Queueing Syst (2009) 63: 241–252
DOI 10.1007/s11134-009-9134-9

On the speed of convergence to stationarity of the Erlang loss system

Erik A.van Doorn · Alexander I. Zeifman

Received: 1 December 2008 / Revised: 15 July 2009 / Published online: 14 August 2009

Abstract We consider the *Erlang loss system*, characterized by N servers, Poisson arrivals and exponential service times, and allow the arrival rate to be a function of N. We discuss representations and bounds for the rate of convergence to stationarity of the number of customers in the system, and display some bounds for the total variation distance between the time-dependent and stationary distributions. We also pay attention to time-dependent rates.

Keywords Charlier polynomials · Rate of convergence · Total variation distance

Mathematics Subject Classification (2000) Primary 60K25 · Secondary 90B22

1 Introduction

We consider the $M/M/N/N$ service system, characterized by Poisson arrivals, exponential service times, and $N \geq 1$ servers but no waiting room. The system is also known as the *Erlang loss system* after A.K. Erlang who was the first to analyse the model in [5]. We allow the arrival rate $\lambda \equiv \lambda(N)$ to be a function of N. With μ denoting the service rate per server, the number of customers in this system is a birth-death process $\mathcal{X} \equiv \{X(t),\ t \geq 0\}$ taking values in $S := \{0, 1, \ldots, N\}$, with birth and death

E.A.van Doorn (✉)
Department of Applied Mathematics, University of Twente, P.O. Box 217, 7500 AE Enschede,
The Netherlands
e-mail: e.a.vandoorn@utwente.nl

A.I. Zeifman
Institute of Informatics Problems RAS, VSCC CEMI RAS, and Vologda State Pedagogical
University, S. Orlova 6, Vologda, Russia
e-mail: zai@uni-vologda.ac.ru

rates

$$\lambda_j = \lambda, \quad 0 \le j < N, \quad \text{and} \quad \mu_j = j\mu, \quad 0 < j \le N,$$

respectively. We write $p_j(t) \equiv \Pr\{X(t) = j\}$, $j \in S$, and let the vector $\boldsymbol{p}(t) \equiv (p_0(t), p_1(t), \ldots, p_N(t))$ represent the state distribution at time $t \ge 0$. The stationary distribution of $\mathcal{X}$ is a truncated Poisson distribution, represented by the vector $\boldsymbol{\pi} \equiv (\pi_0, \pi_1, \ldots, \pi_N)$, where

$$\pi_j := c\frac{(\lambda/\mu)^j}{j!}, \quad j \in S,$$

and c is a normalizing constant. For any initial distribution $\boldsymbol{p}(0)$ the vector $\boldsymbol{p}(t)$ converges to $\boldsymbol{\pi}$ as $t \to \infty$.

In what follows we will be interested in the behaviour of

$$d(t) \equiv d_{tv}\big(\boldsymbol{p}(t), \boldsymbol{\pi}\big) := \sup_{A \subset S}\left\{\left|\sum_{j \in A} p_j(t) - \sum_{j \in A} \pi_j\right|\right\},$$

the *total variation distance* between $\boldsymbol{p}(t)$ and $\boldsymbol{\pi}$, and more specifically in

$$\beta := \sup\big\{a > 0 : d(t) = \mathcal{O}\big(e^{-at}\big) \text{ as } t \to \infty \text{ for all } \boldsymbol{p}(0)\big\}, \tag{1}$$

the *rate of convergence of* $\boldsymbol{p}(t)$ *to* $\boldsymbol{\pi}$, also known as the rate of convergence (or *decay parameter*) of $\mathcal{X}$, and the asymptotic behaviour of $\beta \equiv \beta(N)$ as $N \to \infty$. It is well known (and easy to see) that

$$d(t) = \frac{1}{2}\sum_{j \in S}\big|p_j(t) - \pi_j\big|, \quad t \ge 0, \tag{2}$$

so the total variation distance between $\boldsymbol{p}(t)$ and $\boldsymbol{\pi}$ is essentially equivalent to the L^1-norm of $\boldsymbol{p}(t) - \boldsymbol{\pi}$.

The plan of the paper is as follows. We give a survey of representations and bounds for β in Sect. 2, and discuss asymptotic results for β as $N \to \infty$ in Sect. 3. Some upper bounds on $d(t)$ will subsequently be displayed in Sect. 4. Finally, in Sect. 5 we describe some generalizations of the preceding results to the Erlang loss model with time-dependent rates. As an aside we note that the total variation distance between $\boldsymbol{p}(t)$ and $\boldsymbol{\pi}$ may exhibit very interesting behaviour if t and N tend to infinity simultaneously. A discussion of these issues is outside the scope of this paper (but see, for example, [6, 21] and [20]).

