Strengthening Ergodicity to Geometric Ergodicity for Markov Chains

F.M. Spieksma, University of Leiden
R.L. Tweedie, Colorado State University

TECHNICAL 92/3
Strengthening Ergodicity to Geometric Ergodicity for Markov Chains

F.M. Spiessma† R.L. Tweedie‡

June 1992

Keywords
Irreducible Markov chains, Foster-Lyapunov functions, ergodicity, geometric ergodicity, queues, networks, polling systems

Abstract
In this paper we find conditions under which ergodic Markov chains are also geometrically ergodic: that is, converge to their limits geometrically quickly. We show that if the increment distributions of the chain have uniform exponential tails in an appropriate sense, then the stronger convergence rates hold, whilst if the stationary measure \( \pi \) of the chain has geometric tails then the same conclusions hold under auxiliary conditions. We give examples to show that, in particular, \( \pi \) may have geometric tails but the chain need not be geometrically ergodic. We conclude with a number of examples from queueing and network theory covered by the results, indicating the use of the results when there is a known Foster-Lyapunov function and also when the hitting times of finite sets are merely known to be finite.

†This work was commenced whilst the first author was visiting Colorado State University, with the support of Royal Oil/Shell Research and the Netherlands Organisation of Scientific Research

‡Postal address: Department of Applied Mathematics and Computer Science, University of Leiden, P.O. Box 9512, 2300RA Leiden, Netherlands

簪Postal Address: Department of Statistics, Colorado State University, Fort Collins CO 80524
1 Types of Ergodic Chain

We consider a time-homogeneous Markov chain $\mathcal{X} = \{\mathcal{X}_n, n \in \mathbb{Z}_+\}$ on a countable space $X$ with $n$-step transition matrix $P^n = (P^n(i, j))$ so that

$$P^n(i, j) = P(\mathcal{X}_n = j | \mathcal{X}_0 = i), \quad n \in \mathbb{Z}_+, i, j \in X.$$  

We will assume throughout that $\mathcal{X}$ is ergodic: that is, aperiodic, irreducible and positive recurrent. This is well-known to be equivalent to the existence of an invariant probability measure $\pi = (\pi(j))$ satisfying the equations

$$\pi(j) = \sum_k \pi(k) P(k, j), \quad j \in X \quad (1)$$

with $\pi(j) > 0$ for all $j \in X$ and with the property that

$$|P^n(k, j) - \pi(j)| \to 0, \quad n \to \infty \quad (2)$$

for every $k, j \in X$.

Our main aim in this paper is to provide auxiliary conditions on either $P$ or $\pi$ which will ensure that the chain is geometrically ergodic: that is, which will ensure that the convergence in (2) can be taken to occur geometrically quickly.

In fact this enables the strengthening of (2), not just to give better pointwise rates of convergence, but also to give sharper bounds on the distance of the $n$-step transition probabilities from the stationary measure. If we define the uniform $V$-norm of a signed matrix $A$ by

$$\|A\|_V \Delta \sup_k [V(k)]^{-1} \sum_j |A(k, j)V(j)|$$

where $V$ is a positive function on $X$ then we have not only individual geometric rates as in (2) but in fact the uniform bound

$$\|P^n(k, \cdot) - \pi(\cdot)\|_V \leq B \rho^n, \quad n \to \infty \quad (3)$$

for some $B < \infty$, some $\rho < 1$ and some $V \geq 1$: the fact that in the geometrically ergodic case there exists a function $V$ giving geometric convergence in the uniform $V$-norm has been recently proven in [6] and is called there $V$-uniform ergodicity.

This form of convergence is even more valuable than the normal form of pointwise convergence as in (2) or convergence in total variation [13]: this is discussed in Chapters 16-17 of [11]. In the cases we consider it will be usual to find $V^*$-uniform convergence for exponential functions $V^*$, and this enables us to assert that there is geometric convergence of reward functionals of the chain where the reward can be polynomial of any order, or even exponential in the state of the chain.
The key observation in this paper is that it is possible to link the solutions of "drift conditions" for ergodicity and geometric ergodicity.

It is well known (see for example [17, 11]) that ergodicity for aperiodic irreducible chains is equivalent to the existence of a solution to Foster's criterion: there exists some \( V \geq 0 \) such that for some (and then any) finite set \( C \),

\[
\begin{align*}
\sum_j P(i, j)V(j) &\leq V(i) - 1, \quad i \in C^c, \\
\sum_j P(i, j)V(j) &< \infty, \quad i \in C.
\end{align*}
\]  

(4)

One reason for this equivalence is that (see [17]) the function \( V_C(i) \triangleq E_i[\tau_C] \) is the minimal solution to (4), if we take \( V_C(i) = 0 \), for \( i \in C \). We shall in particular use this fact when the state space is \( \mathbb{Z}_+^n \) and the excluded set is \( C = \{0\} \) and denote the minimal solution in this case by

\[
V_0(i) \triangleq E_i[\tau_0], \quad i > 0; \quad V_0(0) = 0.
\]

Our techniques for strengthening ergodicity to geometric ergodicity depend, essentially, on finding conditions ensuring that, given a function \( V \) solving (4), the function \( V^*(i) = e^{\delta V(i)} \) also solves

\[
\begin{align*}
\sum_j P(i, j)V^*(j) &\leq \lambda V^*(i), \quad i \in C^c \\
\sum_j P(i, j)V^*(j) &< \infty, \quad i \in C,
\end{align*}
\]  

(5)

for some \( \delta > 0 \) and \( \lambda < 1 \).

It is perhaps less well-known that the existence of such a solution \( V^* \geq 1 \) to (5), first developed by Popov [14], is equivalent to geometric ergodicity: see [19, 6, 11] for detailed approaches to these conditions.

Once we have established (5) then it follows from [6] that the chain is in fact \( V^* \)-uniformly ergodic, with the same \( V^* \) occurring in (3) as in (5).

The conditions we consider on \( P \) require that the "right tails" of the transition probabilities are bounded exponentially with respect to the rate of growth of \( V \) in (2): essentially we rescale the space using the function \( V \) and then require exponential bounds on the "positive increments". In particular, we require for some \( \beta > 0 \) and \( c < \infty \)

\[
\sum_{j: V(j) \geq V(k)} P(k, j)e^{\beta(V(j) - V(k))} \leq c,
\]  

(6)

whilst for the "negative increments" to the "left" we require at least the variances of the transitions to be bounded: that is, for some \( d < \infty \)

\[
\sum_{j: V(j) < V(k)} P(k, j)(V(j) - V(k))^2 \leq d.
\]  

(7)
We give in Section 4 examples which indicate that some such auxiliary conditions are needed to move to geometric ergodicity from ergodicity: these show that in general conditions such as (6) or (7) cannot be avoided.

There are some results available which can be seen as precursors to our conditions. One obvious way of ensuring that $P$ has the characteristics in (6) and (7) is to require that for some $b$

$$|V(j) - V(k)| \geq b \Rightarrow P(k, j) = 0$$  \hspace{1cm} (8)

that is, the increments from any state are of bounded range, with the range allowed depending on $V$. In [10] it is shown (although rather inexplicitly, and using quite a different and less elementary approach) that under the bounded range condition (8) an ergodic chain is in effect geometrically ergodic. Our results thus extend these results substantially. For chains with a spatially homogeneous structure, such as random walks on the half-line $\mathbb{Z}_+$, the results we give here are also known (see [19]) and the methods we use are essentially based on, and extend, the random-walk concepts.

The conditions we consider on $\pi$ are similar in flavour, although this does not seem to be an approach addressed previously in the literature. The typical condition we require is that $\pi$ satisfy an analogue of (6), namely for some $\beta > 0$ and $c < \infty$

$$\sum_j \pi(j) e^{\beta V(j)} \leq c$$  \hspace{1cm} (9)

we show that under (9), together with some spatial homogeneity in the transition of the chain, (6) holds so that the results follow from the conditions on $P$.

For a special class of models a converse result can be shown as well. If a "skipfree to the left" property holds then geometric ergodicity implies that $\pi$ has "exponential" tails, and in particular in these circumstances (9) is satisfied if we take $V = V_0$, the minimal solution to (5) with $C = \{0\}$.

The paper concludes in Section 5 and Section 6 with examples of the use of the results in one-dimensional and multidimensional queueing models in both discrete and continuous time. For most of these an explicit expression for the Foster-Lyapunov function is not known. In this case our method consists of coupling arguments to show that the increments of the first hitting times of state 0 are uniformly bounded. This appears to be a novel application of coupling arguments, and illustrates the power of the proposed technique for verifying geometric ergodicity.

Many applied models satisfy the skipfree property to the left mentioned above as well, and we are thus able to strengthen the known ergodicity results to give geometric ergodicity and geometric tails on the distribution $\pi$ under appropriate conditions on service time distributions.
2 Conditions on $P$ for geometric ergodicity

The key result we prove is

Theorem 2.1 If $\Phi$ is ergodic and $V$ satisfies (4), and if $P$ satisfies (6) and (7) for some $c, d < \infty$ and $\beta > 0$ then $\Phi$ is $V^*$-uniformly ergodic, where $V^*(f) = e^{\delta V(f)}$ for some $\delta < \beta$.

