P^{k-1} Pf(x)| < \varepsilon$$

for $n \geq N_1$ and uniformly in x, given any $\varepsilon > 0$. The case $k = 1$ states that there is some $N_2 \in \mathbb{N}$ such that

$$|P_n f(x) - Pf(x)| < \varepsilon$$

for $n \geq N_2$ and uniformly in x. Since P_n and thus every iterate is stochastic, the absolute value of the first term in (6) is bounded by ε, too.
□

At this point we should take a look at the periodicity of the embedded Markov chains $\mathcal{Z}$ and $\mathcal{Z}_n$. The times of jumps correspond to all arrivals and departures of the system processes. Hence $\mathcal{Z}$ and $\mathcal{Z}_n$ have period 2. The state space is partitioned as $E = E_0 \cup E_1$ with E_0 and E_1 comprising all states with even and odd numbers of users in the system, respectively. Let π', π'_n (resp. π'', π''_n) denote the stationary distributions of the embedded chains with transition matrix P^2 which are supported by E_0 (resp. E_1). Then we can write

$$\pi = \frac{1}{2}(\pi' + \pi'') = \frac{1}{2}(\pi' + \pi' P) \tag{7}$$

Since P is weak Feller, it now suffices to show $\pi'_n \to \pi'$ in weak convergence. Note that lemma 2 with $k = 2$ now yields the equivalent to lemma 1. Denote the embedded Markov chains with transition matrices P^2, P_n^2 and support E_0 by $\mathcal{Z}'$ and $\mathcal{Z}'_n$, respectively. These are aperiodic.

The approximation $F_n \to F$ of the service time distribution can be chosen in such a way that F_n is stochastically smaller than F (we write $F_n \leq_d F$, see Stoyan [20]) for all $n \in \mathbb{N}$. Writing $\mathbb{E}(F) = \int t \, dF(t)$, this implies $\mathbb{E}(F_n) \leq \mathbb{E}(F)$ such that

$$\lambda \cdot \mathbb{E}(F_n) < c \tag{8}$$

for all $n \in \mathbb{N}$. For a kernel K and a measure μ on $(E, \mathcal{E})$, denote the measure $\int K(x, .) d\mu(x)$ by μK. Let δ_α denote the Dirac measure on $\alpha := \{(0, \mathbf{0})\}$. The above condition (8) guarantees convergence

$$\|\delta_\alpha P_n^{2k} - \pi'_n\| \to 0 \quad \text{and} \quad \|\delta_\alpha P^{2k} - \pi'\| \to 0 \qquad \text{as } k \to \infty \tag{9}$$

against invariant probability measures π'_n and π' (see Orey [17], with C–set or atom α). This convergence is in total variation and thus entails weak convergence.

The next lemma compares the convergence speed of positive recurrent Markov chains with an atom. Since it may be of independent interest, too, it is formulated in slightly more general terms.

Lemma 3 Let $\mathcal{Y} = (Y_n : n \in \mathbb{N}_0)$ and $\mathcal{Y}' = (Y'_n : n \in \mathbb{N}_0)$ denote positive recurrent Markov chains with the same state space E and an atom $\alpha \in E$. Denote their stationary distributions by π and π', respectively. Further denote their stationary versions (with initial distribution π and π') by $\mathcal{Y}^s$ and $\mathcal{Y}'^s$, and the versions with initial distribution δ_α (being the Dirac measure on α) by $\mathcal{Y}^\alpha$ and $\mathcal{Y}'^\alpha$, respectively. Define coupling times T and T' by

$$T := \min\{n \in \mathbb{N} : Y_n^s = Y_n^\alpha = \alpha\} \quad \text{and} \quad T' := \min\{n \in \mathbb{N} : Y_n'^s = Y_n'^\alpha = \alpha\}$$

Let τ_α and τ'_α denote the recurrence times to state α for the chains $\mathcal{Y}^\alpha$ and $\mathcal{Y}'^\alpha$, respectively. Then $T' \leq_d T$ if $\tau'_\alpha \leq_d \tau_\alpha$.

Proof: Let $(f_k : k \in \mathbb{N})$ and $(f'_k : k \in \mathbb{N})$ denote the distribution of τ_α and τ'_α, respectively. Note that

$$\pi'(\alpha) = \left(\sum_{k=1}^{\infty} k f'_k \right)^{-1} \geq \left(\sum_{k=1}^{\infty} k f_k \right)^{-1} = \pi(\alpha)$$

This shows that for any $i, j \in \mathbb{N}$, the stochastic inequality

$$1_\alpha(Y_i^s) \leq_d 1_\alpha(Y_j'^s) \tag{10}$$

holds. Of course it does not hold in general for all pairs (i, j) simultaneously, but only for any choice of $i, j \in \mathbb{N}$. We will specify this choice below.