In what follows $\mathbf{0}$ and $\mathbf{1}$ denote row vectors of zeros and ones, respectively, inequality for vectors indicates elementwise inequality, and superscript ^{T} denotes transpose.

2 Representations and bounds for β

It is well known that the supremum in (1) is in fact a maximum, and that $-\beta$ can be identified with the largest nonzero eigenvalue of

$$Q := \begin{pmatrix} -\lambda & \lambda & 0 & \cdots & 0 & 0 \\ \mu & -(\lambda+\mu) & \lambda & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & -(\lambda+(N-1)\mu) & \lambda \\ 0 & 0 & 0 & \cdots & N\mu & -N\mu \end{pmatrix}, \tag{3}$$

the q-matrix of $\mathcal{X}$. It has also been observed (see Riordan [16, p. 84] or Kijima [12]), that the nonzero eigenvalues of $-Q$ can be identified with the zeros of the polynomial

$$c_N\left(\frac{x}{\mu}-1, \frac{\lambda}{\mu}\right),$$

where

$$c_n(x,a) := \sum_{m=0}^{n} (-1)^m \binom{n}{m}\binom{x}{m}\frac{m!}{a^m}, \quad n \geq 0, \tag{4}$$

are the *Charlier polynomials* (see, for example, Chihara [4, Sect. VI.1]). Since the zeros of a Charlier polynomial are real (and positive), we have the following representation for β.

Theorem 1 *The rate of convergence β of the Erlang loss model with N servers, arrival rate λ and service rate μ per server, is given by*

$$\beta = \mu + \mu\xi_{N,1}\left(\frac{\lambda}{\mu}\right), \tag{5}$$

where $\xi_{N,1}(a)$ denotes the smallest zero of the Charlier polynomial $c_N(x,a)$.

Remark Exploiting Karlin and McGregor's [11] representation for the transition probabilities of a birth-death process, it was shown in [2] that β can be identified with the smallest zero of the polynomial

$$S(x) := \frac{\lambda}{x}\left\{c_{N+1}\left(\frac{x}{\mu}, \frac{\lambda}{\mu}\right) - c_N\left(\frac{x}{\mu}, \frac{\lambda}{\mu}\right)\right\}.$$

But since Charlier polynomials satisfy the recurrence relation

$$ac_{n+1}(x,a) - ac_n(x,a) + xc_n(x-1,a) = 0, \quad n \geq 0,$$

we can actually write

$$S(x) = -c_N\left(\frac{x}{\mu} - 1, \frac{\lambda}{\mu}\right),$$

in accordance with the previous result.

No explicit expression for $\xi_{N,1}(a)$ seems to be available for general a, but efficient algorithms for the numerical evaluation of $\xi_{N,1}(a)$—and hence of β—have been proposed (see, for example, [12]).

Charlier polynomials being orthogonal with respect to a measure consisting of point masses at the points $0, 1, \dots,$ it follows from the theory of orthogonal polynomials (see [4, Chap. 2]) that the ith smallest zero of $c_N(x,a)$ is larger than $i-1$, for $i = 1, 2, \dots$. This leads to some simple bounds for β. First, we must have $\xi_{N,1}(a) > 0$, and hence

$$\beta > \mu. \tag{6}$$

Then, since $c_N(x,a) = c_x(N,a)$ for natural x, we have $c_N(1,N) = c_1(N,N) = 0$. So, the second smallest zero of $c_N(x,N)$ being larger than 1, we must have $\xi_{N,1}(N) = 1$, and hence, for all $N \geq 1$,

$$\lambda = \mu N \quad\Longrightarrow\quad \beta = 2\mu.$$

Since $\xi_{N,1}(a) > 0$ is strictly increasing in a (see, for example, [12]), it follows that

$$\beta \mathrel{\substack{<\\=\\>}} 2\mu \quad\Longleftrightarrow\quad \lambda \mathrel{\substack{<\\=\\>}} \mu N. \tag{7}$$

Further upper and lower bounds have been derived in the literature. Specifically, translating (part of) the Theorems 3 and 5 of Krasikov [14] in terms of β by means of (5), we get the following results.

Theorem 2 *Let $N > 2$, then the rate of convergence β of the Erlang loss model with N servers, arrival rate λ and service rate μ per server, satisfies*

$$\beta < 5\mu \quad \textit{if } \lambda \leq \mu\big(\sqrt{N}+1\big)^2. \tag{8}$$

Moreover, if $\lambda \geq \mu(\sqrt{N}+1)^2$, then

$$\beta > \mu + \big(\sqrt{\lambda} - \sqrt{\mu N}\big)^2 + \sqrt{\mu}\left(\sqrt{\gamma} + \frac{1}{2}\big(\sqrt{\gamma}\big(\sqrt{\lambda} - \sqrt{\mu N}\big)^2\big)^{1/3}\right), \tag{9}$$

where $\gamma := \lambda/(4N)$.