Proof For positive $\delta < \beta$ we have

$$
\frac{1}{V^*(k)} \sum_j P(k, j) V^*(j) = \sum_j P(k, j) e^{\delta(V(j) - V(k))}
$$

$$
= \sum_j P(k, j) \left\{ 1 + \delta(V(j) - V(k)) + \frac{\delta^2}{2}(V(j) - V(k))^2 e^{\theta_k(V(j) - V(k))} \right\}
$$

for some $\theta_k \in [0, 1]$, by using a second order Taylor expansion. As $V$ is assumed to satisfy (4), the right hand side of (10) is bounded for $k \notin C$ by

$$
1 - \delta + \frac{\delta^2}{2} \left\{ \sum_{j: V(j) < V(k)} P(k, j)(V(j) - V(k))^2
$$

$$
+ \sum_{j: V(j) \geq V(k)} P(k, j)((V(j) - V(k))^2 e^{\delta(V(j) - V(k))}) \right\}
$$

$$
\leq 1 - \delta + \frac{\delta^2}{2} d + \frac{\delta^2 - \epsilon}{2} \sum_{j: V(j) \geq V(k)} P(k, j) e^{\delta + \epsilon / 2(V(j) - V(k))}
$$

$$
\leq 1 - \delta + \frac{\delta^2 - \epsilon}{2}(d + c')
$$

for some $\epsilon \in (0, 1)$ such that $\delta^{1+\epsilon / 2} < \beta$, and some constant $c' > 0$, by virtue of (6) and (7), and the fact that $x^2$ is bounded by $e^x$ on $\mathbb{R}_+$. This proves the theorem, since we have

$$
1 - \delta + \frac{\delta^2 - \epsilon}{2}(d + c') < 1
$$

for sufficiently small $\delta > 0$, and thus (5) holds.

Many examples have the structure of a random walk on $\mathbb{Z}_+^n$ with reflecting barriers. When in addition $V$ has bounded increments with respect to the $\ell_1$-norm $\| \cdot \|_1$ on $X$, then the result of Theorem 2.1 has a simpler formulation: since in many cases the test function $V(x) = c\|x\|_1$ is used, it seems worthwhile to give this formulation separately.
Theorem 2.2 Suppose that $X = \mathbb{Z}^n_+$, that $\Phi$ is ergodic and that $V$ satisfies (4), and for some $c > 0$

$$|V(j) - V(k)| \leq c\|j - k\|_1. \quad (12)$$

(i) If for some $b > 0$

$$\|j - k\|_1 \geq b \Rightarrow P(k, j) = 0 \quad (13)$$

then $\Phi$ is $V^*$-uniformly ergodic.

(ii) If there are finitely many different increment distributions, that is, there is a finite set $S \subset X$, such that

$$P(k, k + l) = P(s, s + l), \quad l \in \mathbb{Z}^n, \quad (14)$$

for some $s \in S$, then $\Phi$ is $V^*$-uniformly ergodic, provided that for some $\delta > 0$

$$\sum_l P(s, s + l)e^{\delta\|l\|_1} < \infty, \quad s \in S. \quad (15)$$

Proof For the proof of (i) (6) and (7) are straightforward to check.

For (ii)

$$\sum_{l: V(k + l) \geq V(k)} P(k, k + l)e^{\delta(V(k + l) - V(k))} \leq \sum_l P(k, k + l)e^{\delta\|l\|_1}$$

$$= \sum_l P(s, s + l)e^{\delta\|l\|_1},$$

for some $s \in S$ by (12) and (14). The latter expression is finite by (15), and thus (6) holds; and (7) follows analogously.

Theorem 2.2 remains valid, if instead of (14) we require the increment distributions to be stochastically decreasing in an appropriate sense together with uniformly bounded increments "to the left". However, the conditions used here, suffice for most queueing applications.

We shall see in Section 4 that some conditions such as (6) or (7) are needed for these conclusions to apply.

3 Conditions on $\pi$ for geometric ergodicity

We now extend these results to provide conditions in terms, not of $P$, but of the stationary probability measure $\pi$ itself. We do however need some spatial homogeneity in the structure of $P$ for $\pi$ to describe the rate of convergence.
For convenience we take $X = \mathbb{Z}^n$. Suppose that the chain contains a finite set $S$, and that the positive $V$-increments are comparable in the rather strong sense that there exist $\gamma$ and $c > 0$ such that for all $j$ with $V(k+j) \geq V(k)$

$$P(k, k+j)e^{\gamma(V(k+j) - V(k))} \leq cP(s, s+j)e^{\gamma(V(s+j) - V(s))}$$

(16)

for some $s \in S$. Then we have

**Theorem 3.1** Suppose that $\Phi$ is ergodic and $V$ is a solution to (4). If $\pi$ satisfies (9) for some $\beta > 0$ whilst $P$ satisfies (7) and the uniform condition (16) for some finite set $S \subset X$, then $\Phi$ is $V^*$-uniformly ergodic.

**Proof** Since (9) implies for $\gamma < \beta$

$$\infty > \sum_j \pi_j e^{\gamma V(j)} = \sum_k \pi_k \sum_j P(k, j)e^{\gamma V(j)} = \sum_k \pi_k e^{\gamma V(k)} \sum_j P(k, j)e^{\gamma(V(j) - V(k))}$$

(17)

we have in particular that

$$\sum_j P(s, s+j)e^{\gamma(V(s+j) - V(s))} < \infty$$

for all $s \in S$. Now using the bound in (16) we have for $k \in X$ and some $s \in S$

$$c^{-1} \sum_{j: V(k+j) \geq V(k)} P(k, k+j)e^{\gamma(V(k+j) - V(k))} \leq \sum_{j: V(k+j) \geq V(k)} P(s, s+j)e^{\gamma(V(s+j) - V(s))} \leq \sum_j P(s, s+j)e^{\gamma(V(s+j) - V(s))} < \infty.$$

As $S$ is finite, we thus have that (6) holds, and the result then follows from Theorem 2.1.

□

When the function $V$ has bounded increments then again the result of Theorem 3.1 can be reformulated, and this time we have a considerably simpler condition for the result to hold.

**Theorem 3.2** Suppose that $X = \mathbb{Z}^n_+$, that $\Phi$ is ergodic and $V$ satisfies (4). If $V$ satisfies (12) and $P$ satisfies (14), then when $\pi$ has exponential tails, that is

$$\sum_j \pi(j)e^{\epsilon||j||} < \infty,$$

(18)

for some $\epsilon > 0$, the chain $\Phi$ is $V^*$-uniformly ergodic.
Proof In a same manner as in the proof of Theorem 3.1 (18) is easily seen to imply (15). The result then follows from Theorem 2.2.

\[\Box\]

In this context we have a converse of Theorem 3.2, under conditions, which we shall see are frequently satisfied by models in queueing applications in particular.

**Theorem 3.3** Suppose that \( X = \mathbb{Z}_+^n \), and that \( \Phi \) is geometrically ergodic and satisfies the following "skipfree to the left" property: there is some \( n^* \in \mathbb{Z}_+ \), such that for all \( i, j \in X \)

\[
\|j\|_1 - \|i\|_1 - n^* \Rightarrow P(i, j) = 0. \tag{19}
\]

Then (18) holds for some \( \varepsilon > 0 \) and (9) holds for \( V = V_0 \) the minimal solution to (4).

**Proof** Geometric ergodicity implies that the function \( V(z, i) \triangleq \sum_{n=0}^{\infty} P_i(\tau_0 = n)z^n \) solves (5) for all sufficiently small \( z > 1 \) (see [14, 19]). Because of the left skipfree property (19),

\[
n \leq \frac{\|i\|_1}{n^*} \Rightarrow P_i(\tau_0 = n) = 0
\]

so that \( V(z, i) \geq z\|i\|_1/n^* \).

As \( \sum_i \pi(i)V(z, i) < \infty \) by virtue of Theorem 1 in [20] we have thus proved the first part of the theorem for \( e^z = \sqrt{n}/n^* \).

To show the second part, observe that the function \( f(k) = z^k \) is convex in \( k \in \mathbb{Z}_+ \), so that \( V(z, i) = E_z[\tau_0] \geq E_z[\tau_0] \) by Jensen's inequality. Since \( V_0(i) = E_z[\tau_0] \) is the minimal solution to (4) when \( \{0\} \) is the excluded finite set, the validity of (9) immediately follows. \( \Box \)

There are many chains for which the conditions in these theorems are satisfied: we give a number in Section 5 by way of illustration, but prior to that we show that the various bounds on \( P \) cannot be altogether omitted in moving from ergodicity to geometric ergodicity.