Now consider a coupling for τ_α and τ'_α. Denote their distribution functions by F and F', and define the quantile functions $h := F^{-1}$ and $h' := (F')^{-1}$ as generalized inverses of F and F'. By Lindvall [14], section IV.3, the inequality $\tau'_\alpha \leq_d \tau_\alpha$ means that $h'(U) \leq h(U)$ for a uniform random variable $U \sim U(0, 1)$.

By definition of T, we obtain the inclusion

$$\{T = n\} \subset \{Y_n^\alpha = \alpha\} \cap \{Y_n^s = \alpha\}$$

for every $n \in \mathbb{N}$. For every element of $\{Y_n^\alpha = \alpha\}$ there is an integer $k \leq n$ and realisations $u_1, \ldots, u_k$ of iid uniform random variables $U_1, \ldots, U_k$ such that $n = \sum_{i=1}^k h(u_i)$. This integer k indicates the number of visits to α before the visit at time n. The random variables $U_1, \ldots, U_k$ can be chosen as iid, since the successive recurrence times to α are iid themselves. The coupling between τ_α and τ'_α now implies that there is an integer

$$m := \sum_{i=1}^{k} h'(u_i) \leq n$$

such that $Y_m'^\alpha = \alpha$. Hence under this coupling we have

$$\{Y_n^\alpha = \alpha\} \subset \{Y_m'^\alpha = \alpha\}$$

for some $m \leq n$.

Choosing a coupling for (10) with $i = n$ and $j = m$ yields

$$\{Y_n^s = \alpha\} = \{1_\alpha(Y_n^s) = 1\} \subset \{1_\alpha(Y_m'^s) = 1\} = \{Y_m'^s = \alpha\}$$

Altogether we obtain

$$\{T = n\} \subset \{Y_n^\alpha = \alpha\} \cap \{Y_n^s = \alpha\} \subset \bigcup_{m=1}^{n} (\{Y_m'^\alpha = \alpha\} \cap \{Y_m'^s = \alpha\})$$
$$\subset \bigcup_{m=1}^{n} \{T' \leq m\} \subset \{T' \leq n\}$$

which shows that $T' \leq_d T$.
□

Lemma 4 The convergence $\|\delta_\alpha P_n^{2k} - \pi_n'\| \to 0$ for $k \to \infty$ is uniform in $n \in \mathbb{N}$.

Proof: Choose any index $n \in \mathbb{N}$. Let $\mathcal{X}$ denote the queueing process with respect to the service time distribution F and $\mathcal{X}_n$ the one with service time distribution F_n. Both are completely determined by an initial distribution and the sequences of inter–arrival and service times.

We couple both processes to the same probability space in the following way. Let $X_0 = X_0^{(n)} = (0, \mathbf{0})$ for all paths. Also the Poisson arrival process shall be pathwise identical for $\mathcal{X}$ and $\mathcal{X}_n$. Let S_j and $S_j^{(n)}$ denote the service time for the jth user in $\mathcal{X}$ and $\mathcal{X}_n$, respectively. According to Stoyan [20], proposition 1.2.1, the assumption $F_n \leq_d F$ implies that we can choose our common probability space such that $S_j^{(n)} \leq S_j$ for all $j \in \mathbb{N}$ and all paths. Define the function h on the state space $E = \mathbb{N}_0 \times \mathbb{R}_+^c$ by

$$h(n, x) := n + \frac{1}{c} \sum_{i=1}^{c} \frac{x_i}{x_i + 1}$$

for all $n \in \mathbb{N}_0$ and $x = (x_1, \ldots, x_c) \in \mathbb{R}_+^c$. The queue $\mathcal{X}$ is empty at time t, i.e. $X_t = (0, \mathbf{0})$, if and only if $h(X_t) = 0$. By the above construction of $\mathcal{X}$ and $\mathcal{X}_n$, we obtain $h(X_t^{(n)}) \leq h(X_t)$ for all times t and all paths.

Regarding the embedded Markov chains $\mathcal{Z}'$ and $\mathcal{Z}_n'$ with transition kernels P^2 and P_n^2, respectively, the same coupling implies that the recurrence time $\tau_\alpha^{(n)}$ to the atom $\alpha = (0, \mathbf{0})$ is pathwise (and hence stochastically) smaller in $\mathcal{Z}_n'$ than its analogue τ_α in $\mathcal{Z}'$. Denote the coupling times for $\mathcal{Z}'$ and $\mathcal{Z}_n'$ by T and T_n, respectively.