Our second representation for β is classic, and involves the stationary distribution $\boldsymbol{\pi} \equiv (\pi_0, \pi_1, \dots, \pi_N)$ of $\mathcal{X}$. It may be obtained by observing that the matrix DQD^{-1}, where

$$D := \operatorname{diag}\big(\sqrt{\pi_0}, \sqrt{\pi_1}, \dots, \sqrt{\pi_N}\big),$$

is symmetric. Since 0 is the largest and $-\beta$ the second largest eigenvalue of Q, and hence of DQD^{-1}, the Courant-Fischer theorem for symmetric matrices (see, for example, Meyer [15, p. 550]) tells us that

$$-\beta = \min_{\dim \mathcal{V}=n-1} \max_{\substack{\boldsymbol{y}\in\mathcal{V}\\ \boldsymbol{y}\neq\boldsymbol{0}}} \frac{\boldsymbol{y}DQD^{-1}\boldsymbol{y}^T}{\boldsymbol{y}\boldsymbol{y}^T}.$$

Since $\boldsymbol{\pi}Q=\boldsymbol{0}$, the vector $\boldsymbol{\pi}D^{-1}$ is a left eigenvector of DQD^{-1} corresponding to the eigenvalue 0. Hence, choosing $\mathcal{V}$ to be the space orthogonal to $\boldsymbol{\pi}D^{-1}$ we have

$$-\beta \leq \max_{\substack{\boldsymbol{y}D^{-1}\boldsymbol{\pi}^T=\boldsymbol{0}\\ \boldsymbol{y}\neq\boldsymbol{0}}} \frac{\boldsymbol{y}DQD^{-1}\boldsymbol{y}^T}{\boldsymbol{y}\boldsymbol{y}^T}.$$

But, in fact, equality holds, since we may choose $\boldsymbol{y}$ to be a left eigenvector of DQD^{-1} corresponding to the eigenvalue $-\beta$. Subsequently writing $\boldsymbol{x}=\boldsymbol{y}D$ and $\Pi=D^2$ we obtain the following representation, which was first established by Beneš [1] by reference to a result in the setting of symmetric operators. (As indicated by Beneš, the representation is implied by an observation of Kramer's [13] in the setting of reversible Markov chains.)

Theorem 3 *The rate of convergence β of the Erlang loss model with N servers, arrival rate λ and service rate μ per server, satisfies*

$$\beta = \min_{\substack{\boldsymbol{x}\boldsymbol{1}^T=\boldsymbol{0}\\ \boldsymbol{x}\neq\boldsymbol{0}}} \frac{\boldsymbol{x}(-Q)\Pi^{-1}\boldsymbol{x}^T}{\boldsymbol{x}\Pi^{-1}\boldsymbol{x}^T}, \tag{10}$$

where $\Pi \equiv \mathrm{diag}(\pi_0,\pi_1,\ldots,\pi_N)$ and Q is the matrix (3).

It follows in particular that for any vector $\boldsymbol{x}$ satisfying $\boldsymbol{x}\neq\boldsymbol{0}$ and $\boldsymbol{x}\boldsymbol{1}^T=0$ one has

$$\beta \leq \frac{\boldsymbol{x}(-Q)\Pi^{-1}\boldsymbol{x}^T}{\boldsymbol{x}\Pi^{-1}\boldsymbol{x}^T}. \tag{11}$$

Beneš [1] observed that choosing $x_i=(i-m)\pi_i/\sigma$, where

$$m=\frac{\lambda}{\mu}(1-\pi_N) \quad \text{and} \quad \sigma^2 = m-\frac{\lambda}{\mu}(N-m)\pi_N \tag{12}$$

are the mean and variance, respectively, of the number of busy servers in steady state, gives $\boldsymbol{x}(-Q)\Pi^{-1}\boldsymbol{x}^T=\mu m/\sigma^2$ and $\boldsymbol{x}\Pi^{-1}\boldsymbol{x}^T=1$, so that (11) leads to the bound

$$\beta \leq \frac{\mu m}{\sigma^2} = \frac{(1-\pi_N)\mu}{1-(N-m+1)\pi_N}. \tag{13}$$

Beneš observes that the bound can be used to approximate β if $\lambda<\mu N$.