### 4 Counterexamples

It is of course well-known that exponential tails on the increments of the chain are not needed to ensure geometric ergodicity. To demonstrate this consider any chain on \( \mathbb{Z}_+ \) with the transition matrix

\[
P(j, 0) = 1 - \beta, \quad j \in \mathbb{Z}_+. \tag{20}
\]
and with arbitrary structure elsewhere.

This is a so-called "Markov matrix" and is well-known to be ergodic. Here we see directly that

\[ P_j(\tau_0 > n) = \beta^n \]

for every \( j \): hence, no matter what the tail structure of the matrix, the chain is geometrically ergodic, and indeed strongly ergodic (cf. [6]).

However, unless we have such a Markov matrix structure the various extra conditions in Theorem 2.1 and Theorem 3.1 are not unnecessary, especially when we have more spatial homogeneity in the transition law of the chain.

The following examples indicate some of the structures that are possible when we do not have the conditions of Theorem 2.1 or Theorem 3.1. In order to develop these we rely on a characterisation of geometric ergodicity in terms of the distributions of the hitting times

\[ \tau_\alpha = \min\{n > 0 : \Phi_n = \alpha\} \]

where \( \alpha \) is some specific state in \( X \).

**Proposition 4.1** If \( \Phi \) is geometrically ergodic and \( V \geq 1 \) satisfies (5) then there exists \( \lambda < 1 \) such that for all \( j \in X \), and all \( n \)

\[ P_j(\tau_\alpha > n) \leq V(j)\lambda^n \]

**Proof** Since \( V \geq 1 \) we can rewrite (5) as

\[ V(j)\lambda \geq \sum_{j \neq \alpha} P(j, k)V(k) \geq P_j(\tau_\alpha > 1) \]

iterating this gives the result for each \( n \), as is shown in [6], Proposition 2.3. \( \square \)

Thus in order to assert that a chain is not geometrically ergodic it is enough to verify that the tails of the hitting time distributions on \( \{\alpha\} \) are not uniformly geometrically bounded. We now develop two examples where such an approach is straightforward, and which illustrate the need for some control over all of the aspects of the chain that we have imposed.

**Counterexample 1**

Consider a chain on \( \mathbb{Z}_+ \) with the transition matrix

\[
\begin{align*}
P(0,j) &= \gamma_j, & j &\in \mathbb{Z}_+ \\
P(j,j) &= \beta_j, & j &\in \mathbb{Z}_+ \\
P(j,0) &= 1 - \beta_j, & j &\in \mathbb{Z}_+.
\end{align*}
\] (21)
where $\sum_j \gamma_j = 1$.

The mean return time from zero to itself is given by

$$E_0[\tau_0] = \sum_j \gamma_j [1 + (1 - \beta_j)^{-1}]$$

and the chain is thus ergodic if $\gamma_j > 0$ for all $j$ (ensuring irreducibility and aperiodicity), and

$$\sum_j \gamma_j (1 - \beta_j)^{-1} < \infty.$$  \hspace{1cm} (22)

In this example

$$P_j(\tau_0 > n) = \beta_j^n$$

and so if $\beta_j \to 1$ as $n \to \infty$, from Proposition 4.1 the chain is not geometrically ergodic regardless of the structure of the distribution $\{\gamma_j\}$, provided only that $\gamma_n \to 0$ sufficiently fast to ensure that (22) holds.

Clearly here the right hand increments are uniformly bounded in relation to $V$ for $j > 0$: but if we consider the minimal solution to (4), namely

$$V_0(j) = E_j[\tau_0] = [1 - \beta_j]^{-1}, \quad j > 0$$

then we find that

$$\sum_{j<i} P(i,j)(V_0(j) - V_0(i))^2 = P(i,0)[1 - \beta_i]^{-2} = [1 - \beta_i]^{-1} \to \infty, \quad i \to \infty.$$ 

Hence, perhaps surprisingly, the bounded variance condition (7) is necessary in this particular model for the conclusion of Theorem 2.1 to be valid.

Moreover, for this chain, since we know that $\pi(j)$ is proportional to

$$E_0[\text{Number of visits to } j \text{ before return to 0}]$$

we have

$$\pi(j) \propto \gamma_j [1 - \beta_j]^{-1}$$

and so for suitable choice of $\gamma_j$ such as

$$\gamma_j \propto \rho^j [1 - \beta_j], \quad \rho < 1,$$

we can clearly ensure that the tails of $\pi$ are geometric: but this certainly need not imply geometric ergodicity without the bounded variance condition.

Counterexample 2

We can also ensure lack of geometric ergodicity if the drift to the right is not controlled in terms of $V$, even if the ergodic property is not generated by large negative jumps: this indicates that conditions such as the uniform exponential
tails bound of (6) are needed when the chain has some translation invariance properties.

To see this we consider a chain on \( \mathbb{Z}_+ \) with the transition matrix given by, for each \( i \in \mathbb{Z}_+ \),

\[
P(0, i) = \alpha_i,
\]

\[
P(i, i - 1) = \gamma_i,
\]

\[
P(i, i + n) = [1 - \gamma_i][1 - \beta_i]^{n}, \quad n \in \mathbb{Z}_+.
\]

(23)

where \( \sum \alpha_i = 1 \) and \( \gamma_i, \beta_i \) are both less than unity for all \( i \).

Here we have ergodicity since \( V(s) = x/\varepsilon \) satisfies Foster's criterion for some \( \varepsilon \in (0, 1) \), provided \( \sum i \alpha_i < \infty \) and we choose \( \gamma_i \) sufficiently large that

\[
[1 - \gamma_i] \beta_i / [1 - \beta_i] - \gamma_i \leq -\varepsilon:
\]

this can be done if we choose

\[
\gamma_i \geq \beta_i + \varepsilon [1 - \beta_i].
\]

In this example we certainly then have bounded variances for the left tails of the increment distributions, but no uniformity in the exponential tails of the right increments.

And now if we choose \( \beta_j \to 1 \) as \( j \to \infty \) we see again from Proposition 4.1 that the chain is not geometrically ergodic: we have for any \( j \)

\[
P_j(\tau_0 > n) \geq [1 - \gamma_j][1 - \beta_j]^{n}
\]

so there is no uniform rate of convergence of these tails.

5 Models with known test functions

We next illustrate the use of these theorems through the analysis of three queueing systems, for which the Foster-Lyapunov drift functions are known explicitly.

Typically in all of our examples the key extra assumption needed to ensure geometric ergodicity is a geometric tail on the distributions involved.

Let us say that the distribution function \( G \) of a random variable on \( [0, \infty) \) is in \( \mathcal{G}(\alpha) \) if \( G \) has a Laplace-Stieltjes transform convergent in \( (-\alpha, \alpha) \): that is, if

\[
\int_0^\infty e^{st}G(dt) < \infty, \quad |s| < \alpha.
\]

We will show that for a number of queueing models the following related results hold.
(E1) \( \Phi \) is geometrically ergodic iff \( \Phi \) is ergodic and the service time distributions are in \( \mathcal{G}(\epsilon) \) for some \( \epsilon > 0 \);

(E2) the minimal solution \( V_0 \) to (4) satisfies \( \epsilon_1 ||i||_1 \leq V_0(i) \leq \epsilon_2 ||i||_1 \) for some \( \epsilon_1, \epsilon_2 > 0 \);

(E3) geometric ergodicity implies that \( \Phi \) is \( V^* \)-uniformly ergodic for \( V^*(i) = e^{V_0(i)} \), hence \( V^* \) is lower- and upperbounded by exponential functions of \( ||i||_1 \);

(E4) if the service time distributions are in \( \mathcal{G}(\alpha) \) then \( \pi \) has exponential tails as described in Theorem 3.3.

By showing \( V^* \)-uniform ergodicity where \( V^* \) is a function that increases exponentially fast in the state variable, we prove a stronger result than just geometric ergodicity. Indeed, it implies geometric convergence of all polynomial functions of the states.

The first two examples are standard. We consider the Markov chain \( N_n \) associated with the number of customers in the \( M/G/1 \) queue at the instants just after a service completion, and the embedded \( GI/M/1 \) queue \( N^*_n \) representing the number of customers in the system just before the instants of an arrival, both of which are well known to constitute Markov chains.

It is well known [5, 2] that if \( \lambda^{-1} \) is the mean inter-arrival time and \( \mu^{-1} \) is the mean service time, then for both of these chains \( \lambda < \mu \) is a necessary and sufficient condition for ergodicity.

(i) The embedded \( M/G/1 \) queue \( N_n \)

**Theorem 5.1** Let \( \Phi \) be the Markov chain \( N_n \). If the service time distributions are in \( \mathcal{G}(\epsilon) \) for some \( \epsilon > 0 \) then (E1-E4) all hold.

**Proof** It is simple to check that \( V(i) \propto i \) is a solution to (4) with \( C = \{0\} \), so that \( V \) has bounded increments and (12) applies. In fact, since the minimal solution \( V_0 \) is known to be asymptotically linear from [3], (E2) holds with no other assumptions.

