Error Bounds for Perturbing Nonexponential Queues

Nico M. van Dijk
University of Amsterdam
Amsterdam, The Netherlands

Masakiyo Miyazawa
Tokyo University of Science
Noda, Chiba 278, Japan

Abstract

A general framework is provided to derive analytic error bounds for the effect of perturbations and inaccuracies of nonexponential service or arrival distributions in single- and multi-server queues. The general framework is worked out in detail for the three types of finite or infinite buffer queues: $GI/G/1/N$, $M/G/c/N$, and $GI/M/c/N$.

First, for the standard $GI/G/1/N$ queue, it is illustrated how the general error bound result can lead to error bounds for different performance measures like the throughput, mean queue length and stationary queue length distribution. Next, for the $M/G/c/N$ queue, an error bound and monotonicity result are established for the throughput. $M/G/c/N$ queues can so be compared even when hazard rates are not ordered. Finally, for the $GI/M/c/N$ queue, a similar result is obtained with a perturbation of the inter-arrival time distribution.

The error bound results are supported by asymptotic expressions for the $M/M/c/N$ queue and numerical results for the $GI/G/1/N$ queue.

Keywords: Queue, Finite waiting room, Nonexponential system, Stationary distribution, Perturbation, Error bound, Markov reward process, Many server.

MSC 2000 classification: 60K25, 60K20

OR/MS classification: **Queues** Approximation, Multichannel

1. Introduction

As nonexponential stochastic service systems such as $M/G/c/N$ and $GI/G/c/N$-queues cannot be solved analytically, there has been a huge number of studies on approximations for characteristics of interests. Reviews of such approximations can be found, among others, in Tijms [18] for single server queues (see also [8] for the $GI/G/1/N$ queue) and Kimura [6] for multi-server queues. These approximations are often based on closed form expressions that are analytically tractable (see, e.g., [9]). However, approximations can still be limited, complicated and computationally expensive. It thus remains of interest to compare nonexponential stochastic service systems under (slightly) different parameter and service assumptions, most notably,:

(i) To compare it with service cases that are more tractable, either analytically or numerically,

(ii) To compare its performance, most notably to quantify its difference, with that for the simple exponential case,

(iii) To take into account the natural feature that the underlying parameter estimates and service form assumption are usually based on data and thus involve some amount of imprecision or 'mis-specification'.

Stochastic comparison approach For $M/G/c/N$-queues, there has been enormous literature on closed form expressions and accurate approximations. Amazingly enough, relatively little attention has been paid to accuracy or sensitivity questions for (i), (ii) and (iii). One positive exception in this direction has been set by the extensive interest over the last three decades for stochastic comparison results (see, e.g., [5, 7]). Early stochastic comparison results for $GI/G/c$-queues can already be found in Daley and Moran [2] (single server), Stidham [15] (single server), Jacobs and Schach [4] (multi-servers), and Yu [28] (multi-Erlangian servers).

An elegant survey of bounds and approximations for $GI/G/c$-queues along this line is given in Stoyan [16], followed by a more extensive treatment of stochastic comparison results in his excellent book [17] (see also [11, 27] for more recent accounts). In these references infinite buffer queues were considered. Extensions for the finite buffer case can be found in Sonderman [13, 14] based on his thesis [12]. In [13] the effect of strongly stochastically ordered interarrival and service time distributions is studied with counterexamples referred to for specific cases. In [14] the effect of changing the number of servers is investigated.

In those stochastic comparisons, sample path and weak coupling arguments have been used extensively. Despite the elegance of such a sample path approach, a price to be paid here is that (strong) stochastic ordering conditions may have to be imposed. These may be hard to verify. Particularly, service distributions that are not stochastically ordered can not be considered. A second even more important drawback of a stochastic comparison approach is that no quantification or error bound is provided for the difference of characteristics to be compared such as the throughput or response time.

Error bound approach In Van Dijk and Puterman [23], a Markovian reward approach was introduced and applied to provide error bounds for the effect of perturbations or imprecisions such as in $M/M/\cdot/\cdot$ queues with different arrival and service rates. This reward approach has meanwhile been applied in a variety of non-solvable queueing network cases. For example, analytic error bounds are provided for truncating or expanding finite capacity networks (cf. [20, 22]).

So far, however, this approach has only been applied to exponential queues for establishing bounds and error bounds when modifying the system parameters or protocols (buffer sizes, number of servers, arrival rates, blocking protocol). Here, by exponential queue, it is meant that the inter-arrivals and service times are exponentially distributed. In Miyazawa and Van Dijk [10] a first step has been set to extend the Markov reward approach to nonexponential service systems. Error bounds were obtained for changing the buffer sizes in $GI/M/c/N$ queues. However, the arrival and service distributions were kept unchanged.

Extension In the present paper, the Markov reward approach is extended to also investigate the effect of changes in nonexponential arrival and service distributions in multi-server queues. More precisely, *analytic error bounds* are derived for the effect of *different nonexponential* distributions. This extension involves two essentially new technicalities:

1. A Markovian reward approach which also deals with states that change continuously in time in stead of only at discrete times.

2. The bounding of so-called bias-terms, as essential step of the Markov reward approach, with a continuous-state variable.

Neither of these two aspects has been dealt with before nor seems evident. As technical restriction, the service distributions are assumed to have bounded hazard rates. Nevertheless, these distributions are wide enough in the sense that they cover any phase type distribution with a finite state space. These in turn are dense within the set of all distributions.

The error bound results that will be obtained are not meant or can not be expected to be accurate. They are developed to provide secure bounds, orders of magnitude and comparison results also for queues that are not stochastically ordered.

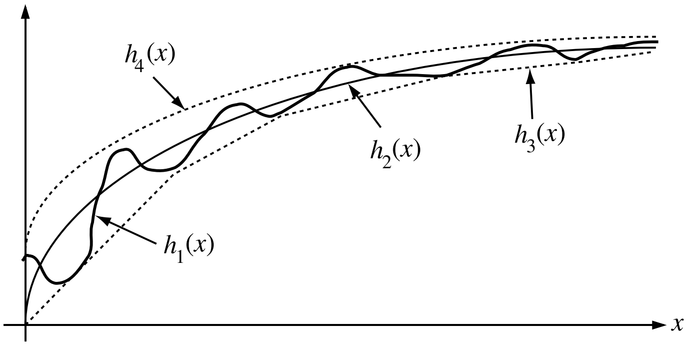

Figure 1: Comparison of non-ordered hazard rate functions

Example As an example, compare two $M/G/c/N$ queues with different service distributions G_1 and G_2. These distributions or rather their hazard rates h_1 and h_2 may not be ordered (see Figure 1). Nevertheless, a comparison will be made up to an inaccuracy (error bound) which can be expressed analytically by the difference in bounding hazard rates h_3 and h_4.

Results In order to develop, apply and illustrate these extensions as well as to show the quality of the error bounds, this paper contains three types of results:

1. A general error bound for the effect of changes in nonexponential distributions for multi-server queues.
2. An application of this result for a number of nonexponential service systems (single server case, nonexponential arrival and nonexponential service case).
3. Support by analytic asymptotic expressions and numerical results for a number of situations.

The technical details of these applications are highly complex and even in compact form constitute a substantial part of the paper. A comparison with some analytic cases illustrates that the (analytic) error bounds have the order that can at best be expected.

Outline and detailed results First we develop a general framework and derive a general error bound under certain technical conditions in Section 2. Next we consider its applications to the $GI/G/1/N$, $M/G/c/N$ and $GI/M/c/N$ queues in Sections 3, 4 and 5 respectively. In these sections, the technical conditions are verified. Particularly, error bounds for the so-called bias-terms are established. Analytic error bounds are obtained for perturbations in the service or arrival distribution. Different measures of interest are considered: the throughput, the steady state queue length distribution, and the mean queue length for the $GI/G/1/N$ case (Theorems 3.1, 3.2 and 3.3), the throughput for the $M/G/c/N$ case (Theorem 4.1) and the mean queue length for the $GI/M/c/N$ case (Theorem 5.1). As side results, Theorems 3.3 and 3.4 also include monotonicity results that could also have been concluded by stochastic comparison results, provided the queues are ordered. Moreover, in Corollary 4.2, a monotonicity result is established for queues that are not ordered. Some numerical support is provided in Section 6. An evaluation of the error bound results and some remarks on their limitations and possible extensions conclude the paper.

2. Nonexponential perturbation and error bound results

This section contains the general framework and the basic error bound result. In the subsequent sections, error bound results will be concluded for a number of specific situations and performance measures by verifying the technical conditions.

2.1 Formulation and notation

Consider a multi-server queue with a nonexponential arrival stream, nonexponential services and a finite or infinite capacity. More precisely, a $GI/G/c/N$ queue is considered, in which the interarrival and service times are independent and identically distributed with an arrival distribution function $F(t)$ and a service distribution $G(t)$, respectively, and there are c servers and a waiting room of size $M \equiv N - c$. Here, N may be infinite. When the waiting room is fully occupied, arriving jobs are rejected and lost. Servers are numbered as $1, 2, \ldots, c$, and jobs are processed in First Come First Served order. As we need to keep track of the individual services and their evolution, for convenience we make the following assumptions. When ℓ jobs are in service with $\ell \leq c$, servers $1, 2, \ldots, \ell$ are occupied. When a job arrives, it is assigned to server $\ell + 1$ if $\ell < c$, and otherwise added to the end of the waiting line. When the job at server $i \leq c$ completes service, the first job in the waiting line instantaneously moves to server i. If there are no waiting jobs, the jobs at servers $i+1, \ldots, \ell$ shift to servers $i, \ldots, \ell - 1$. This service policy is referred to as a packing rule. Note that this packing rule is equivalent to renumbering servers upon arrivals and service completions, since service speeds are identical for all the servers. However, the packing rule is more convenient for our purposes.

Furthermore, we impose the assumption that both the arrival and service distributions $F(t)$ and $G(t)$ are absolutely continuous with density functions $f(t)$ and $g(t)$, and have bounded hazard rate functions $\lambda(t)$ and $h(t)$. That is, for some finite constants H_f and H_g,

$$\begin{aligned} \lambda(t) &= \frac{f(t)}{1 - F(t)} \leq H_f, \qquad && t \geq 0, \\ h(t) &= \frac{g(t)}{1 - G(t)} \leq H_g, \qquad && t \geq 0. \end{aligned}$$

Since this class of distributions is dense in the class of one dimensional distributions on the nonnegative half line, the assumption is not very restrictive. For instance, it includes any phase type distribution with a finite state space.

In order to present a Markovian description in continuous time, we need to keep track of the state $\boldsymbol{x} = (n, a, \boldsymbol{s})$ with

n : the total number of jobs present which includes the jobs in service,

a : the elapsed time since the last arrival,

$\boldsymbol{s} = (s_1, s_2, \ldots, s_\ell)$: the ℓ-dimensional vector with $\ell \leq c$ with by s_i denoting the attained service time of the job that is presently at server i. If $\ell = 0$, equivalently, $n = 0$, $\boldsymbol{s}$ is a null vector with no component. We denoted it by the number 0 for convenience.

Performance measures Let $\{X(t) | t \geq 0\}$ be the corresponding Markov process. Let $\nu(t)$ be the number of the state changes up to time t due to arrivals or service completions. We will distinguish two types of possible reward functions.