At this point we mention a lower bound, derived by Jagerman [10, Theorem 13] by algebraic techniques, that may also be usable as an approximation to β, namely

$$\beta \geq \mu + \frac{\mu N}{\zeta_1 + \sqrt{(N-1)(N\zeta_2 - \zeta_1^2)}}, \tag{14}$$

where

$$\zeta_1 := N! \sum_{i=1}^{N} \frac{(\mu/\lambda)^i}{i(N-i)!}, \qquad \zeta_2 := \zeta_1^2 - 2N! \sum_{i=2}^{N} \frac{(\mu/\lambda)^i}{i(N-i)!} \sum_{j=1}^{i-1} \frac{1}{j}. \tag{15}$$

Further representations for β may be obtained by particularizing a result for ergodic birth-death processes that, in its full generality, was first stated by one of us in [23], and later by Chen [3]. We refer to [18] and [19] for more information on the various methods by which the result (or part of it) can be proven, and more references.

Theorem 4 *The rate of convergence β of the Erlang loss model with N servers, arrival rate λ and service rate μ per server, satisfies*

$$\beta = \max_{\boldsymbol{x}>\boldsymbol{0}} \left\{ \min_{1\leq i\leq N} \alpha_i(\boldsymbol{x}) \right\} = \min_{\boldsymbol{x}>\boldsymbol{0}} \left\{ \max_{1\leq i\leq N} \alpha_i(\boldsymbol{x}) \right\}, \tag{16}$$

where $\boldsymbol{x} \equiv (x_1, x_2, \ldots, x_N)$, and

$$\alpha_i(\boldsymbol{x}) := \left(1 - \frac{x_{i+1}}{x_i}\right)\lambda + \left(i - (i-1)\frac{x_{i-1}}{x_i}\right)\mu, \quad 1 \leq i \leq N, \tag{17}$$

with $x_0 = x_{N+1} = 0$.

It follows that for any vector $\boldsymbol{x} > \boldsymbol{0}$

$$\min_{1\leq i\leq N} \alpha_i(\boldsymbol{x}) \leq \beta \leq \max_{1\leq i\leq N} \alpha_i(\boldsymbol{x}). \tag{18}$$

For example, if $\lambda > \mu N$, we can choose $x_i = (\sqrt{\mu N/\lambda})^i$, $1 \leq i \leq N$, and find after a little algebra that

$$2\sqrt{\lambda\mu/N} - \mu \leq \beta - \left(\sqrt{\lambda} - \sqrt{\mu N}\right)^2 \leq (N+1)\sqrt{\lambda\mu/N}, \tag{19}$$

giving some supplementary information to Theorem 2.

One of the methods for proving Theorem 4 exploits the fact that $-\beta$ is in fact the *largest* eigenvalue of the matrix

$$C := \begin{pmatrix} -(\lambda+\mu) & \mu & 0 & \cdots & 0 & 0 \\ \lambda & -(\lambda+2\mu) & 2\mu & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & -(\lambda+(N-1)\mu) & (N-1)\mu \\ 0 & 0 & 0 & \cdots & \lambda & -(\lambda+N\mu) \end{pmatrix}, \tag{20}$$

which can be interpreted as the q-matrix of a transient birth-death process and therefore has only negative eigenvalues. The argument is given, for example, in [19, Sect. 4] in the more general setting of finite birth-death processes. Since C is a sign-symmetric tridiagonal matrix we can employ the results in [17] on representations and bounds for the largest eigenvalue of such matrices. It appears that [17, Theorem 1] leads to the min-max representation of Theorem 4, but [17, Theorem 5] (see also [9, Theorem 2]) leads to a new result.

Theorem 5 *The rate of convergence β of the Erlang loss model with $N > 1$ servers, arrival rate λ and service rate μ per server, satisfies*

$$\beta = \lambda + \frac{1}{2}\mu + \max_{x} \min_{1 \le i < N} \left\{ i\mu - \frac{1}{2}\sqrt{\mu^2 + \frac{4i\lambda\mu}{(1-x_i)x_{i+1}}} \right\}, \tag{21}$$

where $\boldsymbol{x} \equiv (x_1, x_2, \ldots, x_N)$ is such that $x_1 = 0$, $x_N = 1$, and $0 < x_i < 1$ for $1 < i < N$.

If $N = 2$ we can write down the exact value of $\beta \equiv \beta(N)$ directly from Theorem 5, namely

$$\beta(2) = \lambda + \frac{3}{2}\mu - \frac{1}{2}\sqrt{\mu^2 + 4\lambda\mu}. \tag{22}$$

For $N > 2$ and any vector $\boldsymbol{x}$ satisfying the requirements of Theorem 5 we obviously have

$$\beta \ge \lambda + \frac{1}{2}\mu + \min_{1 \le i < N} \left\{ i\mu - \frac{1}{2}\sqrt{\mu^2 + \frac{4i\lambda\mu}{(1-x_i)x_{i+1}}} \right\}. \tag{23}$$

For instance, by letting $x_i = \frac{1}{2}$, $1 < i < N$, we obtain the lower bound

$$\beta \ge \lambda + \frac{1}{2}\mu + \min_{1 \le i < N} \left\{ i\mu - \frac{1}{2}\sqrt{\mu^2 + 8e_i\lambda\mu i} \right\}, \tag{24}$$

where $e_i = 1$ if $i = 1, N-1$ and $e_i = 2$ otherwise.