Let us now assume that the service time distribution \( G \in \mathcal{G}_e \).

Clearly (14) holds by assumption for \( S = \{0, 1\} \). We will prove (15). Application of Theorem 2.2 then proves \( V^* \)-uniform ergodicity of the embedded Markov chain for \( V^*(i) = e^{\beta i} \), and some \( \beta > 0 \).

Let \( \alpha_k \) denote the probability of \( k \) arrivals within one service.

For \( l \geq -1 \) and \( k \geq 0 \) we have

\[
P(k, k + l) = a_{l+1} = \frac{1}{(l+1)!} \int_0^\infty e^{-\lambda t} (\lambda t)^{l+1} dG(t).
\]
Let \( \delta > 0 \), so that

\[
\sum_{l \geq -1} P(1, 1 + l) e^{\delta(l+1)} = \int_{0}^{\infty} \exp\{(e^{\delta} - 1) \lambda t\} dG(t)
\]

which is assumed to be finite for \((e^{\delta} - 1) \lambda < \epsilon\). A similar derivation holds for the initial state \(0 \in S\). Consequently we have (E3).

From [8] we know the converse: geometric ergodicity of the embedded \( M/G/1 \) queue is equivalent to convergence of the Laplace-Stieltjes transform of the service time distribution in a neighbourhood of 0. In fact we can also obtain the latter from geometric ergodicity by using similar arguments to those in the proof of Theorem 3.3. Thus we have (E1).

Since the embedded chain has a left skipfree property, we may now apply Theorem 3.3 to obtain (E4), which is a well-known result [2] for the embedded \( M/G/1 \) queue.

\( \square \)

(ii) The embedded \( GI/M/1 \) queue \( N_n^* \)

**Theorem 5.2** Let \( \Phi \) be the Markov chain \( N_n^* \). Then (E1-E4) all hold.

**Proof** It is again easy to check that \( V(i) \propto i \) is a solution to (4) with \( C = \{0\} \), and as before (E2) is established from [3].

In this case the easiest approach to (E1) and (E3) is to show directly the existence of \( \beta < 1 \) and positive \( \delta \), such that \( \sum_{j \geq 0} P(k, j)V^*(j)/V^*(k) \leq \beta \) for \( V^*(i) = e^{\delta i} \).

Let \( b_k \) be the probability that \( k \) customers are served between two successive arrivals. We then have

\[
\sum_{j \geq 0} P(k, j) \frac{V^*(j)}{V^*(k)} = \sum_{l=0}^{k} b_l e^{\delta(1-l)} + \sum_{l>k} b_l e^{-\delta k},
\]

which is finite for all \( k \) and decreasing in \( k \). Thus it is sufficient to show the existence of \( k^* \), such that there exists \( \delta > 0 \) for which (24) is smaller than 1. For \( \delta = 0 \) (24) equals 0; taking the derivative gives

\[
\frac{d}{d\delta} \left\{ \sum_{l=0}^{k} b_l e^{\delta(1-l)} + \sum_{l>k} b_l e^{-\delta k} \right\}_{\delta=0} = \sum_{l=0}^{k} b_l (1-l) e^{-\delta l} - ke^{-\delta(k+1)} \sum_{l>k} b_l |_{\delta=0}
\]

\[
= 1 - \sum_{l=0}^{k} lb_l - (k+1) \sum_{l>k} b_l \leq 1 - \sum_{l=0}^{k} lb_l,
\]

which converges to \( 1 - \mu/\lambda \) as \( k \to \infty \). Choose \( k^* \), such that the righthandside of (25) is negative; then there exists \( \delta > 0 \), such that (24) is smaller than 1 for this \( k^* \).
This establishes the existence of a solution $V^*$ to (5) with $C = \{0, 1, \ldots, k^*\}$, so that the $V^*$-uniform ergodicity property holds for $V^*$ an exponential function of the states. Thus we have strengthened the geometric ergodicity result in [8]. Note that the left skipfree property does not hold for this model, although $\pi$ is known to have a product form (E4) by direct calculation of the stationary distribution [2, 11].

\[\square\]

(iii) Gated-limited polling system

We next analyse a multidimensional queueing model. A detailed description of the polling system we consider here can be found in [1].

Consider a system consisting of $K$ infinite capacity queues and a single server. The system is observed at each instant the server arrives at queue 1 and then the queue lengths at the respective queues are recorded. We thus have a $K$-dimensional state description $\Phi_n = \Phi_n^k$, where $\Phi_n^k$ stands for the number of customers in queue $k$ at the server's $n^{th}$ visit to queue 1. Note that in [1] the system is modelled slightly differently, with arrivals of the server at each gate defining the times of the embedded process.

The arrival stream at queue $k$ is a Poisson stream with parameter $\lambda_k$; the amount of service given to a queue $k$ customer is drawn from a general distribution with mean $\mu_k^{-1}$. During a visit to queue $k$ the server serves $\min(x, \ell_k)$ customers, where $x$ is the number of customers present at queue $k$ at the instant the server arrives there: thus $\ell_k$ is the "gate-limit". For simplicity, in this illustration of our methods we do not assume any set-up times, but we do incorporate the walking times between gates. Thus the time needed to walk from queue $k$ to queue $k + 1$ has a general distribution with mean $d_k$.

To make the process $\Phi$ a Markov chain we assume that the sequence of service times to queue $k$ are i.i.d. random variables, as are the sequence of walking times from queue $k$ to $k + 1$. Moreover, the arrival streams, walking times and service times are assumed to be independent of each other.

**Theorem 5.3** The gated-limited model $\Phi$ described above is geometrically ergodic provided

\[1 > \rho \triangleq \sum_k \lambda_k / \mu_k\]  \hspace{1cm} (26)

\[(1 - \rho)\ell_k / \mu_k > \sum_{l=1}^{K} d_l \rho\]  \hspace{1cm} (27)

and the service-time distributions and the walking time distributions are in $G(s)$ for some fixed common $s$. The model then satisfies (E2-E4) so that for the invariant measure $\pi$ we have for some $c > 0$

\[\sum \pi(i) e^{c||i||} < \infty.\]
Proof In [1] it is shown that \( \Phi \) is ergodic for the gated-limited service discipline when \((26)\) and \((27)\) hold, by identifying a Foster-Lyapunov function that is linear in the number of customers in the respective queues: specifically \( V(i) \propto \sum_{k=1}^{K} i_{k} / \mu_{k} \) where \( i \) is a \( K \)-dimensional vector with \( k^{th} \) component \( i_{k} \), is shown to satisfy \((4)\). Consequently \( V \) satisfies \((12)\).

To apply the results in this paper, observe that for this embedded chain there are only finitely many different values of the one-step displacements, depending on whether \( \Phi_{n}^{k} \) exceeds \( \ell_{k} \) or equals \( x < \ell_{k} \), so that \((14)\) holds. To ensure validity of \((15)\), it is straightforward to check that convergence of the Laplace-Stieltjes transforms of the service-time distributions and the walking times in a common neighbourhood of \( 0 \) is sufficient to achieve this.

The conditions of Theorem 3.3 are then also satisfied for \( n^{*} = \sum_{k} \ell_{k} \), which finishes the proof.

\( \square \)

For the gated-limited version our result on geometric ergodicity appears to be new. This is essentially the extension one would seek of the single-server queue results above.

In [1] it is shown that for specific service disciplines, which do not include the gated-limited version, there is a linear solution to \((5)\) also, thus implying \( V \)-uniform ergodicity for a linear \( V \) function. However, in the case considered in [1] the chain considered is analogous to considering the process at the renewals generated by the end of cycles rather than at the end of service times: thus there is a real difference between the time-scale with relation to which those embedded Markov models converge geometrically and the time scale of our geometric convergence.

6 Models with no explicit drift function

In many cases there is no closed form expression for the Foster-Lyapunov function, and different methods have been used to characterise the ergodicity conditions. However, once ergodicity has been established, we do know that the function \( V_{0}(i) \) is a solution to \((4)\) with \( C = \{0\} \). For the subsequent queueing models we will study properties of this function without direct calculation.

Instead of studying the continuous time processes natural in all these models, we will use a discrete time Markov chain approximating the process at times \( h, 2h, \ldots \). We call this chain the \( h \)-approximation to the continuous time Markov process. This approach is no restriction, since \( V^{*} \)-uniform ergodicity of the approximating Markov chain implies the same property for the Markov process (cf. [16], Theorem 3.3).

The models we consider in this section satisfy the left skip-free property, and also satisfy \((E2)\) with \( V(i) \sim c \|i\|_{1} \) for some constant \( c \). As a consequence \( V^{*} \)
increases exponentially fast in all components of the states, and Theorem 3.3 applies to ensure that the tails of $\pi$ are exponentially decreasing as in (E4).