$r_1(\boldsymbol{x})$: a reward rate when the system state is $\boldsymbol{x}$,

$r_2(\boldsymbol{x}, \boldsymbol{y})$: an instantaneous reward when the system state changes from $\boldsymbol{x}$ to $\boldsymbol{y}$.

The average performance measures A of interest can then be expressed by

$$A = \lim_{t\to\infty} \frac{1}{t} E\Big[\int_0^t r_1(X(u))du + \int_0^t r_2(X(u-), X(u+))d\nu(u) \Big| X(0) = \boldsymbol{x}_0 \Big]. \tag{2.1}$$

Here, E denotes an expectation, and the limit is assumed to exist and to be independent of the initial state $\boldsymbol{x}_0$. As usual, $X(u-)$ and $X(u+)$ are the value of $X(u)$ just before and just after time t, respectively.

The distinction of the reward functions r_1 and r_2 is convenient for representing different performance measures. For example, A represents the *mean queue length* by setting

$$r_1(n, a, \boldsymbol{s}) = n$$

and $r_2 \equiv 0$. A represents the *throughput* of the system by setting $r_1 \equiv 0$ and

$$\begin{aligned} r_2((n,a,\boldsymbol{s}),(n',a',\boldsymbol{s}')) = &\sum_{\ell=1}^{c} 1(n' = n-1 \geq c, a' = a, s'_\ell = 0, s'_j = s_j > 0 \text{ for } j \neq \ell) \\ &+1(c > n' = n-1 \geq 0, a' = a, \boldsymbol{s}'_n = 0, s'_j = s_j > 0 \text{ for } j < n), \end{aligned}$$

where $1(\mathcal{S})$ is the indicator function of the statement $\mathcal{S}$. In most applications, either $r_1 \equiv 0$ or $r_2 \equiv 0$. But, this is not necessary for the general case.

We are interested in the effects on these performance measures when the system, more precisely the hazard rate function λ or h, is perturbed. In order to study these effects, we will present a general error bound (Lemma 2.2). First, let us introduce some notation and reformulate the continuous time Markov process as a discrete time Markov process.

Notation We use the following operations for vectors. For $\boldsymbol{s} = (s_1, \cdots, s_\ell)$ and for $x \geq 0$, define operations $\oplus, \ominus$ and $\vee$ by

$$\begin{cases} \boldsymbol{s} \oplus \boldsymbol{u}_j(x) = (s_1, \cdots, s_{j-1}, x, s_j, \cdots, s_\ell) & \text{for } j \leq \ell + 1 \\ \boldsymbol{s} \ominus \boldsymbol{u}_i = (s_1, \cdots, s_{i-1}, s_{i+1}, \cdots, s_\ell) & \text{for } i \leq \ell \\ \boldsymbol{s} \ominus (\boldsymbol{u}_i \vee \boldsymbol{u}_j) = (s_1, \cdots, s_{i-1}, s_{i+1}, \cdots, s_{j-1}, s_{j+1}, \cdots, s_\ell) & \text{for } i < j \leq \ell\,, \end{cases}$$

particularly, if $\ell < c$, then

$$\boldsymbol{s} \oplus \boldsymbol{u}_{\ell+1}(x) = (s_1, \cdots, \cdots, s_\ell, x)\,.$$

The operation $\oplus$ presents to add a job in service, and $\ominus$ presents to remove a job from service. Those operations are always performed from the left to the right, so parentheses for multiple operations are omitted unless it causes any confusion. For instance, with $i < j < \ell$:

$$\begin{aligned} &\boldsymbol{s} \ominus \boldsymbol{u}_i \oplus \boldsymbol{u}_i(x) = (s_1, \cdots, s_{i-1}, x, s_{i+1}, \cdots, s_\ell) \\ &\boldsymbol{s} \oplus \boldsymbol{u}_i(x) \oplus \boldsymbol{u}_j(y) = (s_1, \cdots, s_{i-1}, x, s_{i+1}, \cdots, s_{j-2}, y, s_{j-1}, \cdots, s_\ell)\,. \end{aligned}$$

Note that the order of these operations can not be changed. For instance, for $\boldsymbol{s} = (1, 2, 3)$ with $n = 3$, we have

$$\boldsymbol{s} \oplus \boldsymbol{u}_1(0) \oplus \boldsymbol{u}_3(0) = (0, 1, 0, 2, 3) \neq (0, 1, 2, 0, 3) = \boldsymbol{s} \oplus \boldsymbol{u}_3(0) \oplus \boldsymbol{u}_1(0)\,.$$

Furthermore, for any $v > 0$ and vector $\boldsymbol{s} = (s_1, \ldots, s_\ell)$, let

$$\boldsymbol{s}_v = (s_1 + v, s_2 + v, \ldots, s_\ell + v).$$

Uniformization In our arguments, we will evaluate the error bounds by induction based upon discrete-time cumulative reward functions. To this end, we first transform the continuous time Markov process into a discrete-time Markov chain in line with an extended uniformization result obtained in [19].

Consider an alternative Markov process, denoted by $\{Y(t)|t \geq 0\}$, which has jumps according to a Poisson process with rate Q, where Q is any number such that

$$H_f + cH_g \leq Q.$$

This Markov process has state description $(n, a, \boldsymbol{s})$. The transition kernel is as follows. Given that the state just before the jump is $(n, a, \boldsymbol{s})$, at the jump the state will change into $(n', a', \boldsymbol{s}')$ with probability:

$$P((n,a,\boldsymbol{s}),(n',a',\boldsymbol{s}')) = \begin{cases} Q^{-1}\lambda(a), & \begin{cases} n' = n+1 \leq c, a' = 0, \boldsymbol{s}' = \boldsymbol{s} \oplus \boldsymbol{u}_{n+1}, \\ n' = n+1 \geq c+1, a' = 0, \boldsymbol{s}' = \boldsymbol{s}, \\ n' = n = N, a' = 0, \boldsymbol{s}' = \boldsymbol{s}, \end{cases} \\ Q^{-1}h(s_i) & \begin{cases} n' = n-1 < c, a' = a, \boldsymbol{s}' = \boldsymbol{s} \ominus \boldsymbol{u}_i, \\ n' = n-1 \geq c, a' = a, \boldsymbol{s}' = \boldsymbol{s} \ominus \boldsymbol{u}_i \oplus \boldsymbol{u}_i(0), \\ (i = 1, 2, \ldots, \min(n,c)), \end{cases} \\ 1 - Q^{-1}[\lambda(a) + \sum_{i=1}^{\min(n,c)} h(s_i)], & \qquad (n', a', \boldsymbol{s}') = (n, a, \boldsymbol{s}). \end{cases}$$

For instance, the third case, i.e., $n' = n = N, a' = 0, \boldsymbol{s}' = \boldsymbol{s}$, represents that a customer arrives but finds no space to enter. Between the jumps, the continuous components of $Y(t)$ are assumed to increase at rate 1. That is, if the system is in state $(n, a, \boldsymbol{s})$ just after the jump and if there is no jump during a time interval of length v, then the system will be in state $(n, a+v, \boldsymbol{s}_v)$ at time v after the jump. The following fact is then an immediate consequence of [19], of which the proof is based upon showing that both processes have the same infinitesimal generators.

Lemma 2.1 The processes $\{X(t)|t \geq 0\}$ and $\{Y(t)|t \geq 0\}$ are stochastically equivalent, i.e., they have the same joint probability distributions.

As a direct consequence of this equivalence result, we can evaluate the average values A in a discrete manner by just using the jump transition probability function P. To this end, define functions $V_k(n, a, \boldsymbol{s})$ for $k = 0, 1, \ldots$ as

$$V_k(n,a,\boldsymbol{s}) = E\Big(\int_0^{\tau_k} r_1(X(u))du + \sum_{i=1}^{k} r_2(X(\tau_i-), X(\tau_i+))\Big| X(0) = (n,a,\boldsymbol{s})\Big),$$

where τ_i is the i-th transition epoch under the uniformization. In words, V_k collects the rewards up to the k-th state change at rate r_1 and instantaneous rewards r_2. We therefore refer to V_k as a cumulative reward function. Let $R_v(n, a, \boldsymbol{s})$ be the expected reward up to time v since the last jump instant, given that the system was in state $(n, a, \boldsymbol{s})$ directly after this jump instant and that the next jump takes place after time v. That is, R_v is defined as

$$\begin{aligned} R_v(n, a, \boldsymbol{s}) = & \int_0^v r_1(n, a + u, \boldsymbol{s}_u) du \\ & + \sum_{(n', a', \boldsymbol{S}')} P((n, a + v, \boldsymbol{s}_v), (n', a', \boldsymbol{s}')) r_2((n, a + v, \boldsymbol{s}_v), (n', a', \boldsymbol{s}')). \end{aligned} \tag{2.2}$$

The cumulative rewards V_k can then be computed iteratively by $V_0(\cdot) \equiv 0$ and, for $k = 1, 2, \ldots$,

$$\begin{aligned} V_k(n, a, \boldsymbol{s}) = & \int_0^\infty dv Q e^{-Qv} \Big[R_v(n, a, \boldsymbol{s}) \\ & + \sum_{(n', a', \boldsymbol{S}')} P((n, a + v, \boldsymbol{s}_v), (n', a', \boldsymbol{s}')) V_{k-1}(n', a', \boldsymbol{s}') \Big]. \end{aligned} \tag{2.3}$$

Here, V_k has the stochastic interpretation of the expected cumulative reward over k-steps where each step requires an exponential time with parameter Q, and thus with expected step time Q^{-1}. This also explains the factor Q in (2.4) below. Based upon the equivalence result by Lemma 2.1, the average reward A can then be computed by

$$A = \lim_{k \to \infty} \frac{V_k(n, a, \boldsymbol{s})}{E(\tau_k)} = \lim_{k \to \infty} \frac{Q}{k} V_k(n, a, \boldsymbol{s}), \tag{2.4}$$

for an arbitrary initial state $(n, a, \boldsymbol{s})$. Note that the right-hand side of (2.4) does not depend on Q, because A does not depend on Q by Lemma 2.1.

2.2 General error bound result

Now let the arrival and service hazard rates, $\lambda(t)$ and $h(t)$, of the multi-server queue described above be perturbed into hazard rates $\overline{\lambda}(t)$ and $\overline{h}(t)$, respectively. Here, we assume that also $\overline{\lambda}$ and $\overline{h}$ are bounded, say by constants $\overline{H}_f$ and $\overline{H}_g$, respectively, and that Q is chosen in such a way that

$$\overline{H}_f + \overline{H}_g \leq Q.$$

For simplicity, the capacity N is kept unchanged. But it is easy to implement such a perturbation as well (see, e.g., [20]). We refer to this modified system as the perturbed system, and all notation and symbols are carried over for the perturbed system with an upper bar symbol, e.g., $\overline{P}$, $\overline{A}$ and $\overline{V}_k$.