Our final representation for β is similar to (10), but involves the matrix C of (20) rather than the matrix Q. It is obtained by symmetrizing the matrix C by a suitable similarity transform and applying the Courant-Fischer theorem to characterize the *largest* eigenvalue of the resulting matrix. This procedure amounts to applying a variant of [17, Theorem 8] to C, and leads to the following result.

Theorem 6 *The rate of convergence β of the Erlang loss model with N servers, arrival rate λ and service rate μ per server, satisfies*

$$\beta = \min_{x \ne \mathbf{0}} \frac{\boldsymbol{x}(-C)\tilde{\Pi}\boldsymbol{x}^T}{\boldsymbol{x}\tilde{\Pi}\boldsymbol{x}^T}, \tag{25}$$

where $\tilde{\Pi} \equiv \mathrm{diag}(\pi_0, \pi_1, \ldots, \pi_{N-1})$ and C is the matrix (20).

It follows that for any vector $\boldsymbol{x} \neq \mathbf{0}$ one has

$$\beta \leq \frac{\boldsymbol{x}(-C)\tilde{\Pi}\boldsymbol{x}^T}{\boldsymbol{x}\tilde{\Pi}\boldsymbol{x}^T}. \tag{26}$$

In particular, choosing $\boldsymbol{x} = \boldsymbol{e}_1$, the first unit vector, in (26) we find that

$$\beta \leq \lambda + \mu. \tag{27}$$

(Note that equality holds if $N = 1$.) A subtler approach is to minimize the upper bound (26) over all vectors $\boldsymbol{x}$ with two, adjacent, nonzero components. A little algebra then reveals for $N > 1$ the upper bound

$$\beta \leq \lambda + \frac{1}{2}\mu + \min_{1 \leq i < N} \left\{ i\mu - \frac{1}{2}\sqrt{\mu^2 + 4i\lambda\mu} \right\}. \tag{28}$$

This concludes our survey of representations and bounds for β. In the next section we will say more about the asymptotic behaviour of $\beta \equiv \beta(N)$ as $N \to \infty$.

3 Asymptotic results

The next theorem gives us the asymptotic behaviour of $\beta \equiv \beta(N)$ as $N \to \infty$ if $\lambda \equiv \lambda(N)$ is in some sense small. It encompasses in particular the case λ is constant.

Theorem 7 *If there is a constant $c < \mu$ such that $\lambda \leq cN$ for N sufficiently large, then the rate of convergence $\beta(N)$ of the Erlang loss model with N servers, arrival rate λ and service rate μ per server, satisfies*

$$\lim_{N \to \infty} \beta(N) = \mu. \tag{29}$$

Proof From [19, Theorem 12] we know that (29) holds true if $\lambda = cN$ for some $c < \mu$. Since $\xi_{N,1}(a) > 0$ is strictly increasing in a, the statement is implied by Theorem 1. □

In view of (7) we cannot improve upon the bound on c in this theorem. The asymptotic analysis of the linear case $\lambda = cN$ is completed by assuming $c > \mu$. The lower bound (19) (or, for N sufficiently large, the lower bound (9)) then tells us that

$$\beta(N) > \left(\sqrt{c} - \sqrt{\mu}\right)^2 N. \tag{30}$$

The following result, which was stated in [6] (without proof) and proven in [19], establishes that, actually, both sides of (30) are asymptotically equal.

Theorem 8 *If $c > \mu$, then the rate of convergence $\beta(N)$ of the Erlang loss model with N servers, arrival rate $\lambda = cN$ and service rate μ per server, satisfies*

$$\lim_{N \to \infty} \frac{\beta(N)}{N} = \left(\sqrt{c} - \sqrt{\mu}\right)^2.$$

We finally look into the case

$$\lambda = \mu N + a\sqrt{N} + o\bigl(\sqrt{N}\bigr) \quad \text{as } N \to \infty, \tag{31}$$

for some constant $a \in \mathbb{R}$. The scaling (31) is known as the *Halfin-Whitt regime*, after Halfin and Whitt [8] who introduced it in the setting of a multiserver queueing system (with negative a). In the setting at hand we have, for N sufficiently large,

$$\beta(N) < \begin{cases} 2\mu & \text{if } a < 0 \\ 5\mu & \text{if } a < 2\mu \end{cases}, \tag{32}$$

in view of (7) and Theorem 2, respectively. More refined statements may be obtained by applying the full [14, Theorem 5], but the main conclusion is that $\beta(N)$ is bounded whenever $a < 2\mu$. When $a > 2\mu$, Theorem 2 tells us that

$$\beta(N) > \frac{3}{2}\mu + \frac{1}{4}\left(\frac{a^2}{\mu^2} + \sqrt[3]{a^2}\right), \tag{33}$$

but it is not known for which values of $a \geq 2\mu$, if any, $\beta(N)$ is bounded.