(i) Jackson networks

The first example of this kind is the open Jackson network with $\tau$ nodes and infinite capacity queues. The arrival stream at node $k$ is a Poisson $(\lambda_k)$ stream. A job at node $k$ receives an exponentially distributed amount of service with rate $\mu_k$. Upon finishing service the job is routed to node $l$ with probability $r_{kl}$ or he leaves the network with probability $1 - \sum_l r_{kl}$. Each job is assumed to leave the network eventually. As the service discipline we take FIFO and we allow more than one server in each node.

The chain we analyse keeps track of the number of jobs at the various nodes. The transition probabilities of the $h$-approximation are the first order approximations of the continuous time transition probabilities and they are specified as follows. Let $h < (\sum_k \lambda_k + \sum_k s_k \mu_k)^{-1}$, where $s_k$ is the number of servers at node $k$, and let $e_k$ denote the $k^{th}$ unit vector. Then the only non-zero $P(i, j)$ are

\[
\begin{align*}
P(i, i + e_k) &= \lambda_k h \\
P(i, i^- e_k + e_l) &= \min\{i_k, s_k\} \mu_k r_{kl} h \\
P(i, i - e_k) &= \min\{i_k, s_k\} \mu_k (1 - \sum_l r_{kl}) h \\
P(i, i) &= 1 - \sum_{i \neq i} P(i, l),
\end{align*}
\]

for $i \in X = \mathbb{Z}_+^\tau$. Given this discretisation, in one unit of time at most one customer can leave or enter the network or move between two nodes. This implies (13) for $b = 3$.

Let us assume ergodicity (the ergodicity conditions of the Jackson network are well-known, see [7], but we will not need them here). Then $V_0(i) = E_i[\tau_0]$ solves (4). Using coupling of sample paths similar to [16] it is easy to see that

\[0 \leq V_0(i + e_k) - V_0(i) \leq c\]

for some constant $c$, and all $i \in X$, $k = 1, 2, \ldots, K$. Hence, $V_0$ is monotonic and has bounded increments, so that (12) is satisfied. As a consequence we may apply Theorem 2.2 and we conclude that

**Theorem 6.1** The $h$-approximation of the multiserver Jackson network is $V^*$-uniformly ergodic for

\[V^*(i) = e^{E_i[\tau_0]}.

Note that the left skipfree property is satisfied for $n^* = 1$, so that $E_i[\tau_0] \geq \|i\|_1$. 


This result for the Jackson network is not entirely new. Fayolle et al. [4] construct a quadratic function solving a generalised version of Foster's criterion, but allowing only one server in each node, and this quadratic function is used to show geometric ergodicity; whilst [16] uses a different method to show essentially the same result proved here.

(ii) Single server queue with phase-type service time distribution

We next consider queues with more general service time distributions. To maintain the Markov chain structure, it is convenient to use phase-type distributions [12]: this is a class which is dense in the class of all distributions.

We first analyse a simple extension of exponential queueing models. Jobs arrive at a service facility according to a Poisson process with parameter $\lambda$. The service facility has infinite capacity and uses the FIFO service discipline. There is one server; with probability $p_k$ any job requires $k$ exponentially distributed phases of service each with mean $\nu$. Thus the mean service time $\mu^{-1} = \sum_{k=1}^{\infty} k p_k / \nu$, and $\lambda < \mu$ is known to be necessary and sufficient for ergodicity.

This process is a Markov process on the state space

$$X = \{ i = (i_1, i_2) \mid i_1, i_2 \in \mathbb{Z}_+ \}$$

where $i_1$ stands for the number of jobs in the queue and $i_2$ for the remaining number of phases of service the job in service is to get: see [12] in the finite phase case and [18] in the more general case.

As we did with the Jackson network we consider an $h$-approximation, which has the following non-zero transition probabilities

$$P(0, 0 + le_2) = \lambda p_l h,$$
$$P(i, i + e_1) = \lambda h, \quad i_1, i_2 > 0$$
$$P(i, i - e_2) = \nu h, \quad i_1 > 0, i_2 > 1$$
$$P(i, i - e_1 + le_2) = \nu p_l h, \quad i_1 > 0, i_2 = 1$$
$$P(i, i) = 1 - \sum_{j \neq i} P(i, j),$$

where $h < (\lambda + \nu)^{-1}$.

We now use a coupling argument to show for $V_0(i) = E[\tau_0]$ that

$$V_0(i + e_2) - V_0(i) = c,$$  \hspace{1cm} (28)
$$V_0(i + e_1) - V_0(i) = c' := c \sum_{l=1}^{\infty} l p_l,$$  \hspace{1cm} (29)

for some constant $c > 0$, so that $V_0(i) = c'i_1 + ci_2$ and is thus linear in both components of the state variable.
Hence, \( V_0 \) is monotonically increasing with respect to componentwise ordering and has uniformly bounded increments.

The coupling construction is standard for such models (see [15]) and clearly is ideal for our type of argument. First we generate sample paths of \( \Phi \) drawing from two i.i.d. sequences \( U^1 = \{U^1_n\}_n, U^2 = \{U^2_n\}_n \) of random variables having a homogeneous distribution on \((0, 1)\). The first one generates arrivals and phase-completions, the second one generates the amount of phases of service that will be given to a customer starting service. The procedure is as follows. If \( U^1_n \in (0, \lambda h) \) an arrival is generated in \((n h, (n + 1) h)\); if \( U^1_n \in (\lambda h, \lambda h + \nu h) \) a phase completion is generated, and otherwise nothing happens. On the other hand, if \( U^2_n \in (\sum_{i=0}^{k-1} p_i, \sum_{i=0}^k p_i) \) \( k \) phases will be given to the \( n^{th} \) job starting service. This stochastic process has the same probabilistic behaviour as \( \Phi \).

To prove (28) we compare two sample paths, say \( \phi^k = \{\phi^k_n\}_n, k = 1, 2 \), with \( \phi^1_0 = i \) and \( \phi^2_0 = i + e_2 \), generated by one realisation of \( U^1 \) and \( U^2 \). Clearly \( \phi^2_0 = \phi^1_0 + e_2 \), until the first moment that \( \phi^1 \) hits 0, say at time \( n^* \). But then \( \phi^2_{n^*} = (0, 1) \). This holds for all realisations \( \phi^1 \) and \( \phi^2 \) and we conclude that \( V_0(i + e_2) = E_{i+e_2}[\tau_0] = E_{i}[\tau_0] + E_{e_2}[\tau_0] = V_0(i) + c \), for \( c = E_{e_2}[\tau_0] \).

If \( \phi^2 \) starts in \( i + e_1 \) we have a slightly different result. In fact, \( \phi^2_{n^*} = (0, 1) \) with probability \( p_i \), so that \( V_0(i + e_2) = V_0(i) + \sum_i p_i E_{e_2}[\tau_0] = V_0(i) + \sum_i p_i \cdot lc \).

With this coupling we have enough to prove

**Theorem 6.2** The \( h \)-approximation of the \( M/PH/1 \) queue is geometrically ergodic whenever it is ergodic, provided that the phase-distribution of the service times is in \( \mathcal{G}(s) \) for some \( s > 0 \).

In particular if there are a finite number of phases ergodicity is equivalent to geometric ergodicity for the \( h \)-approximation.

Moreover, the condition that the service time distribution is in \( \mathcal{G}(s) \) is also necessary for geometric ergodicity, and in the geometrically ergodic case the invariant measure has geometric tails.

**Proof** Clearly (14) holds. Combination of (29) and (28) proves (12) and (15) if we assume that the service time distribution is in \( \mathcal{G}(s) \) for some \( s > 0 \), giving sufficiency of this condition for geometric ergodicity by virtue of Theorem 2.2. It is not hard to prove that the service time distribution being in \( \mathcal{G}(s) \) is necessary for geometric ergodicity. Indeed, \( P_{e_2}(\tau_0 = n) = 0 \) for \( n < l \), so that \( \sum_{n=1}^{\infty} P_{e_2}(\tau_0 = n) z^n \geq z^l \). If \( \Phi \) is geometrically ergodic then the last expression is finite for some \( z > 1 \) (cf. [6]). Hence

\[
\infty > \sum_{n=1}^{\infty} P_0(\tau_0 = n) z^n = z(1 - \lambda h) + z\lambda h \sum_{i=0}^{\infty} p_i \sum_{n=1}^{\infty} P_{e_2}(\tau_0 = n) z^n \\
\geq z(1 - \lambda h) + z\lambda h \sum_{i=0}^{\infty} p_i z^i
\]

and thus \( E[e^{-s\tau}] \) (which equals \( \sum_{i} p_i (\nu/s + \nu)^i \)) converges for some \( s < 0 \).
Finally we note that $\Phi$ has the left skipfree property, hence the conclusions of Theorem 3.3 hold, giving the exponential tails for the invariant measure. \qed

7 Retrial queues with $s$ servers and phase-type service

We consider as a final and more complex example a service system with $s$ servers and a finite buffer with capacity $C$. The arrival stream of jobs is a Poisson process with parameter $\lambda$. When an arriving job finds the buffer filled to capacity, it joins the "orbit". The time elapsed between two successive retrials from the orbit for entering the service centre is exponentially distributed with rate $f(j)$, when $j$ jobs are waiting in the orbit; $f(j)$ is assumed to be increasing in $j$ and upperbounded. With probability $p_k$ a job has $k$ exponentially ($\nu$) distributed phases of work to be done by a server and the service discipline is assumed to be FIFO.