We aim to establish error bounds for $|A - \overline{A}|$. To this end, we impose the natural assumption that both the original and perturbed systems have stationary distributions with multi-dimensional distribution functions:

$$\Pi(n, a, \boldsymbol{s}) \qquad \text{and} \qquad \overline{\Pi}(n, a, \boldsymbol{s}).$$

By π and $\overline{\pi}$, we denote the marginal mass functions with respect to the first components of Π and $\overline{\Pi}$, i.e.,

$$\begin{aligned} \pi(n) &= \Pi(n, +\infty, +\infty) - \Pi(n-1, +\infty, +\infty)1(n \geq 1), \\ \overline{\pi}(n) &= \overline{\Pi}(n, +\infty, +\infty) - \overline{\Pi}(n-1, +\infty, +\infty)1(n \geq 1). \end{aligned}$$

Furthermore, for a nonnegative function ψ in $(n, a, \boldsymbol{s})$, we define the scalar $\langle \Pi, \psi \rangle$ by

$$\langle \Pi, \psi \rangle = \int d\Pi(n, a, \boldsymbol{s})\psi(n, a, \boldsymbol{s}),$$

and, for all $v > 0$, we define the operators M_v for ψ by

$$M_v\psi(n, a, \boldsymbol{s}) = \sum_{n'} \int_{a', \boldsymbol{s}'} P((n, a+v, \boldsymbol{s}_v), (n', da', d\boldsymbol{s}'))\psi(n', a', \boldsymbol{s}').$$

Clearly, the operator M_v are linear and bounded under the sup norm. For $k = 0, 1, \dots$ and $v \geq 0$, define the difference function Z_v^k as

$$Z_v^k(n, a, \boldsymbol{s}) = [\overline{R}_v - R_v](n, a, \boldsymbol{s}) + [\overline{M}_v - M_v]V_k(n, a, \boldsymbol{s}). \tag{2.5}$$

Lemma 2.2 (i) If, for some nonnegative function δ,

$$|Z_v^k(n, a, \boldsymbol{s})| < \delta(n, a+v, \boldsymbol{s}_v)Q^{-1} \qquad \text{for all } k, v, \text{ and } (n, a, \boldsymbol{s}), \tag{2.6}$$

then

$$|A - \overline{A}| \leq \langle \overline{\Pi}, \delta \rangle. \tag{2.7}$$

(ii) If $Z_v^k(n, a, \boldsymbol{s}) \geq 0$ for all k, v and $(n, a, \boldsymbol{s})$, then

$$A \leq \overline{A}.$$

Remark 2.1 (a) To bound $Z_v^k(n, a, \boldsymbol{s})$, $\delta(n, a+v, \boldsymbol{s}_v)$ instead of $\delta(n, a, \boldsymbol{s}_v)$ is used in (2.6). As will be apparent from the proof below, this enables one to express the error bound (2.7) in terms of $\overline{\Pi}$ instead of the embedded distribution just after a jump instant, which will be denoted by $\overline{\Pi}^+$. The latter distribution is more difficult to get than $\overline{\Pi}$. The bounding functions in our applications will fit the form (2.6).
(b) It may look strange that the bound in (2.7) needs $\overline{\Pi}$ of the perturbed system. However, by using symmetry arguments in Π and $\overline{\Pi}$, we can replace $\overline{\Pi}$ by Π. Furthermore, full information on $\overline{\Pi}$ (or Π) may not be required.

PROOF. By comparing the one step relation (2.3) for the original and perturbed system, for any k, we obtain

$$\begin{aligned}
&(\overline{V}_{k+1} - V_{k+1})(n, a, \boldsymbol{s}) \\
&\quad = \int_0^\infty dvQe^{-Qv}\Big[(\overline{R}_v - R_v)(n, a, \boldsymbol{s}) + (\overline{M}_v\overline{V}_k - M_vV_k)(n, a, \boldsymbol{s})\Big] \\
&\quad = \int_0^\infty dvQe^{-Qv}\Big[(\overline{R}_v - R_v)(n, a, \boldsymbol{s}) + (\overline{M}_v - M_v)V_k(n, a, \boldsymbol{s})\Big] \\
&\qquad\qquad + \int_0^\infty dvQe^{-Qv}\overline{M}_v(\overline{V}_k - V_k)(n, a, \boldsymbol{s}). \qquad (2.8)
\end{aligned}$$

Let $\{\overline{Y}(t)|t \geq 0\}$ be the uniformized Markov process for the perturbed system, corresponding with $\{Y(t)|t \geq 0\}$ of the original system. Let $\overline{\Pi}^+$ be the stationary distribution of $\{\overline{Y}(\overline{\tau}_i+)|i = 0, 1, 2, \ldots\}$, where $\overline{\tau}_i$ is the i-th jump instant of $\overline{Y}(t)$. Since those jump instants constitute a Poisson process, the stationary distribution $\overline{\Pi}$ of $\overline{Y}(t)$ is identical with the one just before the jump instants due to the well known property PASTA (Poisson Arrivals See Time Average). Hence, we have

$$\overline{\Pi}^+(n, a, \boldsymbol{s}) = \int_{(n',a',\boldsymbol{s}')} d\overline{\Pi}(n', a', \boldsymbol{s}')P((n', a', \boldsymbol{s}'), (n, a, \boldsymbol{s})).$$

Then, for any function ψ,

$$\begin{aligned}
\int_{(n,a,\boldsymbol{s})} d\overline{\Pi}^+(n, a, \boldsymbol{s}) \int_0^\infty dvQe^{-Qv}\overline{M}_v\psi(n, a, \boldsymbol{s}) &= \int_{(n,a,\boldsymbol{s})} d\overline{\Pi}^+(n, a, \boldsymbol{s})\psi(n, a, \boldsymbol{s}), \\
\int_{(n,a,\boldsymbol{s})} d\overline{\Pi}^+(n, a, \boldsymbol{s}) \int_0^\infty dvQe^{-Qv}\psi(n, a + v, \boldsymbol{s}_v) &= \int_{(n,a,\boldsymbol{s})} d\overline{\Pi}(n, a, \boldsymbol{s})\psi(n, a, \boldsymbol{s}),
\end{aligned}$$

where the second equation is obtained again by PASTA. So taking absolute values of both sides of (2.8) and integrating with respect to $\overline{\Pi}^+$ yield

$$\langle\overline{\Pi}^+, |\overline{V}_{k+1} - V_{k+1}|\rangle \leq Q^{-1}\langle\overline{\Pi}, \delta\rangle + \langle\overline{\Pi}^+, |\overline{V}_k - V_k|\rangle.$$

Summing for $k = 0, 1, \ldots, m - 1$ yields

$$\langle\overline{\Pi}^+, |\overline{V}_m - V_m|\rangle \leq mQ^{-1}\langle\overline{\Pi}, \delta\rangle, \qquad m = 1, 2, \ldots.$$

Multiplying both sides by Q/m, letting $m \to \infty$, and applying Fatou's lemma gives

$$\langle\overline{\Pi}^+, \liminf_{m\to\infty} \frac{Q}{m}|\overline{V}_m - V_m|\rangle \leq \liminf_{m\to\infty}\langle\overline{\Pi}^+, \frac{Q}{m}|\overline{V}_m - V_m|\rangle \leq \langle\overline{\Pi}, \delta\rangle.$$

By the definitions of A and $\overline{A}$ and the assumption that A and $\overline{A}$ are independent of the initial values, this proves part (i) of Lemma 2.2. Part (ii) follows similarly by not taking absolute values. □

Lemma 2.2 may seem impractical as $V_k(n, a, \boldsymbol{s})$ will generally grow linearly in k while the bound $\delta(n, a + v, \boldsymbol{s}_v)$ in (2.6) must be independent of k. However, by using the fact that $\overline{P}_v$ and P_v are transition probability functions, we can write

$$\begin{aligned}
Z_v^k(n,a,\boldsymbol{s}) &= [\overline{R}_v - R_v](n,a,\boldsymbol{s}) \\
&\quad + Q^{-1}\Big[\overline{\lambda}(a+v) - \lambda(a+v)\Big]\Big\{1(n<c)\Big[V_k(n+1,0,\boldsymbol{s}_v \oplus \boldsymbol{u}_{n+1}(0)) - V_k(n,a+v,\boldsymbol{s}_v)\Big] \\
&\qquad + 1(c \le n < N)\Big[V_k(n+1,0,\boldsymbol{s}_v) - V_k(n,a+v,\boldsymbol{s}_v)\Big] \\
&\qquad + 1(c = N)\Big[V_k(n,0,\boldsymbol{s}_v) - V_k(n,a+v,\boldsymbol{s}_v)\Big]\Big\} \\
&\quad + Q^{-1}\sum_{i=1}^{\min(n,c)}\Big[\overline{h}(s_i+v) - h(s_i+v)\Big] \\
&\qquad \times\Big\{1(n\le c)\Big[V_k(n-1,a+v,\boldsymbol{s}_v \ominus \boldsymbol{u}_i) - V_k(n,a+v,\boldsymbol{s}_v)\Big] \\
&\qquad + 1(n>c)\Big[V_k(n-1,a+v,\boldsymbol{s}_v \ominus \boldsymbol{u}_i \oplus \boldsymbol{u}_i(0)) - V_k(n,a+v,\boldsymbol{s}_v)\Big]\Big\}. \qquad (2.9)
\end{aligned}$$

Expression (2.9) more explicitly shows the effect of a perturbation. More importantly, it enables one to transform conditions on V_k, which are generally unbounded in k, into conditions on difference terms (so called bias terms) of V_k. These difference terms, in turn, are generally uniformly bounded in k, as will be proved in the subsequent sections. In fact, in what follows, the main focus will be to obtain explicit analytic bounds for these difference terms.

Remark 2.2 As is directly seen from the proof of Lemma 2.2, the role of the original and modified systems can be interchanged. This observation can be useful when relative errors are considered.

3. $GI/G/1/N$ queue

In this section we will investigate a single server queue with a perturbation of the service distribution. The more general s-server case will be dealt with in the next section. However, the single server case is dealt with separately first to illustrate the technicalities of interest. Furthermore, both asymptotic and numerical results can best be evaluated for the single server case.

Throughout this and subsequent sections it is assumed that $h(t)$ nondecreasing in t. By a symmetry of arguments, this can be replaced by a monotonicity assumption of $\overline{h}(t)$. This monotonicity assumption is one of prices to be paid to obtain analytical error bounds.