4 Upper bounds on $d(t)$

Applying [19, Theorem 9] (which is implied by [22, Theorem 1] or [23, Theorem 3.2]) to the Erlang loss model, and recalling (2), gives us the following upper bound on the total variation distance between the time-dependent and stationary distributions.

Theorem 9 *For any initial distribution* $\boldsymbol{p}(0)$ *and vector* $\boldsymbol{x} \equiv (x_1, x_2, \ldots, x_N)$ *such that* $x_{\min} := \min_i\{x_i\} > 0$, *the total variation distance* $d(t)$ *between the distribution at time* t *and the stationary distribution in the Erlang loss model with* N *servers, arrival rate* λ *and service rate* μ *per server, satisfies*

$$d(t) \leq C(\boldsymbol{x})\, d(0)\, e^{-\min_i\{\alpha_i(\boldsymbol{x})\}t}, \quad t \geq 0, \tag{34}$$

where $C(\boldsymbol{x}) := 4\sum_{i=1}^{N}(x_i/x_{min})$, *and* $\alpha_i(\boldsymbol{x})$ *is given by* (17).

A simple corollary of this theorem (mentioned already in [22]) is obtained by choosing $x_i = 1$ for all i.

Corollary 10 *The total variation distance* $d(t)$ *between the distribution at time* t *and the stationary distribution in the Erlang loss model with* N *servers, arrival rate* λ *and service rate* μ *per server, satisfies*

$$d(t) \leq 4N\, d(0)\, e^{-\mu t}, \quad t \geq 0. \tag{35}$$

Of course, this result is particularly relevant in a setting such as that of Theorem 7, where μ is the limiting rate of convergence as $N \to \infty$. In the specific case $\lambda = cN$ with $c < \mu$, it was shown in [6, Proposition 10] by employing a coupling technique that, actually,

$$d(t) \le (N+1)\, d(0)\, e^{-\mu t}, \quad t \ge 0. \tag{36}$$

Continuing with the linear case $\lambda(N) = cN$, we now assume $c > \mu$. Theorem 9 leads to a bound—already mentioned in less explicit form in [19]—that slightly improves upon [6, Proposition 6].

Corollary 11 *If $c > \mu$, then the total variation distance between the distribution at time t and the stationary distribution in the Erlang loss model with N servers, arrival rate cN and service rate μ per server, satisfies*

$$d(t) \le C\, d(0)\, e^{-Mt}, \quad t \ge 0, \tag{37}$$

where

$$C := 4\frac{(\sqrt{c/\mu})^N - 1}{\sqrt{c/\mu} - 1} \quad \text{and} \quad M := \big(\sqrt{c} - \sqrt{\mu}\big)^2 N + 2\sqrt{c\mu} - \mu.$$

Proof Choosing $x_i = (\sqrt{\mu/c})^i$, $1 \le i \le N$, the quantities $\alpha_i(\boldsymbol{x})$ of (17) satisfy

$$\alpha_i(\boldsymbol{x}) = \begin{cases} (c - \sqrt{c\mu})N - (\sqrt{c\mu} - \mu)i + \sqrt{c\mu}, & 1 \le i < N, \\ cN - (\sqrt{c\mu} - \mu)N + \sqrt{c\mu}, & i = N, \end{cases}$$

so that $\min_i\{\alpha_i(\boldsymbol{x})\} = \alpha_{N-1}(\boldsymbol{x})$. The result follows readily from Theorem 9 by substitution. □

Note that, in view of Theorem 8, the exponent in (37) is asymptotically sharp as $N \to \infty$.

5 Time-dependent rates

In this section we allow the arrival rate $\lambda(t) \equiv \lambda(N,t)$ as well as the service rate per server $\mu(t)$ to be functions of time, and assume them to be nonnegative and locally integrable on $[0,\infty)$. Employing the approach of [23] and [7], we then obtain the following generalization of Theorem 9.