As before we analyse the $h$-approximation that is associated with the number of jobs in the orbit, the buffer and with the amount of work at the servers. This process is an irreducible Markov chain on the state space

$$X = \{i = (i_1, \ldots, i_{s+2}) | i_1, \ldots, i_{s+2} \in \mathbb{Z}_+, i_2 \leq C, \exists k \geq 3 : i_k = 0 \Rightarrow i_2 = 0\},$$

with state $i = (i_1, \ldots, i_k)$ corresponding to $i_1$ jobs in the orbit, $i_2$ jobs in the buffer and $i_k$ phases of work at server $(k - 2)$, $k = 3, \ldots, s + 2$; the buffer is assumed to be empty if there is an idle server. The transition probabilities for this Markov chain are

$$P(i, i + e_1) = \lambda h, \quad i_2 = C, \quad i_k \geq 1, k \geq 3$$
$$P(i, i + e_2) = \lambda h, \quad i_2 < C, \quad i_k \geq 1, k \geq 3$$
$$P(i, i + le_{k^*}) = \frac{1}{m} \lambda p_l h, \quad i_k = 0, m = \#\{k \geq 3 | i_k = 0\}$$
$$P(i, i - e_1 + e_2) = f(i_1) h, \quad i_1 \geq 1, i_2 < C, i_k \geq 1, k \geq 3$$
$$P(i, i - e_1 + le_{k^*}) = \frac{1}{m} f(i_1) p_l h, \quad i_1 \geq 1, i_k = 0, m = \#\{k \geq 3 | i_k = 0\}$$
$$P(i, i - e_k) = \nu h, \quad i_k > 1 \lor \{i_k = 1, i_2 = 0\}, k \geq 3$$
$$P(i, i - e_2 + le_{k^*}) = \nu p_l h, \quad i_{k^*} = 1, i_2 > 0$$
$$P(i, i) = 1 - \sum_{j \neq i} P(i, j),$$

for $k^* \geq 3$ and $h < (\lambda + s\mu + \sup_j f(j))^{-1}$ the length of the time discretisation interval. A verifiable ergodicity condition for $C = 0$, $s = 1$ and $p_1 = 1$ (i.e. for exponentially distributed workloads) has been derived in [9], but in this general case it does not seem to be known.
Our approach is to assume ergodicity, so that \( V(i) = E_i[\tau_0] \) is a Foster-Lyapunov function, and then show that under appropriate conditions ergodicity always implies geometric ergodicity: even in the simpler case of exponential service times this geometric ergodicity result is new.

We shall prove

**Theorem 7.1** The \( h \)-approximation of the \( M/PH/s \) queue with retrial is geometrically ergodic whenever it is ergodic, provided that the phase-distribution of the service times is in \( \mathcal{G}(s) \) for some \( s > 0 \). Moreover, the condition that the service time distribution is in \( \mathcal{G}(s) \) is also necessary for geometric ergodicity, and in the geometrically ergodic case the invariant measure has geometric tails.

The proof of these connections occupies the remainder of the section.

First we will show that \( V_0 \) is monotonically increasing, in the sense that

**Lemma 7.2**

\[
\begin{align*}
V_0(i + e_{k^*}) & \geq V_0(i), \quad k^* \geq 1. \\
V_0(i) & \geq V_0(i - e_1 + e_{k^*}), \quad k^* \geq 2. \\
V_0(i) & \geq V_0(i - e_2 + e_{k^*}), \quad k^* \geq 3.
\end{align*}
\]

**Proof** Here we have to use a more refined coupling than in the previous model. Sample paths of a bivariate Markov chain \((\Phi^1, \Phi^2)\) are generated by drawing from three independent i.i.d. sequences \( U^1, U^2 \) and \( U^3 \) of homogeneously distributed r.v.'s on \((0,1]\), such that \( \Phi^1 \) and \( \Phi^2 \) have the same transition law as \( \Phi \).

Let \( u^s_n, \phi^k_n = (\phi^k_{n,1}, \ldots, \phi^k_{n,s+2}) \), \( n = 1, \ldots \) denote realisations of \( U^s \) and \( \Phi^k \), \( s = 1, 2, 3 \), \( k = 1, 2 \). Similar to the previous model, \( u^s_n \) determines the number of phases of work of the \( n^{th} \) job starting service, \( n = 1, 2, \ldots \), for both processes \( \Phi^1 \) and \( \Phi^2 \). Similarly the \( n^{th} \) job entering the servers will be sent to a randomly selected idle server determined by \( u^3_n \) in the following way. Suppose there are \( l \) empty servers, then the job is sent to the \( k^{th} \) of these (in sequential order), if \( u^3_n \in ((k - 1)/l, k/l] \).

Finally, \( u^1_n \) specifies phase completions, retrials and arrivals in the \( n^{th} \) time interval for the \( \Phi^k \) process, \( k = 1, 2 \). If \( u^1_n \in (0, \lambda t) \) an arrival is generated and, depending on \( \phi^k_n \), assigned to an idle server, if possible, otherwise to the buffer and if the buffer is filled to capacity, to the orbit; if \( u^1_n \in (\lambda t, \lambda t + f(j)h) \) a retrial from the orbit is generated and assigned similarly, provided \( \phi^k_{n,1} \geq j \). If \( u^1_n \in T_l := (1 - l\nu h, 1 - (l - 1)\nu h] \) a phase completion is generated at server \( l \) in
the $\Phi^1$ process and in the $\Phi^2$ process at a coupled server. Coupling of servers only depends on the state $(\phi^1_n, \phi^2_n)$ of the bivariate Markov chain, and is achieved through the labelling procedure described below.

Write $\ell_n(m)$ for the label of $m \in \{3, \ldots, s + 2\}$ at time $n$. Then $m - 2$ and $\ell_n(m) - 2$ are coupled servers at time $n$ in the $\Phi^1$ and $\Phi^2$ processes.

**Labelling procedure:**

1. Set $r = 0$, $\ell_n(m) = 0$, $m = 3, \ldots, s + 2$ and $L^1_n = L^2_n = \{3, \ldots, s + 2\}$.

2. Let $L^1_n$ be the set of labels $l \in \{3, \ldots, s + 2\}$ that have not been assigned yet, and $L^1_n = \{m \geq 3 \mid \ell_n(m) = 0\}$ corresponds to the set of unlabelled servers. If $\phi^1_{n,m} > \phi^2_{n,l}$, $\forall m \in L^1_n$, $l \in L^2_n$, go to 5; otherwise go to the next step.

3. For $m = 3, \ldots, s + 2$ do: if $m \in L^1_n$ choose the smallest value of $l \in L^2_n$, such that $\phi^1_{n,m} = \phi^2_{n,l} - r$; set $\ell_n(m) := l$, $L^1_n := L^1_n \setminus \{m\}$ and $L^2_n := L^2_n \setminus \{l\}$.


5. Assign the $l$th smallest unassigned label (in $L^2_n$) to the $l$th smallest number in $L^1_n$, $\forall l$. Set $L_n := L^1_n$ and stop.

Note that the procedure always generates sets $L_n$ with minimum cardinality. This set corresponds to the servers that have more remaining phases of service in the $\Phi^1$ process than their counterparts in the $\Phi^2$ process. As we shall see below, the procedure is devised to bar jobs in the $\Phi^1$ process from being overtaken by "corresponding" jobs in the $\Phi^2$ process (which is meant to be the slower process). To this end we list some properties of the labelling procedure.

Let $v^1, v^2 \in \mathbb{R}^{s+2}$. In the labelling procedure replace $\phi^k_n$ by $v^k$, and write $\ell$ for the labelling obtained from the procedure and $L(v^1, v^2)$ for the set $\{k \geq 3 \mid v^k_1 > v^k_2 \}$. For an arbitrary labelling $\ell'$ we define the set $L'$ analogously, that is $L' = \{k \geq 3 \mid v^k_1 > v^k_2 \}$. The following properties are easily checked by induction.

**L.1.** If $L(v^1, v^2) = \{i_1, \ldots, i_r\}$, then there is a subset of $L'$, say $\{j_1, \ldots, j_r\}$ with $v^1_{i_l} \leq v^1_{j_l}$, $l = 1, \ldots, r$. We write $v^1_{L(v^1, v^2)} \preceq v^1_{L'}$ to denote this relation. Note that the relation implies $|L(v^1, v^2)| \leq |L'|$. Thus replacing an element in $L'$ by an element from $L(v^1, v^2)$ gives a lower value of the corresponding $v^1$ component.