3.1 Bias terms

Consider an original and perturbed $GI/G/1/N$ in which the hazard rate of the service times is perturbed from $h(t)$ to $\overline{h}(t)$. By writing out $Z_v(n,a,\boldsymbol{s})$ as per (2.9) with $c = 1$

and $\lambda(t) = \overline{\lambda}(t)$, where in the notation the vector $\boldsymbol{s}$ becomes a scalar s which represents the attained service time for the job in service, we find

$$\begin{aligned} Z_v(n,a,s) &= [\overline{R}_v(n,a,s) - R_v(n,a,s)] \\ &\quad + Q^{-1}[\overline{h}(s+v) - h(s+v)]1(n>0)[V_k(n-1,a,0) - V_k(n,a,s+v)]. \end{aligned} \tag{3.1}$$

In order to apply the error bound result, it would be sufficient to bound the bias terms $V_k(n-1,a,0)) - V_k(n,a,s+v)$. We evaluate these differences using recursive expressions in k as by (2.3). This in turn also requires to consider differences $V_k(n-1,a,0)) - V_k(n-1,a,z)$. We introduce the following notation.

$$\Delta_z V_k(n,a,s) = \begin{cases} V_k(n+1,a,s+z) - V_k(n,a,s), & n>0, \\ V_k(1,a,z) - V_k(0,a,0), & n=0, s=0 \\ 0 & \text{otherwise}, \end{cases} \tag{3.2}$$

$$\delta_z V_k(n,a,s) = \begin{cases} V_k(n,a,s+z) - V_k(n,a,s), & n>0, \\ 0 & \text{otherwise}, \end{cases} \tag{3.3}$$

To derive recursive expressions for the bias terms in (3.2), now consider a state (n,a,s) and value k. Let $\alpha = 1/Q$. Then by virtue of (2.3), we obtain

$$\begin{aligned} V_k(n,a,s) &= \int_0^\infty dvQe^{-Qv}\Big[R_v(n,a,s) \\ &\quad + \alpha\lambda(a+v)1(n<N)V_{k-1}(n+1,0,s+v) \\ &\quad + \alpha\lambda(a+v)1(n=N)V_{k-1}(n,0,s+v) \\ &\quad + \alpha h(s+v)1(n>0)V_{k-1}(n-1,a+v,0) \\ &\quad + [1-\alpha\lambda(a+v) - \alpha h(s+v)]1(n>0)V_{k-1}(n,a+v,s+v) \\ &\quad + [1-\alpha\lambda(a+v)]1(n=0)V_{k-1}(0,a+v,0)\Big]. \end{aligned} \tag{3.4}$$

Similarly, for state $(n+1,a,s+z)$, we find

$$\begin{aligned} V_k(n+1,a,s+z) &= \int_0^\infty dvQe^{-Qv}\Big[R_v(n+1,a,s+z) \\ &\quad + \alpha\lambda(a+v)1(n+1<N)V_{k-1}(n+2,0,s+v+z) \\ &\quad + \alpha\lambda(a+v)1(n+1=N)V_{k-1}(n+1,0,s+v+z) \\ &\quad + \alpha h(s+v+z)V_{k-1}(n,a+v,0) \\ &\quad + [1-\alpha\lambda(a+v) - \alpha h(s+v+z)]V_{k-1}(n+1,a+v,s+v+z)\Big]. \end{aligned} \tag{3.5}$$

Now in order to subtract (3.4) from (3.5), it would be convenient if transitions from (3.4) and (3.5) could be paired. To this end, first note that we only need to consider $n<N$ as $n+1 \leq N$. Therefore, in (3.4), we can rewrite the second term in the right hand side as

$$\begin{aligned} &\alpha\lambda(a+v)1(n<N)V_{k-1}(n+1,0,s+v) \\ &= \alpha\lambda(a+v)\Big[1(n+1<N)V_{k-1}(n+1,0,s+v) + 1(n+1=N)V_{k-1}(n+1,0,s+v)\Big], \end{aligned}$$

and the last two terms as

$$\begin{aligned}
&[1-\alpha\lambda(a+v)-\alpha h(s+v)]1(n>0)V_{k-1}(n,a+v,s+v)\\
&\quad+[1-\alpha\lambda(a+v)]1(n=0)V_{k-1}(0,a+v,0)\\
&\quad\quad=[1-\alpha\lambda(a+v)-\alpha h(s+v+z)]V_{k-1}(n,a+v,s+v)\\
&\quad\quad\quad+\alpha\Big[h(s+v+z)-h(s+v)\Big]1(n>0)V_{k-1}(n,a+v,s+v)\\
&\quad\quad\quad+\alpha h(s+v+z)1(n=0)V_{k-1}(0,a+v,0).
\end{aligned}$$

Then, after these substitutions have been made and by subtracting (3.4) from (3.5), we find the following expression after pairwise arranging terms with the same coefficients, while the one but last term is indeed equal to 0 but left in for clarity of the derivation.

$$\begin{aligned}
&V_k(n+1,a,s+z)-V_k(n,a,s)\\
&\quad=\int_0^\infty dvQe^{-Qv}\Big\{\Big[R_v(n+1,a,s+z)-R_v(n,a,s)\Big]\\
&\quad\quad+\alpha\lambda(a+v)1(n+1<N)\Big[V_{k-1}(n+2,0,s+v+z)-V_{k-1}(n+1,0,s+v)\Big]\\
&\quad\quad+\alpha\lambda(a+v)1(n+1=N)\Big[V_{k-1}(n+1,0,s+v+z)-V_k(n+1,0,s+v)\Big]\\
&\quad\quad+\alpha h(s+v)1(n>0)\Big[V_{k-1}(n,a+v,0)-V_{k-1}(n-1,a+v,0)\Big]\\
&\quad\quad+\alpha\Big[h(s+v+z)-h(s+v)\Big]1(n>0)\Big[V_{k-1}(n,a+v,0)-V_{k-1}(n,a+v,s+v)\Big]\\
&\quad\quad+\alpha h(s+v+z)1(n=0)\Big[V_{k-1}(0,a+v,0)-V_{k-1}(0,a+v,0)\Big]\\
&\quad\quad+[1-\alpha\lambda(a+v)-\alpha h(s+v+z)]\\
&\quad\quad\quad\times\Big[V_{k-1}(n+1,a+v,s+v+z)-V_{k-1}(n,a+v,s+v)1(n>0)\Big]\Big\}. \qquad (3.6)
\end{aligned}$$

By using the notation from (3.2) and (3.3), the bias (or difference) term $\Delta_z V_k$ can be expressed, in the bias terms $\Delta_z V_{k-1}$ and $\delta_s V_{k-1}$ in an exact manner. More precisely,

$$\begin{aligned}
\Delta_z V_k(n,a,s)=&\int_0^\infty dvQe^{-Qv}\Big\{\Big[R_v(n+1,a,s+z)-R_v(n,a,s)\Big]\\
&+\alpha\lambda(a+v)1(n+1<N)\Delta_z V_{k-1}(n+1,0,s+v)\\
&+\alpha\lambda(a+v)1(n+1=N)\delta_z V_{k-1}(n+1,0,s+v)\\
&+\alpha h(s+v)1(n>0)\Delta_0 V_{k-1}(n-1,a+v,0)\\
&+\alpha\Big[h(s+v+z)-h(s+v)\Big]1(n>0)\Big[-\delta_{s+v}V_{k-1}(n,a+v,0)\Big]\\
&+[1-\alpha\lambda(a+v)-\alpha h(s+v+z)]\Delta_z V_{k-1}(n,a+v,s+v)\Big\}. \qquad (3.7)
\end{aligned}$$

Note here, as announced in Section 3.1 that a difference of the form $\Delta_k V_k$, that is in number of jobs, also necessarily leads to a difference of the form $\delta_z V_{k-1}$, that is, in attained times. We thus necessarily have to analyze both.

To this end, along the same lines and leaving the details of the various steps to the reader, we can also derive a recursion relation for $\delta_z V_k$, which becomes

$$\begin{aligned}\delta_z V_k(n,a,s) = \int_0^\infty dvQe^{-Qv}\Big\{&\Big[R_v(n+1,a,s+z) - R_v(n,a,s)\Big]\\ &+\alpha\lambda(a+v)1(n<N)\delta_z V_{k-1}(n+1,0,s+v)\\ &+\alpha\lambda(a+v)1(n=N)\delta_z V_{k-1}(n,0,s+v)\\ &+\alpha\Big[h(s+v+z)-h(s+v)\Big]1(n>0)\Big[-\Delta_{s+v}V_{k-1}(n-1,a+v,0)\Big]\\ &+[1-\alpha\lambda(a+v)-\alpha h(s+v+z)1(n>0)]\delta_z V_{k-1}(n,a+v,s+v)\Big\}. \quad (3.8)\end{aligned}$$

3.2 Error bounds

In principle, by (3.7) and (3.8) we have exact expressions to evaluate the effect of perturbations in the arrival or service distribution, as based upon Lemma 2.2. However, this would require the exact (recursive) computation of (3.7) and (3.8) for sufficiently large k and all possible states (n,a,s). Clearly, this is of at least the same complexity as computing the stationary distribution Π itself, which is generally impossible.

We therefore aim to use the recursive expressions (3.7) and (3.8) to inductively prove bounds. In this section this will be illustrated for different performance measures of interest, as determined by different reward functions. Though the steps are similar for different measures, different technicalities are involved and require attention for each of the cases separately as conflicting terms appear (see the minus signs in (3.7) and (3.8)). The following cases will be considered.

Mean queue length	Reward rate: $r_1(n,a,s) = n$,
Tail probabilities	Reward rate: $r_1(n,a,s) = 1(n \geq \ell)$, for a given ℓ,
Throughput	Instantaneous reward: $r_2((n,a,s),(n',a',s')) = 1(n' = n-1 \geq 0, a' = a, s' = 0)$.

In what follows, in addition to the nondecreasing assumption on h, we frequently need further assumptions. To this end, we introduce the following two symbols.

$$\lambda^*_{\max} \equiv \sup_{a>0} \int_0^\infty \lambda(a+v)Qe^{-Qv}dv \quad (3.9)$$

$$h_Q \equiv \int_0^\infty h(v)Qe^{-Qv}dv \quad (3.10)$$

and assume that

$$\lambda^*_{\max} < h_Q. \quad (3.11)$$

This condition seems to be stronger than the stability condition,

$$\lambda \equiv \left(\int_0^\infty vf(v)dv\right)^{-1} < \mu \equiv \left(\int_0^\infty vg(v)dv\right)^{-1}, \quad (3.12)$$

for the infinite capacity queue, i.e., $N = \infty$. More specifically, we conjecture that, if h is nondecreasing, then, for h_Q defined by (3.10),

$$h_Q \leq \mu. \tag{3.13}$$

For instance, it can be proved that, if h is a single step function, i.e., $h(t) = h_1 + (h_2 - h_1)1(t \geq a)$ for constants $a > 0$ and $h_2 > h_1 > 0$, then (3.13) holds true (see Appendix A). In our experience, h_Q is relatively close to μ in many cases (e.g., see Section 6 and Appendix C). Condition (3.11) will thus be somehow stronger than the stability condition (3.12), and is thus somewhat restrictive for finite capacity queues. This is a cost to be paid for deriving analytical error bounds.