Theorem 12 *For any two initial distributions $\boldsymbol{p}^{(1)}(0)$ and $\boldsymbol{p}^{(2)}(0)$, and any vector $\boldsymbol{x} \equiv (x_1, x_2, \dots, x_N)$ such that $x_{min} := \min_i\{x_i\} > 0$, the total variation distance between the distributions $\boldsymbol{p}^{(1)}(t)$ and $\boldsymbol{p}^{(2)}(t)$ in the Erlang loss model with N servers, and arrival rate $\lambda(\tau)$ and service rate $\mu(\tau)$ per server at time τ, satisfies*

$$d_{tv}\big(\boldsymbol{p}^{(1)}(t), \boldsymbol{p}^{(2)}(t)\big) \le C(\boldsymbol{x})\, d_{tv}\big(\boldsymbol{p}^{(1)}(0), \boldsymbol{p}^{(2)}(0)\big) e^{-\int_0^t \min_i\{\alpha_i(\boldsymbol{x},\tau)\}\, d\tau}, \quad t \ge 0, \tag{38}$$

where $C(\boldsymbol{x}) := 4\sum_{i=1}^{N}(x_i/x_{min})$, *and*

$$\alpha_i(\boldsymbol{x},\tau) := \left(1 - \frac{x_{i+1}}{x_i}\right)\lambda(\tau) + \left(i - (i-1)\frac{x_{i-1}}{x_i}\right)\mu(\tau), \quad 1 \le i \le N, \tag{39}$$

with $x_0 = x_{N+1} = 0$.

Choosing $x_i = 1$ for all i gives us the generalization of Corollary 10 that was stated earlier in [23, Theorem 7.1].

Corollary 13 *For any two initial distributions* $\boldsymbol{p}^{(1)}(0)$ *and* $\boldsymbol{p}^{(2)}(0)$, *the total variation distance between the distributions* $\boldsymbol{p}^{(1)}(t)$ *and* $\boldsymbol{p}^{(2)}(t)$ *in the Erlang loss model with* N *servers*, *and arrival rate* $\lambda(\tau)$ *and service rate* $\mu(\tau)$ *per server at time* τ, *satisfies*

$$d_{tv}\big(\boldsymbol{p}^{(1)}(t), \boldsymbol{p}^{(2)}(t)\big) \le 4N\, d_{tv}\big(\boldsymbol{p}^{(1)}(0), \boldsymbol{p}^{(2)}(0)\big) e^{-\int_0^t \mu(\tau)\,d\tau}, \quad t \ge 0. \tag{40}$$

It follows in particular that the total variation distance between $\boldsymbol{p}^{(1)}(t)$ and $\boldsymbol{p}^{(2)}(t)$ tends to 0 as $t \to \infty$ if $\int_0^\infty \mu(\tau)\,d\tau = \infty$.

Let us finally consider the special case $\lambda(t) = Nc(t),\ t \ge 0$. Choosing $x_i = \delta^i$, with $0 < \delta < 1$ and δ so close to 1 that

$$\delta N c(t) > \Delta\mu(t), \quad t \ge 0, \tag{41}$$

where $\Delta := \delta^{-1} - 1$, it follows readily that

$$\min_i\{\alpha_i(\boldsymbol{x},t)\} = \alpha_{N-1}(\boldsymbol{x},t) = N\Delta\big(\delta c(t) - \mu(t)\big) + (2\Delta + 1)\mu(t), \quad t \ge 0. \tag{42}$$

Hence Theorem 12 leads to the following result.

Corollary 14 *Suppose that* $\delta,\ 0 < \delta < 1,\ c(\cdot)$, *and* $\mu(\cdot)$ *are such that* (41) *holds true. Then for any two initial distributions* $\boldsymbol{p}^{(1)}(0)$ *and* $\boldsymbol{p}^{(2)}(0)$, *the total variation distance between the distributions* $\boldsymbol{p}^{(1)}(t)$ *and* $\boldsymbol{p}^{(2)}(t)$ *in the Erlang loss model with* N *servers*, *and arrival rate* $\lambda(\tau) = Nc(\tau)$ *and service rate* $\mu(\tau)$ *per server at time* τ, *satisfies*

$$d_{tv}\big(\boldsymbol{p}^{(1)}(t), \boldsymbol{p}^{(2)}(t)\big) \le C\, d_{tv}\big(\boldsymbol{p}^{(1)}(0), \boldsymbol{p}^{(2)}(0)\big)\, e^{-\int_0^t M(\tau)\,d\tau}, \quad t \ge 0, \tag{43}$$

where

$$C := 4\Delta^{-1}\big(\delta^{-N} - 1\big) \quad \textit{and} \quad M(\tau) := N\Delta\big(\delta c(\tau) - \mu(\tau)\big) + (2\Delta + 1)\mu(\tau).$$

This corollary is a generalization of Corollary 11, for in the stationary setting $c(t) = c,\ \mu(t) = \mu$ and $c > \mu$, we regain Corollary 11 by choosing $\delta = \sqrt{\mu/c}$. Evidently, Corollary 14 is particularly relevant when the functions $c(t)$ and $\mu(t)$ are such that $\int_0^\infty (\delta c(\tau) - \mu(\tau))\,d\tau = \infty$ for δ sufficiently close to 1.