The coupling inequality for Markov chains (see Asmussen [1], chapter VII, (2.3)) states that

$$\|\delta_\alpha P_n^{2k} - \pi_n'\| \leq \mathbb{P}(T_n > k)$$

for all $k \in \mathbb{N}$. Now lemma 3 states that T_n is stochastically smaller than T, meaning that

$$\mathbb{P}(T_n > k) \leq \mathbb{P}(T > k)$$

for all $k \in \mathbb{N}$. Thus we have obtained a uniform bound for the convergence rates of all the chains $(\mathcal{Z}'_n : n \in \mathbb{N})$ starting from the initial distribution δ_α.
□

Theorem 2 *Assume that $F_n \to F$ weakly and all F_n are stochastically smaller than F. Then the stationary distributions π_n of the M/G/c queues with service time distributions F_n converge weakly to π.*

Proof: Due to equation (7) and the weak Feller property of P it suffices show weak convergence $\pi'_n \to \pi'$. Choose any $\varepsilon > 0$ and any bounded and continuous function $f : E \to \mathbb{R}$. By (9), there is a number $l_0 \in \mathbb{N}$ such that

$$|\delta_\alpha P^{2l} f - \pi' f| < \varepsilon/3$$

for all $l \geq l_0$. We write

$$\pi'_n - \pi' = \left(\pi'_n - \delta_\alpha P_n^{2l}\right) + \left(\delta_\alpha P_n^{2l} - \delta_\alpha P^{2l}\right) + \left(\delta_\alpha P^{2l} - \pi'\right)$$

Lemma 4 yields that further

$$|\delta_\alpha P_n^{2l} f - \pi'_n f| < \varepsilon/3$$

for all $l \geq l_0$ and uniformly in $n \in \mathbb{N}$. Finally, lemma 2 states that for any fixed $l \geq l_0$ there is a number $n_0 \in \mathbb{N}$ such that

$$|\delta_\alpha P_n^{2l} f - \delta_\alpha P^{2l} f| < \varepsilon/3$$

for $n \geq n_0$. Altogether this shows that $\pi'_n \to \pi'$ as $n \to \infty$ in weak convergence.
□

References

[1] S. Asmussen. *Applied Probability and Queues*. New York etc.: Springer, 2003.

[2] S. Asmussen and H. Johansen. Über eine Stetigkeitsfrage betreffend das Bedienungssystem GI/GI/s. *Elektronische Informationsverarbeitung und Kybernetik*, 22(10/11):565–570, 1986.

[3] S. Asmussen and J. Møller. Calculation of the steady state waiting time distribution in GI/PH/c and MAP/PH/c queues. *Queueing Systems*, 37:9–29, 2001.

[4] S. Asmussen, O. Nerman, and M. Olsson. Fitting phase-type distributions via the EM algorithm. *Scand. J. Stat.*, 23(4):419–441, 1996.

[5] P. Billingsley. *Convergence of probability measures*. New York etc.: John Wiley and Sons, 1968.

[6] L. Breiman. *Probability.* Philadelphia, PA: SIAM, 1968.

[7] L. Breuer. Transient and stationary distributions for the GI/G/k queue with Lebesgue–dominated inter–arrival time distribution. *Queueing Systems*, 45:47–57, 2003.

[8] O. Costa. Stationary distributions for piecewise-deterministic Markov processes. *J. Appl. Probab.*, 27(1):60–73, 1990.

[9] M. Davis. Piecewise-deterministic Markov processes: A general class of non- diffusion stochastic models. *J. R. Stat. Soc., Ser. B*, 46:353–388, 1984.

[10] M. Davis. *Markov models and optimization.* London: Chapman & Hall, 1993.

[11] F. Dufour and O. L. Costa. Stability of piecewise-deterministic Markov processes. *SIAM J. Control Optimization*, 37(5):1483–1502, 1999.

[12] J. Kiefer and J. Wolfowitz. On the theory of queues with many servers. *Trans. Am. Math. Soc.*, 78:1–18, 1955.

[13] T. Kimura. Approximations for multi–server queues: system interpolations. *Queueing Systems*, 17:347–382, 1994.

[14] T. Lindvall. *Lectures on the coupling method*. Wiley, Chichester etc., 1992.

[15] D. M. Lucantoni and V. Ramaswami. Algorithms for the Multi–Server Queue with Phase Type Service. *Stochastic Models*, 1(3):393–417, 1985.

[16] M. F. Neuts. *Matrix–Geometric Solutions in Stochastic Models*. Johns Hopkins University Press, Baltimore, 1981.

[17] S. Orey. *Lecture notes on limit theorems for Markov chain transition probabilities.* London etc.: Van Nostrand, 1971.

[18] S. Rachev. *Probability Metrics and the Stability of Stochastic Models*. Wiley, Chichester etc., 1991.

[19] R. Schassberger. *Warteschlangen.* Wien-New York: Springer-Verlag, 1973.

[20] D. Stoyan. *Comparison Methods for Queues and Other Stochastic Models*. Wiley, Chichester etc., 1983.