3.2.1 Mean queue length

Lemma 3.1 Let $r_1(n, a, s) \equiv n$, and suppose that condition (3.11) holds. Then,

$$\begin{aligned} |\Delta_z V_k(n, a, s)| &\leq (n+1)C, \qquad && n \leq N-1, && (3.14) \\ |\delta_z V_k(n, a, s)| &\leq nC, && n \leq N, && (3.15) \end{aligned}$$

where

$$C = \frac{1}{h_Q - \lambda^*_{\max}}\ . \tag{3.16}$$

PROOF. First note that R_v as per (2.2) with $r_1(n, a, s) = n$ is given by

$$R_v(n, a, s) = \int_0^v n du = nv.$$

Now let us apply induction in k. Clearly (3.14) and (3.15) hold for $k = 0$. Suppose that (3.14) and (3.15) are satisfied for $k = m$. Since

$$\int_0^\infty v dv Q e^{-Qv} = 1/Q = \alpha, \tag{3.17}$$

by substituting (3.14) and (3.15) for $k = m$ into (3.7), we find

$$\begin{aligned} |\Delta_z V_{m+1}(n, a, s)| \leq \int_0^\infty & dv Q e^{-Qv} \Big\{ \alpha(n+1) - \alpha n \\ & + \alpha\lambda(a+v)1(n+1 < N)(n+2)C \\ & + \alpha\lambda(a+v)1(n+1 = N)(n+1)C \\ & + \alpha h(s+v)1(n > 0)nC \\ & + \alpha\Big[h(s+v+z) - h(s+v)\Big]1(n > 0)nC \\ & + [1 - \alpha\lambda(a+v) - \alpha h(s+v+z)](n+1)C\Big\} \\ \leq (n+1)C + & \int_0^\infty dv Q e^{-Qv} \alpha[1 + \lambda(a+v)C - h(s+v+z)C] \\ \leq (n+1)C + & \int_0^\infty dv Q e^{-Qv} \alpha[1 + \lambda(a+v)C - h(v)C] \\ \leq (n+1)C, & \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad (3.18) \end{aligned}$$

where the last inequality follows from (3.16) and the assumption that $\lambda^*_{\max} < h_Q$. Similarly, by substituting (3.14) and (3.15) for $k = m$ into (3.8) yields, for $n > 0$,

$$\begin{aligned}
|\delta_z V_{m+1}(n,a,s)| \leq \int_0^\infty dvQe^{-Qv}\Big\{&\alpha(n+1) - \alpha n \\
&+\alpha\lambda(a+v)1(n<N)(n+1)C \\
&+\alpha\lambda(a+v)1(n=N)nC \\
&+\alpha\Big[h(s+v+z) - h(s+v)\Big]1(n>0)nC \\
&+[1-\alpha\lambda(a+v) - \alpha h(s+v+z)1(n>0)]nC\Big\}
\end{aligned}$$

$$\begin{aligned}
&\leq nC + \int_0^\infty dvQe^{-Qv}\alpha[1+\lambda(a+v)C - h(s+v)nC] \\
&\leq nC + \alpha\int_0^\infty dvQe^{-Qv}[1+\lambda(a+v)C - h(v)nC] \\
&\leq nC. \qquad (3.19)
\end{aligned}$$

As (3.15) trivially holds for $n = 0$ by definition of δ_z, we have thus proven (3.14) and (3.15) also for $k = m+1$. Induction completes the proof. □

We are now able to apply Lemma 2.2. This leads to the following result.

Theorem 3.1 Let L and $\overline{L}$ be the mean queue lengths (including a customer in service) of the $GI/G/1/N$ queues with hazard rates h and $\overline{h}$, respectively. Assume that the queues are stable and that L and $\overline{L}$ are finite. Let h be nondecreasing and $\lambda^*_{\max} < h_Q$. Then

$$|\overline{L} - L| \leq C\int_{(n,a,s),n>0}(n+1)\delta(s)d\overline{\Pi}(n,a,s), \qquad (3.20)$$

where C is given by (3.16), and $\delta(t) = |\overline{h}(t) - h(t)|$. Furthermore, if $\overline{h}(t)$ is nondecreasing and if $\delta(t) \leq \delta_0$ for all $t \geq 0$, then

$$\frac{|\overline{L} - L|}{L} \leq \overline{C}\delta_0\left(1 + \frac{1-\pi(0)}{L}\right), \qquad (3.21)$$

where $\overline{C} = (\overline{h}_Q - \overline{\lambda}^*_{\max})^{-1}$.

Proof. Note that $R_v(n,a,s) = nv = \overline{R}_v(n,a,s)$ for all (n,a,s). First suppose that (3.11) is satisfied. Then inequality (3.20) follows immediately from (3.1), (3.14) and Lemma 2.2. The remaining part is easily seen by applying the setting in Remark 2.2. □

Quality of the error bound for the $M/M/1$, $M/M/1/N$ and $M/G/1$ cases The error bound of (3.21) is numerically examined in Section 6. Below we investigate the error bounds analytically for the infinite and finite $M/M/1$ cases. First we study the infinite capacity case. Let $h(t) = \mu$ and $\overline{h}(t) = \mu + \delta_0$. Then, as is well-known, $L = \frac{\lambda}{\mu-\lambda}$. Hence, we find

$$\lim_{\delta_0 \downarrow 0} \frac{\overline{L} - L}{\delta_0} \quad = \quad \frac{\partial}{\partial \mu} \frac{\lambda}{\mu - \lambda} = -\frac{1}{\mu - \lambda} L.$$

Since $\lambda^*_{\max} = \lambda$ and $h_Q = \mu$, we thus get

$$\lim_{\delta_0 \downarrow 0} \frac{1}{\delta_0} \frac{|\overline{L} - L|}{L} = \frac{1}{\mu - \lambda} = C.$$

On the other hand, the corresponding ratio in (3.21) is bounded by

$$\lim_{\delta_0 \downarrow 0} \overline{C} \left(1 + \frac{1 - \pi(0)}{L} \right) = \lim_{\delta_0 \downarrow 0} \frac{1}{\mu + \delta_0 - \lambda} \left(1 + \frac{\mu - \lambda}{\mu} \right) = C \left(2 - \frac{\lambda}{\mu} \right).$$

Thus, for $\rho = \lambda/\mu$ sufficiently close to 1, the error bound in (3.21) has the same asymptotic ratio as δ_0 goes to zero.

Next consider a similar perturbation for the $M/M/1/N$ case with finite $N \geq 1$. Then, similar computations are in order. Let $\rho = \lambda/\mu$. To apply the error bound results, we assume that $\rho < 1$. The mean queue length L then becomes:

$$L = \frac{\rho}{1 - \rho} - \frac{(N+1)\rho^{N+1}}{1 - \rho^{N+1}}.$$

Hence,

$$-\frac{1}{L} \frac{\partial L}{d\mu} = \frac{1}{\mu - \lambda} \frac{\dfrac{\rho}{1 - \rho} - \dfrac{(N+1)\rho^{N+1}}{1 - \rho^{N+1}} \dfrac{(N+1)(1 - \rho)}{1 - \rho^{N+1}}}{\dfrac{\rho}{1 - \rho} - \dfrac{(N+1)\rho^{N+1}}{1 - \rho^{N+1}}}. \tag{3.22}$$

The limiting error bound is thus given by the right-hand side of (3.22). Since $1 - \rho^{N+1} = (1 - \rho)(1 + \rho + \ldots + \rho^N)$, it is not hard to see that

$$\frac{(N+1)\rho^{N+1}}{1 - \rho^{N+1}} \leq \frac{(N+1)\rho^{N+1}}{1 - \rho^{N+1}} \frac{(N+1)(1 - \rho)}{1 - \rho^{N+1}} \leq \frac{\rho}{1 - \rho}.$$

Hence,

$$0 \leq -\lim_{\delta_0 \downarrow 0} \frac{1}{\delta_0} \frac{\overline{L} - L}{L} \leq \frac{1}{\mu - \lambda} = C,$$

where 0 is attained as ρ goes to 1, while C is attained as ρ goes to 0. Thus, we still have a nice asymptotics, in particular, for small ρ.

Similar but somehow degraded quality is obtained for the $M/G/1/\infty$ queue. Since the computations are routine, we defer its consideration to Appendix B.

3.2.2 Tail probabilities of queue length distribution

Next, consider the tail probability $\Pi^{\mathrm{c}}(\ell) \equiv \sum_{n=\ell}^{\infty} \pi(n)$ for a given $\ell \geq 1$ with $\lambda^*_{\max}$ and h_Q as before.

Lemma 3.2 Let $h_Q > \lambda^*_{\max}$. Then, for all k and (n, a, s),

$$\begin{aligned} |\Delta_z V_k(n,a,s)| &\leq \phi(n), & (3.23)\\ |\delta_z V_k(n,a,s)| &\leq \phi(n), & (3.24) \end{aligned}$$

where

$$\phi(n) = \begin{cases} C_1 \rho_Q^{\ell-n}, & n < \ell, \\ C_2, & n \geq \ell, \end{cases} \qquad (3.25)$$

with

$$\rho_Q = \frac{\lambda^*_{\max}}{h_Q} < 1, \qquad C_1 = \frac{1}{h_Q \rho_Q (1-\rho_Q)}, \qquad \text{and} \qquad C_2 = \rho_Q C_1.$$

Proof. Again we use induction in k as in the proof of Lemma 3.1. First note that $R_v(n,a,s)$ as defined by (2.2) is given by

$$R_v(n,a,s) = v1(n \geq \ell).$$

Assume that (3.23) holds for $k \leq m$. Consider expression (3.7) together with the identity (3.17) for the induction. We need to distinguish four cases. Note that $\phi(\ell-1) = C_1\rho_Q = C_2$. Hence, for $n \geq \ell$, (3.23) is directly verified for any C_2 by

$$\begin{aligned} |\Delta_z V_{m+1}(n,a,s)| \leq \int_0^\infty dvQe^{-Qv}\Big\{ & \alpha\lambda(a+v)1(n+1<N)C_2 \\ & +\alpha\lambda(a+v)1(n+1=N)C_2 \\ & +\alpha h(s+v)C_2 \\ & +\alpha\Big[h(s+v+z)-h(s+v)\Big]C_2 \\ & +[1-\alpha\lambda(a+v)-\alpha h(s+v+z)]C_2\Big\} \leq C_2. \end{aligned}$$

For $n = \ell - 1$, we find

$$\begin{aligned} |\Delta_z V_{m+1}(n,a,s)| \leq \int_0^\infty dvQe^{-Qv}\Big\{ & \alpha + \alpha\lambda(a+v)C_2 \\ & +\alpha h(s+v)C_1\rho_Q^2 \\ & +\alpha\Big[h(s+v+z)-h(s+v)\Big]C_1\rho_Q \\ & +[1-\alpha\lambda(a+v)-\alpha h(s+v+z)]C_1\rho_Q\Big\} \\ \leq C_1\rho_Q + \int_0^\infty & dvQe^{-Qv}\alpha[1+\lambda(a+v)(C_2-C_1\rho_Q)-h(s+v)C_1\rho_Q(1-\rho_Q)] \\ \leq C_1\rho_Q + \alpha(& 1-h_QC_1\rho_Q(1-\rho_Q)) = C_1\rho_Q. \end{aligned}$$

For $1 \leq n < \ell - 1$, we similarly derive

$$\begin{aligned}|\Delta_z V_{m+1}(n,a,s)| \leq \int_0^\infty & dvQe^{-Qv}\Big\{\alpha\lambda(a+v)C_1\rho_Q^{\ell-n-1} \\ & +\alpha h(s+v)C_1\rho_Q^{\ell-n+1} \\ & +\alpha\Big[h(s+v+z)-h(s+v)\Big]C_1\rho_Q^{\ell-n} \\ & +[1-\alpha\lambda(a+v)-\alpha h(s+v+z)]C_1\rho_Q^{\ell-n}\Big\} \\ \leq C_1\rho_Q^{\ell-n} + C_1\rho_Q^{\ell-n} & \int_0^\infty dvQe^{-Qv}\alpha[\lambda(a+v)(1-\rho_Q)\rho_Q^{-1} - h(s+v)(1-\rho_Q)] \\ \leq C_1\rho_Q^{\ell-n} + \alpha C_1 & \rho_Q^{\ell-n}(1-\rho_Q)(\lambda^*_{\max}\rho_Q^{-1} - h_Q) = C_1\rho_Q^{\ell-n},\end{aligned}$$

since $\rho_Q = \lambda^*_{\max}/h_Q$. Finally, for $n = 0$, we find

$$\begin{aligned}|\Delta_z V_{m+1}(0,a,s)| \leq \int_0^\infty & dvQe^{-Qv}\Big\{\alpha\lambda(a+v)C_1\rho_Q^{\ell-1} \\ & +\alpha\Big[h(s+v+z)-h(s+v)\Big]C_1\rho_Q^{\ell} \\ & +[1-\alpha\lambda(a+v)-\alpha h(s+v+z)]C_1\rho_Q^{\ell}\Big\} \\ \leq C_1\rho_Q^{\ell} + C_1\rho_Q^{\ell} & \int_0^\infty dvQe^{-Qv}\alpha[\lambda(a+v)(1-\rho_Q)\rho_Q^{-1} - h(s+v)] \\ \leq C_1\rho_Q^{\ell} + \alpha C_1\rho_Q^{\ell} & (\lambda^*_{\max}(1-\rho_Q)\rho_Q^{-1} - h_Q) \\ = C_1\rho_Q^{\ell} + \alpha C_1\rho_Q^{\ell} & (h_Q(1-\rho_Q) - h_Q) \leq C_1\rho_Q^{\ell}.\end{aligned}$$

By induction the proof of (3.23) is hereby completed. Note that, in the above computation, we never explicitly used the condition that $\rho_Q < 1$. However, the condition is required for $C_1 > 0$.