**L.2.** Let $\delta^1 = v^1 + c_1 e_l$, $c_1 \geq 0$. Then $v^1_{L(v^1, v^2)} \preceq v^1_{L \cup \{l\}}$.

**L.3.** Let $\delta^2 = v^2 + c_2 e_{l'(v^2)}$, $c_2 \geq 0$. Suppose, that there is an $l_1 \in L'$ with $v^1_{i_1} \leq \delta^2_{L'(v^2)}$, and, if $l' \notin L'$, an $l_2 \in L'$ with $v^1_{l'} \leq v^2_{L'(l_2)}$. Then $v^1_{L(v^1, v^2)} \preceq v^1_{L' \setminus \{l_1\}}$. 

L4. Let \( \theta^1 = v^1 + c_1 e_1, \theta^2 = v^2 + c_2 e_2 \), \( c_1, c_2 \geq 0 \). Suppose, that there is an \( l_1 \in L^i \), with \( \theta^1_{l_1} \leq \theta^2_{l_1} \). Then \( \theta^1_{L}(s_1, s_2) \leq \theta^2_{L}(s_1, u_1 \cup (l_1 \cup (L_1 \cup (L_1)\).

In fact L2, L3 and L4 follow directly from L1, but are explicitly stated for easy reference. They deal with the cases of a job starting service in the \( \Phi^1 \) process and not in the \( \Phi^2 \) process (L1), the reverse (L2), and of a job starting service at a pair of coupled servers (L3).

Next denote by \( s^k_n \) the total number of jobs that started service in the first \( n-1 \) time intervals in the \( \Phi^k \) process, hence \( s^k_0 = 0 \).

First we consider initial conditions \( i \) and \( i + e_{k^*} \), for some \( k^* \geq 3 \), so there is one phase of workload extra in the \( \Phi^2 \) process at server \( k^* - 2 \) and the total number of jobs in the buffer and the orbit together is the same for both processes. Coupling of the external arrivals in both processes then clearly implies

\[
\sum_{k=1,2} \phi^1_{n,k} + s^1_n = \sum_{k=1,2} \phi^2_{n,k} + s^2_n, n = 0, 1, 2 \ldots \tag{33}
\]

We claim that for all time instants \( n \), the following holds.

1. \( s^1_n \geq s^2_n \),
2. \( \phi^1_{n,1} \leq \phi^2_{n,1} \).
3. \( \phi^1_{n/L_n} \leq u^2_{L_n} \), for \( S_n := \{s^2_n + 1, \ldots, s^1_n\} \), and \( L_n = \emptyset \), if \( s^1_n = s^2_n \).

where the relation \( \leq \) in (34.3) is defined similarly to L1.

Here \( |S_n| \) represents the number of jobs present at time \( n \), that started service in the \( \Phi^1 \) process, but did not yet in the \( \Phi^2 \) process. The inequality (34.1) then means that serving a job starts earlier in the \( \Phi^1 \) process; (34.2) means that in the \( \Phi^1 \) process jobs arrive earlier at the service centre and (34.3) essentially states that the remaining workload of jobs that started service before time \( n \) in both processes is less in the \( \Phi^1 \) process. In particular, if \( m \in L_n \), then the job at server \( l_n(m) - 2 \) in the \( \Phi^2 \) process has already finished service in the \( \Phi^1 \) process. Together (33) and (34) obviously imply that the \( \Phi^2 \) process can never "empty" earlier than the \( \Phi^1 \) process, so that (30) holds for \( k^* \geq 3 \).

The claim (34) is proved inductively. For \( n = 1 \) it holds trivially.

Our procedure for labelling servers ensures that server \( l \) has label \( l + 2 \) (that is, he is coupled to server \( l \) in the \( \Phi^2 \) process). Consequently \( \phi^2_n = \phi^1_n + e_{k^*} \), until the first moment, say \( n^* \), that \( \phi^1_{n^*, k^*} = 1, \phi^2_{n^*, k^*} = 2 \) and \( u^1_{n^*} \in I_{k^* - 2} \), that is, there is a phase completion at server \( k^* - 2 \). Notice that for \( n \leq n^* \) the coupled servers are equal, \( s^1_n = s^2_n, S_n = L_n = \emptyset \) and \( \phi^1_{n,k} = \phi^2_{n,k}, k = 1, 2 \). Hence (34) is true for \( n \leq n^* \).

If \( \phi^1_{n^*,2} = 0 \), then \( \phi^1_{n^*,1,k^*} = 0, \phi^2_{n^*,1,k^*} = 1 \). There is no relabelling of servers and (34) remains true for \( n = n^* + 1 \). However, if \( \phi^1_{n^*,2} > 0 \), different things
happen in both systems: \( \phi_{n+1,k}^1 = u_{n+1,k}^2 \), \( \phi_{n+1,2}^1 = \phi_{n,2}^1 - 1 \), \( \phi_{n+1,k}^2 = 1 \), \( \phi_{n+1,2}^2 = \phi_{n,2}^2 \), \( s_{n+1}^1 - s_{n+1}^2 = 1 \) and \( S_{n+1} = \{ s_{n+1}^1 \} \). Application of L2 then yields (34) for \( \nu^1 = \phi_{n,1}^1, \nu^2 = \phi_{n,2}^2 \), \( l = k \), \( L' = L_n \), \( \emptyset \) on \( l' = l_n \).

Next assume that (34) holds for \( n \leq m \). We will prove that it is valid for \( n = m + 1 \). Suppose an arrival is generated and there are idle servers in both processes: that is the arrival is sent to server \( l - 2 \) in the \( \Phi^1 \) process and to the server labelled \( \ell_{m}(l') - 2 \) in the \( \Phi^2 \) process. Then \( s_{m+1}^1 - s_{m+1}^2 = s_{m}^1 - s_{m}^2 \) and we need to consider several cases.

Suppose \( s_{m}^1 > s_{m}^2 \). Consequently \( S_{m} \neq \emptyset \) and \( S_{m+1} = S_{m} \cup \{ s_{m+1}^1 \} - \{ s_{m}^2 + 1 \} \). Furthermore, assume there is some \( k \in L_m \), such that \( \phi_{m,k}^1 \leq \nu_{m+1}^2 \). Then choosing such a \( k \in L_m \), but with the largest value \( \phi_{m,k}^1 \),

\[
\phi_{m,L_m-k}^1 \leq \nu_{m,L_m-s_{m+1}}^2.
\]

Notice that \( \phi_{m,k}^1 = 0 \), if \( l' \notin L_m \). Apply L3 for \( \nu^1 = \phi_{m,k}^1, \nu^2 = \phi_{m+1}^2 \), \( l = k \), \( L' = L_{m} \), then

\[
\phi_{m,L(L_m,L_{m+1})}^1 \leq \phi_{m,L(L_m,k)}^1.
\]

Similarly we obtain

\[
\phi_{m+1,L_{m+1}}^1 \leq \phi_{m+1,L(L_{m+1},L_{m+1})}^1.
\]

by using L2 and (34.3) follows by combining these inequalities, since

\[
\phi_{m+1,L(L_m,L_{m+1})}^1 = \phi_{m,L(L_m,L_{m+1})}^1\text{--}(l')
\]

and \( \phi_{m+1,l}^1 = \nu_{m+1,s_{m+1}}^2 \). This is the basic argument that will be applied more than once in the remainder of the proof.

If \( s_{m}^1 > s_{m}^2 \), but there is no \( k \in L_m \) with \( \phi_{m,k}^1 \leq \nu_{m+1,s_{m+1}}^2 \), then necessarily \( |L_m| < s_{m}^1 - s_{m}^2 \) and \( \phi_{m,L_m}^1 \leq \nu_{m,L_m-s_{m+1}}^2 \). Similar arguments apply to get the result.

If \( s_{m}^1 = s_{m}^2 \), then \( L_{m+1} = \emptyset \) and there is an idle server in the \( \Phi^1 \) process for every idle server in the \( \Phi^2 \) process. This yields \( L_{m+1} = \emptyset \) and (34) goes through for \( m + 1 \) similarly as before.

It is straightforward to check that the arguments concerning the validity of (34) will not change for any combination of arrival or retrial assignments to different parts of the network in the two processes (as for example happens, if an arrival is assigned to the queue in the \( \Phi^1 \) process and to the servers in the \( \Phi^2 \) process, because of all servers in the \( \Phi^1 \) process being occupied). Note, that (34.2) for \( m \) implies that a retrial in the \( \Phi^1 \) process always induces a retrial in the \( \Phi^2 \) process in the \( m^{th} \) time interval. For the induction step it is crucial that the buffer is empty for all states with an idle server.
Suppose next that a phase completion at server $k - 2$ in the $\Phi^1$ process and at server $L_m(k) - 2$ in the $\Phi^2$ process is generated. If both have at least two remaining phases of work, we may apply L1 to obtain (34).