References

1. Beneš, V.E.: The covariance function of a simple trunk group with applications to traffic measurement. Bell Syst. Tech. J. **40**, 117–148 (1961)
2. Blanc, J.P.C., van Doorn, E.A.: Relaxation times for queueing systems. In: de Bakker, J.W., Hazewinkel, M., Lenstra, J.K. (eds.) Mathematics and Computer Science, Proceedings of the CWI symposium, 1983. CWI Monograph, vol. 1, pp. 139–162. North-Holland, Amsterdam (1986)
3. Chen, M.F.: Variational formulas and approximation theorems for the first eigenvalue in dimension one. Sci. China Ser. A **44**, 409–418 (2001)
4. Chihara, T.S.: An Introduction to Orthogonal Polynomials. Gordon & Breach, New York (1978)
5. Erlang, A.K.: Løsning af nogle problemer fra sandsynlighedsregningen af betydning for de automatiske telefoncentraler. Elektroteknikeren **13**, 5–13 (1917). (Translation: Solution of some problems in the theory of probabilities of significance in automatic telephone exchanges. In: E. Brockmeyer, H.L. Halstrøm, A. Jensen (eds.). The Life and Works of A.K. Erlang. Transactions of the Danish Academy of Technical Sciences, No. 2, pp. 138–155. Copenhagen (1948))
6. Fricker, C., Robert, P., Tibi, D.: On the rates of convergence of Erlang's model. J. Appl. Probab. **36**, 1167–1184 (1999)
7. Granovsky, B.L., Zeifman, A.I.: Nonstationary queues: estimation of the rate of convergence. Queueing Syst. **46**, 363–388 (2004)
8. Halfin, S., Whitt, W.: Heavy-traffic limits for queues with many exponential servers. Oper. Res. **29**, 567–588 (1981)
9. Ismail, M.E.H., Li, X.: Bound on the extreme zeros of orthogonal polynomials. Proc. Am. Math. Soc. **115**, 131–140 (1992)
10. Jagerman, D.L.: Nonstationary blocking in telephone traffic. Bell Syst. Tech. J. **54**, 625–661 (1975)
11. Karlin, S., McGregor, J.L.: Ehrenfest urn models. J. Appl. Probab. **2**, 352–376 (1965)
12. Kijima, M.: On the largest negative eigenvalue of the infinitesimal generator associated with $M/M/n/n$ queues. Oper. Res. Lett. **9**, 59–64 (1990)
13. Kramer, H.P.: Symmetrizable Markov matrices. Ann. Math. Stat. **30**, 149–153 (1959)
14. Krasikov, I.: Bounds for zeros of the Charlier polynomials. Methods Appl. Anal. **9**, 599–610 (2002)
15. Meyer, C.D.: Matrix Analysis and Applied Linear Algebra. SIAM, Philadelphia (2001). (Updates available on http://www.matrixanalysis.com)
16. Riordan, J.: Stochastic Server Systems. Wiley, New York (1962)
17. van Doorn, E.A.: Representations and bounds for zeros of orthogonal polynomials and eigenvalues of sign-symmetric tri-diagonal matrices. J. Approx. Theory **51**, 254–266 (1987)
18. van Doorn, E.A., van Foreest, N.D., Zeifman, A.I.: Representations for the extreme zeros of orthogonal polynomials. J. Comput. Appl. Math. (2009, to appear)
19. van Doorn, E.A., Zeifman, A.I., Panfilova, T.L.: Bounds and asymptotics for the rate of convergence of birth-death processes. Theory Probab. Appl. **54**, 18–38 (2009) (Russian edition)
20. Voit, M.: A note on the rate of convergence to equilibrium for Erlang's model in the subcritical case. J. Appl. Probab. **37**, 918–923 (2000)
21. Xie, S., Knessl, C.: On the transient behavior of the Erlang loss model: heavy usage asymptotics. SIAM J. Appl. Math. **53**, 555–599 (1993)
22. Zeifman, A.I.: Some estimates of the rate of convergence for birth and death processes. J. Appl. Probab. **28**, 268–277 (1991)
23. Zeifman, A.I.: Upper and lower bounds on the rate of convergence for non-homogeneous birth and death processes. Stoch. Process. Appl. **59**, 157–173 (1995)