To verify (3.24), the proofs for $n \geq \ell$ as well as $n \leq \ell - 1$ are checked more directly by noting that there is no reward rate difference in the two states to compare and by using that $\lambda^*_{\max}(1-\rho_Q)\rho_Q^{-1} = h_Q(1-\rho_Q)$. □

Theorem 3.2 Let $\Pi^{\mathrm{c}}(\ell)$ and $\overline{\Pi}^{\mathrm{c}}(\ell)$ be the steady state tail probabilities of the $GI/G/1/N$ queues with service hazard rate functions h and $\overline{h}$, respectively. Let $\delta(t) = |\overline{h}(t) - h(t)|$ and $\delta_0 = \sup_{t>0}|\overline{h}(t) - h(t)|$. Then, if $h(t)$ is nondecreasing in t and if $\lambda^*_{\max} < h_Q$, then

$$\begin{aligned}\left|\overline{\Pi}^{\mathrm{c}}(\ell) - \Pi^{\mathrm{c}}(\ell)\right| \; &\leq \; \int_{n,a,s} d\overline{\Pi}(n,a,s)\phi(n)\delta(s), \\ &\leq \; \delta_0 C_1\Big[\sum_{k=0}^{\ell-1}\overline{\pi}(k)\rho_Q^{\ell-k} + \rho_Q\overline{\Pi}^{\mathrm{c}}(\ell)\Big], \qquad (3.26)\end{aligned}$$

where ϕ is given by (3.25), $\rho_Q = \lambda^*_{\max}/h_Q$ and $C_1 = 1/h_Q[\rho_Q(1-\rho_Q)]$.

Proof. Again, this is immediate by combining Lemma 2.2, the difference relation (3.1) and the inequalities (3.23) and (3.24). □

Quality of the error bound for the $M/M/1$ and $M/M/1/N$ cases Also here we test the quality of the error bound for the $M/M/1/\infty$ case. Let $h(t) \equiv \mu$ and $\overline{h}(t) = \mu + \delta(t)$ with $|\delta(t)| \le \delta_0$. In this case, $\lambda^*_{\max} = \lambda$, $h_Q = \mu$ and $\rho_Q = \lambda/\mu = \rho$. Since $\Pi^{\mathrm{c}}(\ell) = \rho^\ell$, we find

$$\lim_{\delta_0 \downarrow 0} \frac{|\overline{\Pi}^{\mathrm{c}}(\ell) - \Pi^{\mathrm{c}}(\ell)|}{\delta_0} = \left| \frac{\partial}{\partial \mu} \left(\frac{\lambda}{\mu} \right)^\ell \right| = \frac{\ell}{\mu} \rho^\ell.$$

On the other hand, from (3.26) and fact that $\pi(k) = (1-\rho)\rho^k$, the corresponding rate of the error bound is

$$\begin{aligned} \lim_{\delta_0 \downarrow 0} C_1 \Big[\sum_{k=0}^{\ell-1} \overline{\pi}(k) \rho^{\ell-k} + \rho \overline{\Pi}^{\mathrm{c}}(\ell) \Big] &= \frac{1}{\mu\rho(1-\rho)} (\ell(1-\rho) + \rho)\rho^\ell \\ &= \frac{\ell \rho^{\ell-1}}{\mu} \left(1 + \frac{\rho}{\ell(1-\rho)} \right). \end{aligned}$$

The error bound is thus degraded by the factor ρ^{-1} from the exact rate for either large ℓ or small ρ. However, it still reveals the exact proportionality with respect to $\ell\rho^\ell$.

We next consider the same perturbation for the $M/M/1/N$ with finite N. Since

$$\overline{\Pi}(\ell) = \frac{\rho^\ell (1 - \rho^{N-\ell+1})}{1 - \rho^{N+1}},$$

we have

$$\begin{aligned} -\frac{\partial}{\partial \mu} \overline{\Pi}(\ell) &= -\frac{\partial}{\partial \mu} \frac{\mu^{N+1-\ell}\lambda^\ell - \lambda^{N+1}}{\mu^{N+1} - \lambda^{N+1}} \\ &= \frac{\ell \rho^\ell}{\mu(1-\rho^{N+1})} - \frac{(N+1)(1-\rho^\ell)\rho^{N+1}}{\mu(1-\rho^{N+1})^2}. \end{aligned}$$

On the other hand, the error bound of (3.26) is

$$\lim_{\delta_0 \downarrow 0} C_1 \Big[\sum_{k=0}^{\ell-1} \overline{\pi}(k) \rho^{\ell-k} + \rho \overline{\Pi}^{\mathrm{c}}(\ell) \Big] = \frac{\ell \rho^{\ell-1}}{\mu(1-\rho^{N+1})} \left(1 + \frac{\rho(1-\rho^{N+1-\ell})}{\ell(1-\rho)} \right).$$

Thus, for the finite capacity case, the quality of the asymptotic error bound is degraded, but still keep a similar property to the infinite case.

3.2.3 Throughput

Finally, let us consider the throughput by setting

$$r_2((n,a,s),(n',a',s')) = 1(n' = n-1 \ge 0).$$

Then with $R_v(n,a,s)$ of (2.2) we find, using $\alpha = Q^{-1}$,

$$R_v(n,a,s) = h(s+v)\alpha 1(n > 0).$$

Lemma 3.3 For all k, z and (n,a,s), we have

$$0 \leq \Delta_z V_k(n,a,s) \leq 1, \tag{3.27}$$

$$0 \leq \delta_z V_k(n,a,s) \leq 1. \tag{3.28}$$

PROOF. Again, the steps of the proof of Lemma 3.1 are followed. That is, by assuming (3.27) and (3.28) for $k=m$, we obtain, for $k=m+1$,

$$\begin{aligned}\Delta_z V_{m+1}(n,a,s) \leq \int_0^\infty & dvQe^{-Qv}\Big\{[h(s+v+z)-h(s+v)]\alpha \\ & +\alpha\lambda(a+v)+\alpha h(s+v) \\ & +[1-\alpha\lambda(a+v)-\alpha h(s+v+z)]\Big\} \leq 1.\end{aligned}$$

Here, it is used that $-\delta_{s+v}V_m(n,a,s) \leq 0$ by (3.28). Conversely, also the lower bound 0 of (3.27) can be proven for $k=m+1$, by noting that, for $n>0$,

$$\begin{aligned}& [h(s+v+z)-h(s+v)]/Q \\ & \quad +\alpha\Big[h(s+v+z)-h(s+v)\Big][-\delta_{s+v}V_m(n,a+v,0)] \\ & = \alpha\Big[h(s+v+z)-h(s+v)\Big][1-\delta_{s+v}V_m(n,a+v,0)] \geq 0.\end{aligned}$$

The proof of (3.28) for $k=m+1$ goes similarly by noting the fact that for $n>0$ (see (3.8))

$$\begin{aligned}& [h(s+v+z)-h(s+v)]/Q \\ & \quad +\alpha\Big[h(s+v+z)-h(s+v)\Big][-\Delta_{s+v}V_m(n,a+v,0)] \geq 0.\end{aligned}$$

The induction in m then completes the proof. □

Remark 3.1 (Estimates for bias terms) The lower estimates 0 in (3.27) and (3.28) are not developed to obtain monotonicity results but purely to establish the upper bounds 1 in (3.27) and (3.28). The monotonicity is simply a side result.

Theorem 3.3 Let T and $\overline{T}$ be the throughputs of the $GI/G/1/N$ queues with service hazard rate functions h and $\overline{h}$, respectively. Let $\delta(t)=|\overline{h}(t)-h(t)|$ and $\delta_0=\sup_{t>0}|\overline{h}(t)-h(t)|$. Then, if $h(t)$ is nondecreasing in t, then

$$\begin{aligned}|\overline{T}-T| & \leq \int_{n>0,a,s} d\overline{\Pi}(n,a,s)\delta(t) \\ & \leq \delta_0(1-\overline{\pi}(0)).\end{aligned} \tag{3.29}$$

Furthermore, if $\overline{h} \geq (\leq) h$, then

$$\overline{T} \geq (\leq) T. \tag{3.30}$$

Proof. First recall that $R_v(n,a,s) = 1(n>0)h(s+v)/Q$. Hence, $Z_v(n,a,s)$ as per (3.1) becomes

$$Z_v(n,a,s) = [\overline{h}(s+v) - h(s+v)]Q^{-1}\{1 - \Delta_{s+v}V_k(n,a,s)\}.$$

The proof now follows immediately from Lemmas 2.2 and 3.3. □

Remark 3.2 By assuming $|\overline{h}(t) - h(t)| \leq \delta_1 h(t)$ and that $\overline{h}(t)$ instead of $h(t)$ is nondecreasing in t, Lemma 2.2 implies that (3.29) can be replaced by

$$\frac{|\overline{T} - T|}{T} \leq \delta_1. \tag{3.31}$$

Quality of the error bound for the $M/M/1/N$ case As before, to investigate the quality of the error bound, we consider the $M/M/1/N$ case. Let $h(t) \equiv \mu$ and $\overline{h}(t) = \mu(1+\delta(t))$ with $|\delta(t)| \leq \delta_1$. Let $\rho = \lambda/\mu$. Since

$$T = \lambda \frac{1-\rho^N}{1-\rho^{N+1}}.$$

we find, for $\rho \neq 1$,

$$\begin{aligned}
\lim_{\delta_1 \downarrow 0} \frac{|\overline{T} - T|}{\delta_1} &= \lambda \left.\frac{\partial}{\partial x} \frac{(x+1)^{N+1} - (x+1)\rho^N}{(x+1)^{N+1} - \rho^{N+1}}\right|_{x=0} \\
&= \frac{\lambda \rho^N}{(1-\rho^{N+1})^2} \big((N+1)(1-\rho) - (1-\rho^{N+1})\big) \\
&= T \frac{\rho^N}{(1-\rho^N)(1-\rho^{N+1})} \big((N+1)(1-\rho) - (1-\rho^{N+1})\big). \qquad (3.32)
\end{aligned}$$

Comparing (3.32) with (3.31), it is easy to see that the general error bound result is not tight for small ρ. However, the error bound (3.31) becomes asymptotically tight as either ρ goes to infinity or N goes to infinity for $\rho > 1$, since the last term of the above equation converges to T.