The only interesting case remaining is the case of a service completion in the $\Phi^2$ process only, whilst the coupled server in the $\Phi^1$ process is not idle, that is when $\phi_{m,k}^1 > 1$, $\phi_{m,L_m(k)}^2 = 1$. Clearly, if $\phi_{m,2}^2 = 0$ nothing happens, so we assume $\phi_{m,2}^2 > 0$. Then $s_{m+1}^2 = s_m^2 + 1$ and $s_{m+1}^1 = s_m^1$. This case cannot occur if $s_m^1 = s_m^2$, by virtue of (34.3), since $k \in L_m$ necessarily and thus $L_m \neq \emptyset$. If $s_m^1 > s_m^2$, then $\phi_{m+1,L_m(k)}^2 = u_{s_m+1}^2 \geq \phi_{m,k}^1$, so that $L' = \{ l \geq 3 | \phi_{m+1,l}^1 \geq \phi_{m+1,L_m(l)}^1 \} = \emptyset$ and thus $L_{m+1} = \emptyset$ by L1. As $s_{m+1}^1 = s_{m+1}^2$, (34) holds for $m + 1$. If $s_m^1 = s_m^2$, then $S_{m+1} = S_m - \{ s_m^2 + 1 \}$ and we have to distinguish the cases that there is some (appropriately chosen) $l \in L_m$ with $\phi_{m,l}^1 \leq u_{s_{m}^2}^2 + 1$ or not. As in the foregoing we obtain in the first case

\[ \phi_{m+1,L_m+1}^1 \leq \phi_{m+1,L_m-1}^1 \leq \phi_{m,L_m-1}^1 \leq u_{S_m-1}^2, \]

by L3 and because $\phi_{m+1,l}^1 \leq \phi_{m,l}^1$. In the second case $\phi_{m+1,L_m}^1 \leq u_{S_m-1}^2$, and (34.3) follows similarly. Clearly, (34.2) is trivial and (34.1) holds because of $s_m^1 > s_m^2$. Hence (34) holds for $m + 1$.

If there is a service completion in both processes, then $k \not\in L_m$. The different combinations of jobs entering service during the $m^{th}$ time interval in one or both processes are handled similarly to the foregoing analyses. The arguments are similar, whenever there is a service completion in the $\Phi^1$ process and none in the $\Phi^2$ process, or when there is a service or phase completion at only one of a pair of coupled servers, because of the other one being idle already. Consequently (34) holds for $m + 1$. This completes the proof of (30) for $k^* \geq 3$.

Using the same coupling method, it is obvious that the case $k^* = 1$ (one more job in the orbit) is handled similarly to the case $k^* = 2$ (one more job in the queue). We will analyse the latter, i.e. $\Phi^1$ and $\Phi^2$ have initial conditions $i$ and $i + e_2$.

One adaptation of the previous coupling mechanism is needed to avoid the $n^{th}$ job served in the $\Phi^1$ process to be overtaken by the $n^{th}$ job in the $\Phi^2$ process, which might occur because of an extra job already being present in the $\Phi^2$ process. Similarly as before, $u_n^2$ generates the total amount of work of the $n^{th}$ job entering the servers in the $\Phi^2$ process, but of the $(n-1)^{th}$ job in the $\Phi^1$ process instead. Thus we pretend that the extra job in the $\Phi^2$ process has already left the $\Phi^1$ process before time $n = 1$, and we force the $\Phi^2$ process to be “slower” than the $\Phi^1$ process. This is a well-known procedure. Note that $u_n^2$ is only used for the $\Phi^2$ process. Defining $s_1^1 = 1 + s_1^2 = 1$, (33) holds and the proof reduces to the same analysis as in the case of the $\Phi^2$ process starting with one extra phase of workload. Consequently, (34) holds for $n \in \mathbb{Z}$ and we have proven (30).
The more stringent assertions (31) and (32) follow in the same way as the cases
\(k^* = 1, 2\) in (30) by observing that a job always has at least one phase of
workload. This shows the monotonicity of \(V_0\) and we have completed the proof
of Lemma 7.2.

Secondly, and rather more easily, we need \(V_0\) to have bounded increments.

**Lemma 7.3** \(V_0(i + e_{k^*}) \leq V_0(i) + c, \ k^* = 1, 2, \ldots, s + 2.\)

**Proof** Let \(k^* = 2\), and couple the processes with initial conditions \(i\) and \(i + e_2\),
such that the \(n^\text{th}\) job entering the servers gets \(u_{n}^2\) phases of service time in both
processes. The first time of interest, say \(n^*,\) is, when \(d_{n^*,2}^1 = 0, \ d_{n^*,2}^2 = 1,
\phi_{n^*,l}^1 = \phi_{n^*,l}^2 = 1\) and a phase completion at servers \(l - 2\) and \(L_{n^*}(l) - 2\)
is generated (a similar situation occurs when there is a retrial from the orbit
in the \(\Phi^2\) process only and servers \(l - 2\) and \(L_{n^*}(l) - 2\) are idle). In this case
we make the \(\Phi^2\) process slower in reaching state 0 by stopping to serve server
\(L_{n^*}(l) - 2\) after time \(n^*\) until the next arrival at the service centre, which then
occurs in both processes. Obviously, this has no influence on the evolution of the \(\Phi^1\)
process, since server \(l - 2\) is idle until the next arrival. Then we are in the
same situation as before. There is no problem if in the meantime another pair
of coupled servers has become idle; only some relabelling might be necessary if
for example the arrivals are assigned to server \(L_{n^*}(l) - 2\) in the \(\Phi^2\) process and
to an idle server other than \(l - 2\) in the \(\Phi^1\) process. Lemma 7.2 then implies
that the modified \(\Phi^2\) process does not drift more quickly to 0 than the original
one.

The procedure can be repeated indefinitely. Once the \(\Phi^1\) process is empty, the
slowed down version of the \(\Phi^2\) process is in state \(e_1, e_2\) or in \(e_{c_2}\) for some\n\(c > 0, l \geq 3.\) However, \(c\) equals \(m\) with probability \(p_m,\) since it is independent
of the \(\Phi^1\) process (if the \(\Phi^1\) process "empties" at time \(n,\) then the number of
remaining phases of workload in the \(\Phi^2\) process is \(u_{n+1}^2,\) with \(s^2_{n+1} = s^1_n + 1)\).

Consequently,

\[
V_0(i + e_2) \leq V_0(i) + \max\{V_0(e_1), V_0(e_2), \sum_m p_m V_0(me_l), l = 3, \ldots, s + 2\},
\]

which proves Lemma 7.3 for \(k^* = 2.\) The case \(k^* = 1\) is handled analogously.

Suppose that \(k^* \geq 3:\) Once the processes with initial conditions \(i\) and \(i + e_{k^*}\)
have entered states \(\phi_{n^*,k^*}^1 = 1, \phi_{n^*,k^*}^2 = 2, \phi_{n^*,2}^1 > 0\) and a phase completion at
server \(k^* - 2\) is drawn at time \(n^*,\) we slow down the \(\Phi^2\) process by giving server
\((k^* - 2) u_{n+1}^2\) phases of work. Then we are in the situation of \(k^* = 2\) and we
can proceed as before. If \(\phi_{n^*,2}^1 = 0(= \phi_{n^*,2}^2),\) then the first moment different
tings happen in the two processes is, when an arrival or retrial occurs and the
job at server $k^* - 2$ didn't finish service yet (if he did, the two processes evolve together), but as in the foregoing we will then be in the situation of $k^* = 1$ or $k^* = 2$. This proves Lemma 7.3.

To prove geometric ergodicity, assuming that $\Phi$ is ergodic, we first assume that $\sup_j f(j)$ is achieved for some $j^*$. Then we may apply Theorem 2.2 in conjunction with Lemma 7.3. This is standard by now. However, to do so, we need the service time distribution to lie in $G(s)$ for some $s > 0$. Because of $\Phi$ satisfying the left skipfree property, it can then be shown that this is also a necessary condition for geometric ergodicity, similar to the proof for the single server phase-type queue. Moreover the conclusion in Theorem 3.3 applies.

If $f^* \triangleq \sup_j f(j)$ is not achieved, then substituting this value $f$ in the transition probabilities it is easy to show that

$$\sup_{i \in X} \sum_{j \neq i} P(i, i + l) e^{\delta \|i||l\|} < \infty.$$ 

The conclusion in Theorem 2.2 then follows from Lemma 7.3 if the service time distributions lie in $G(s)$ for some $s > 0$. The remainder of the proof is accordingly and we have thus shown that Theorem 7.1 holds.

It seems that geometric ergodicity of the workload process also follows, if we use this special coupling of servers described in the labelling procedure. We conjecture that it applies to a whole class of models, including the Jackson network with phase-type service time distributions and more than one customer class, when analysing geometric ergodicity of the Markov process associated with the number of customers in each station and the remaining service times, or the workload process.

Acknowledgements

We are grateful to Wojtek Szpankowski and Arie Hordijk for discussions on this problem.

References