Consequently, the error bound (3.31) is not generally tight. This is intuitively clear, since the throughput becomes insensitive with respect to the hazard rate when the loss probability is small. The sensitive part may be detected by the loss probability, which determines the throughput. In other words, to get the error bound on T, it is as hard as to get an error bound for the loss probability. Therefore, let us check how the error bound is changed if we use the error bound for the tail probability. From (3.26) for the $M/M/1/N$ queue, we derive

$$\big|\overline{T} - T\big| = \lambda \big|\overline{\Pi}^{\mathrm{c}}(N) - \Pi^{\mathrm{c}}(N)\big| \leq \lambda \delta_0 C_1 \Big[\sum_{k=0}^{N-1} \overline{\pi}(k)\rho^{N-k} + \rho\overline{\pi}(N)\Big].$$

Hence, the asymptotic ratio of the error bound under the perturbation $\overline{h}(t) = \mu(1+\delta(t))$ with $|\delta(t)| \leq \delta_1$ is, for $\rho < 1$, noting the relation $\delta_0 = \mu\delta_1$,

$$\begin{aligned}\lim_{\delta_1\downarrow 0} \lambda\frac{\delta_0}{\delta_1}C_1\Big[\sum_{k=0}^{N-1}\overline{\pi}(k)\rho^{N-k} + \rho\overline{\pi}(N)\Big] &= \lim_{\delta_1\downarrow 0}\frac{\lambda}{\rho(1-\rho)}\Big[\sum_{k=0}^{N-1}\overline{\pi}(k)\rho^{N-k} + \rho\overline{\pi}(N)\Big] \\ &= T\frac{\rho^{N-1}(N+\rho)}{(1-\rho^N)(1-\rho)}. \qquad (3.33)\end{aligned}$$

Comparing (3.33) result with (3.32), we see that the error bound for the tail probability improves the error bound on the throughput when ρ is small.

The advantage of the error bound (3.31) is its simplicity and generality, where the extra condition (3.11) is not required, while still keeping the nice asymptotic tightness for some special cases. These are the type of properties that will be studied in the next section.

4. $M/G/c/N$ queues

With reference to the discussion at the beginning of Section 3, in this section we will investigate a more complex nonexponential multi-server queue case with a perturbation of the service time distribution. In this section, the arrival process is assumed to be Poisson. Furthermore, we restrict ourselves to the throughput for the following two reasons. In the first place, as it is a most natural measure of practical interest. In the second place, it keeps the technical details as transparent as possible, because the bounding functions will appear to be constant as in Lemma 3.3. We follow the steps from Section 3.

4.1 Formulation and perturbation

Consider an $M/G/c/N$-queue with Poisson arrival rate λ, c servers and a finite constraint for at most N jobs (customers) in total. We are interested in the throughput of the queue:

$$T = \lambda(1-B)$$

where B is the loss probability of arriving jobs, or equivalently, the expected number of service completions per unit time in the long run. As in Section 3, we aim to investigate the effect on T when the service distribution is perturbed.

Perturbation
Now consider the $M/G/c/N$ queue with the service hazard rate h perturbed into $\overline{h}$, while the arrival rate λ and queue capacity N are assumed to be the same.

As we do not need to keep track of the elapsed interarrival time while it is also more convenient to keep track of how many jobs are waiting, instead of the notation $(n, a, \boldsymbol{s})$ we use the notation

$(\boldsymbol{s}, w)$ with $w = (n - c)^+$; the number of waiting jobs.

All notation from Section 2 is adopted accordingly, e.g., $V_k(\boldsymbol{s}, w)$, $R_v(\boldsymbol{s}, w)$ and $P((\boldsymbol{s}, w), (\boldsymbol{s}', w'))$. As in Section 3.2.3, in order to evaluate the throughput we use the instantaneous reward

$$\begin{aligned} r_2((\boldsymbol{s}, w), (\boldsymbol{s}', w')) &= 1(w' = w - 1 \geq 0, \boldsymbol{s}' = \boldsymbol{s} \ominus \boldsymbol{u}_i \oplus \boldsymbol{u}_i(0) \text{ for some } i = 1, 2, \ldots, \ell) \\ &\quad +1(w' = w = 0, \boldsymbol{s}' = \boldsymbol{s} \ominus \boldsymbol{u}_i \text{ for some } i = 1, 2, \ldots, \ell). \end{aligned}$$

Hence, by (2.2), $R_v(\boldsymbol{s}, w)$ becomes

$$R_v(\boldsymbol{s}, w) = \sum_{i=1}^{\ell} h(s_i + v)/Q.$$

By writing out $Z_v(\boldsymbol{s}, w)$ as per (2.9) with the modified notation $(\boldsymbol{s}, w)$, we then obtain

$$\begin{aligned} Z_v(\boldsymbol{s}, w) = \sum_{i=1}^{\ell} &[\overline{h}(s_i + v) - h(s_i + v)]Q^{-1} \\ &\times \Big\{ 1 + 1(w = 0)[V_k(\boldsymbol{s}_v \ominus \boldsymbol{u}_i, 0) - V_k(\boldsymbol{s}_v, 0)] \\ &\quad +1(w > 0)[V_k(\boldsymbol{s}_v \ominus \boldsymbol{u}_i \oplus \boldsymbol{u}_i(0), w - 1) - V_k(\boldsymbol{s}_v, w)] \Big\}. \end{aligned} \tag{4.1}$$

To estimate or bound the bias terms, other bias terms will get involved too, as will be expressed in detail in the next section. Once these bias terms are estimated, the error bound theorem can be applied.

4.2 Error bound for throughput

As in Section 3, the second and major step is to bound the bias terms. This is established in the next lemma.

Lemma 4.1 Assume that h is nondecreasing. Then, for all $(\boldsymbol{s}, w)$, s, i and $k \geq 0$,

$$\begin{aligned} 0 \leq {}^1\!\boldsymbol{\Delta}_i^k(\boldsymbol{s}, 0) &:= V_k(\boldsymbol{s}, 0) - V_k(\boldsymbol{s} \ominus \boldsymbol{u}_i, 0) \leq 1 && (n \leq c) && (4.2) \\ 0 \leq {}^2\!\boldsymbol{\Delta}_i^k(\boldsymbol{s}, w)[t] &:= V_k(\boldsymbol{s}, w) - V_k(\boldsymbol{s} \ominus \boldsymbol{u}_i \oplus \boldsymbol{u}_i(t), w) \leq 1 && (t \leq s_i) && (4.3) \\ 0 \leq {}^3\!\boldsymbol{\Delta}_i^k(\boldsymbol{s}, w)[t] &:= V_k(\boldsymbol{s}, w) - V_k(\boldsymbol{s} \ominus \boldsymbol{u}_i \oplus \boldsymbol{u}_i(t), w - 1) \leq 1 && (t \leq s_i, w > 0)\,. && (4.4) \end{aligned}$$

Inequalities (4.2)-(4.4) reflect that the expected total number of departing customers is neither decreased nor increased more than one, if the initial state is altered by either one of the following changes. One customer is reduced, one attained service time is reduced, or both of them occurs. Since h is nondecreasing, inequalities (4.2) and (4.3) may seem intuitively obvious. However, because of the finite capacity, one has to be most careful (for example, standard sample path arguments will fail due to overtaking), and formal proofs are needed. Furthermore, (4.4) is not obvious even intuitively. We defer their proofs to Appendix C, since they are highly technical.

Let $\overline{G}$ be the service time distribution of the modified queue, and suppose that $\overline{G}$ has a hazard rate function $\overline{h}$.

With ℓ the dimension of the vector $\boldsymbol{s}$, i.e., the number of busy servers, as implicitly assumed throughout this section, let, for a real valued function δ,

$$(\sigma \circ \delta)(\boldsymbol{s}) = 1(\ell > 0) \sum_{j=1}^{\ell} \delta(s_j).$$

Theorem 4.1 Assume that the service time distribution has a nondecreasing hazard rate function h for the $M/G/c/N$ queue, and a hazard rate function $\overline{h}$ for the modified $M/G/c/N$ queue. Let T be the throughput of the original system and $\overline{T}$ of the modified system. Then,

i) $h(t) \geq (\leq)\, \overline{h}(t), \quad t \geq 0, \quad \Longrightarrow \quad T \geq (\leq)\, \overline{T}.$

ii) $|h(t) - \overline{h}(t)| \leq \delta(t), \quad t \geq 0$, for some $\delta : I\!R \to I\!R \quad \Longrightarrow \quad |T - \overline{T}| \leq \langle \overline{\Pi}, \sigma \circ \delta \rangle.$

Proof. Consider expression (4.1) for $Z_v(\boldsymbol{s}, w)$. Now note that by Lemma 4.1 the term between the braces in (4.1) is bounded between 0 and 1 as the difference terms are bounded between -1 and 0. Lemma 2.2 and (4.1) then complete the proof. □

Remark 4.1 The monotonicity result in i) is also proven in [13].

Corollary 4.1 With nondecreasing hazard rate function $\overline{h}$ and nonnegative number δ_1,

$$\left| h(t) - \overline{h}(t) \right| \leq \delta_1 h(t) \text{ for all } t \geq 0 \quad \Longrightarrow \quad \left| \frac{T - \overline{T}}{T} \right| \leq \delta_1 \,. \tag{4.5}$$

Proof. Immediate from Theorem 4.1 by changing the roles of the original and perturbed models (see Remark 2.2). □

The following corollary is of interest when service time distributions are compared which are not purely ordered or monotone in hazard rate but up to some (small) discrepancy (function). This corollary can be seen as a relaxation of strict comparison results such as in [13].

Corollary 4.2 Under the setting of Theorem 4.1 without assuming h to be nondecreasing, if $\overline{h} + \delta$ is nonnegative and nondecreasing for some function δ, then

$$\begin{aligned} h \geq \ \overline{h} + \delta &\Longrightarrow T \geq \overline{T} - \langle \overline{\Pi}, \sigma \circ |\delta| \rangle, \\ h \leq \ \overline{h} + \delta &\Longrightarrow T \leq \overline{T} + \langle \overline{\Pi}, \sigma \circ |\delta| \rangle. \end{aligned} \tag{4.6}$$

PROOF. Define $\tilde{h} = \overline{h} + \delta$, then we can apply Theorem 4.1 for this $\tilde{h}$ instead of h and for h instead of $\overline{h}$. Thus, i) of Theorem 4.1 leads to $T \geq (\leq)\ \tilde{T}$, where $\tilde{T}$ is the throughput for the service time distribution with hazard rate $\tilde{h}$. We next apply Theorem 4.1 for $\tilde{h}$ and $\overline{h}$. Then, by ii), we have $|\tilde{T} - \overline{T}| \leq \langle \Pi, \sigma \circ |\delta| \rangle$. Combining these two inequalities, we have (4.6). □

Quality of the error bound for the $M/M/c/c$ case As already discussed for the single server queue in Section 3.2.3, the error bound (4.5) is not generally tight. For the $M/M/c/c$ loss system under the perturbation of Corollary 4.1, we have

$$\lim_{\delta_1 \downarrow 0} \frac{|\overline{T} - T|}{\delta_1} \quad = \quad T \frac{\frac{\rho^c}{c!}\left(cB(c-1) - \rho B(c-2)\right)}{B(c)B(c-1)}, \tag{4.7}$$

where $\rho = \lambda/\mu$ and $B(n) = \sum_{i=0}^{n} \frac{\rho^i}{i!}$. However, for the $M/M/c/N$ with $\rho > c$, if N goes to infinity, we can still check that the error bound (4.5) is asymptotically exact as a ratio with respect to δ_1. This situation is similar to the single server case.

5. $GI/M/c/N$

5.1 Perturbation

As last application, we aim to show that also a perturbation for a nonexponential arrival distribution can be dealt with by the results from Section 2. To merely focus on this aspect, we consider a $GI/M/c/N$ queue, with arrival distribution F with hazard rate function $\lambda(t)$ and exponential service distribution with mean $1/\mu$. We perturb this arrival rate function $\lambda(t)$, and denote the perturbed arrival rate function by $\overline{\lambda}(t)$. Since the service distribution is assumed to be exponential, we can omit the attained service times in the system state description. The state can thus be denoted by (n, a), where n represents the number of customers in the system and a the attained interarrival time. Then, in the setting of Section 2;

$$P((n,a),(n',a')) = \begin{cases} Q^{-1}\lambda(a), & \begin{cases} n' = n+1 \leq N, a' = 0, \\ n' = n = N, a' = 0, \end{cases} \\ Q^{-1}\min(n,c)\mu & \begin{cases} n' = n = 0, a' = a, \\ n' = n-1 \geq 0, a' = a, \end{cases} \\ 1 - Q^{-1}[\lambda(a) + \min(n,c)\mu], & (n',a') = (n,a). \end{cases}$$

Accordingly, Z_v^k of (2.9) here becomes

$$\begin{aligned} Z_v^k(n,a) = {} & [\overline{R}_v - R_v](n,a) \\ & + Q^{-1}\Big[\overline{\lambda}(a+v) - \lambda(a)\Big]\Big[V_k(n+1,0) - V_k(n,a+v)\Big]1(n < N) \\ & + Q^{-1}\Big[\overline{\lambda}(a+v) - \lambda(a)\Big]\Big[V_k(n,0) - V_k(n,a+v)\Big]1(n = N). \end{aligned} \tag{5.1}$$

As before, in order to apply Lemma 2.2, we will first derive bounds for the bias terms $V_k(n+1,0) - V_k(n,a+v)$ as well as $V_k(n,a) - V_k(n,a+v)$.

5.2 Bias terms

We only consider the mean queue length. The throughput and tail probabilities can be obtained similarly. In all three cases, under the assumption that $\lambda(t)$ is nondecreasing in t, we can establish the following bounds for the bias terms if a function Φ is appropriately chosen.

$$0 \leq \Delta_z^* V_k(n,a) \equiv V_k(n+1,a) - V_k(n,a+z) \leq \Phi(n), \qquad n \leq N-1, \tag{5.2}$$

$$0 \leq \delta_z^* V_k(n,a) \equiv V_k(n,a+z) - V_k(n,a) \leq \Phi(n), \qquad n \leq N. \tag{5.3}$$

Here, the lower bounds 0 are easily seen by the stochastic monotonicity of the queue length process with respect to the interarrival times, which can be formally proved in a standard way by sample path arguments. We therefore verify only the upper bounds.

As in Sections 3 and 4, the first step for the verification is to establish recursive relations for the bias terms. These will be used in the subsequent section. For $n < N$, we have

$$\begin{aligned} V_k(n+1,a) &= \int_0^\infty dv Q e^{-Qv} \Big[R_v(n+1,a+v) \\ &\quad + \alpha\lambda(a+v) 1(n+1<N) V_{k-1}(n+2,0) \\ &\quad + \alpha\lambda(a+v) 1(n+1=N) V_{k-1}(n+1,0) \\ &\quad + \alpha \min(n+1,c)\mu V_{k-1}(n,a+v) \\ &\quad + [1-\alpha\lambda(a+v) - \min(n+1,c)\mu] V_{k-1}(n+1,a+v) \Big], \end{aligned} \tag{5.4}$$

and

$$\begin{aligned} V_k(n,a+z) &= \int_0^\infty dv Q e^{-Qv} \Big[R_v(n,a+v+z) \\ &\quad + \alpha\lambda(a+v+z) 1(n<N) V_{k-1}(n+1,0) \\ &\quad + \alpha\lambda(a+v+z) 1(n=N) V_{k-1}(n,0) \\ &\quad + \alpha \min(n,c)\mu V_{k-1}(n-1,a+v+z) \\ &\quad + [1-\alpha\lambda(a+v+z) - \min(n,c)\mu] V_{k-1}(n,a+v+z) \Big]. \end{aligned} \tag{5.5}$$

Subtracting (5.5) from (5.4) leads to, for $n < N$,

$$\begin{aligned}
\Delta^*_z V_k(n,a) &= V_k(n+1,a) - V_k(n,a+z) \\
&= \int_0^\infty dv Q e^{-Qv} \Big[R_v(n+1,a+v) - R_v(n,a+v+z) \\
&\quad +\alpha\lambda(a+v+z) 1(n+1<N) \Delta^*_z V_{k-1}(n+1,0) \\
&\quad +\alpha(\lambda(a+v+z) - \lambda(a+v)) \delta^*_{a+v} V_{k-1}(n+1,0) \\
&\quad +\alpha \min(n,c)\mu \Delta^*_z V_{k-1}(n-1,a+v) \\
&\quad -\alpha\mu 1(n \le c-1) \delta^*_z V_{k-1}(n,a+v) \\
&\quad +[1-\alpha\lambda(a+v+z) - \min(n+1,c)\mu] \Delta^*_z V_{k-1}(n,a+v) \Big]. \qquad (5.6)
\end{aligned}$$

Taking the difference of (5.5) with $z > 0$ and with $z = 0$, we obtain, for $n \le N$,

$$\begin{aligned}
\delta^*_z V_k(n,a) &= V_k(n,a+z) - V_k(n,a) \\
&= \int_0^\infty dv Q e^{-Qv} \Big[R_v(n,a+v+z) - R_v(n,a+v) \\
&\quad +\alpha(\lambda(a+v+z) - \lambda(a+v)) 1(n<N) \Delta^*_{a+v} V_{k-1}(n,0) \\
&\quad +\alpha \min(n,c)\mu \delta^*_z V_{k-1}(n-1,a+v) \\
&\quad +[1-\alpha\lambda(a+v+z) - \min(n,c)\mu] \delta^*_z V_{k-1}(n,a+v) \Big]. \qquad (5.7)
\end{aligned}$$

5.3 Error bound

Let $\lambda_{\max} = \sup_{t\ge 0} \lambda(t)$. Then, for the queue length, we have the following result.

Lemma 5.1 Assume that $\lambda(t)$ is nondecreasing in t, and $c\mu > \lambda_{\max}$. Let $r_1(n,a) \equiv n$, then (5.2) and (5.3) hold with

$$\Phi(n) = \begin{cases} D_1, & n < c, \\ D_1 + (n-c+1)D_2, & n \ge c, \end{cases} \qquad (5.8)$$

where $D_1 = c/(c\mu - \lambda_{\max})$ and $D_2 = 1/(c\mu - \lambda_{\max})$.

Remark 5.1 The condition $c\mu > \lambda_{\max}$ is stronger than the stability condition $c\mu > \lambda^*$ for the infinite capacity queue, where $(\lambda^*)^{-1} = \int_0^\infty f(v)/\lambda(v)dv$, so $\lambda^*_{\max} \ge \lambda^*$. Of course, such a stability condition is not needed for the stability of the finite capacity queue. Here, we again have to pay a price to obtain analytical error bounds.

Proof. Instead of Φ given above, to seek a better bound, we first set

$$\Phi(n) = \begin{cases} D_1, & n < c, \\ D_1 + D_3 + (n-c)D_2, & n \ge c, \end{cases} \qquad (5.9)$$

with arbitrary nonnegative numbers D_1, D_2 and D_3. It will turn out that $D_3 = D_2$ and the given D_1 and D_2 are only possible choices. As before, we prove (5.2) and (5.3) by induction in k. They trivially hold for $k = 0$. Assume that (5.2) and (5.3) hold for $k = m$. Below we only check (5.2), since (5.3) can be obtained similarly. To this end, we substitute the bound of (5.9) into (5.6) with $k = m + 1$. We need to consider four cases separately.

i) $\underline{n \geq c+1}$

$$\begin{aligned}
\Delta_z^* V_{m+1}(n,a) \leq \int_0^\infty & dvQe^{Qv}dv\Big[\alpha \\
& +\alpha\lambda(a+v+z)\left(D_1+D_3+(n+1-c)D_2\right) \\
& +\alpha(\lambda(a+v+z)-\lambda(a+v))\left(D_1+D_3+(n+1-c)D_2\right) \\
& +\alpha c\mu\left(D_1+D_3+(n-1-c)D_2\right) \\
& +(1-\alpha\lambda(a+v+z)-\alpha c\mu)\left(D_1+D_3+(n-c)D_2\right)\Big]
\end{aligned}$$

$$\begin{aligned}
&= \int_0^\infty dvQe^{Qv}dv\Big[\alpha(1+\lambda(a+v+z)D_2-c\mu D_2)+D_1+D_3+(n-c)D_2\Big] \\
&\leq D_1+D_3+(n-c)D_2+\alpha(1-c\mu D_2+\lambda_{\max}D_2).
\end{aligned}$$

Thus, the bound is established only if $D_2 \geq 1/(c\mu - \lambda_{\max})$.

ii) $\underline{n = c}$ Similarly to case i), we have

$$\begin{aligned}
\Delta_z^* V_{m+1}(n,a) \leq \int_0^\infty & dvQe^{Qv}dv\Big[\alpha \\
& +\alpha\lambda(a+v+z)\left(D_1+D_3+D_2\right) \\
& +\alpha(\lambda(a+v+z)-\lambda(a+v))\left(D_1+D_3+D_2\right) \\
& +\alpha c\mu D_1 \\
& +(1-\alpha\lambda(a+v+z)-\alpha c\mu)\left(D_1+D_3\right)\Big] \\
\leq\ & D_1+D_3+\alpha(1-c\mu D_3+\lambda_{\max}D_2).
\end{aligned}$$

Thus, the bound is established only if

$$D_3 \geq \frac{1+\lambda_{\max}D_2}{c\mu} \geq \frac{1}{c\mu-\lambda_{\max}} = D_2.$$

iii) $\underline{n = c-1}$

$$\begin{aligned}
\Delta_z^* V_{m+1}(n,a) \leq \int_0^\infty & dvQe^{Qv}dv\Big[\alpha \\
& +\alpha\lambda(a+v+z)\left(D_1+D_3\right) \\
& +\alpha(\lambda(a+v+z)-\lambda(a+v))\left(D_1+D_3\right) \\
& +\alpha(c-1)\mu D_1 \\
& +(1-\alpha\lambda(a+v+z)-\alpha c\mu)D_1\Big] \\
\leq\ & D_1+\alpha(1-c\mu D_1+\lambda_{\max}D_3).
\end{aligned}$$

Thus, the bound is established only if

$$D_1 \geq \frac{1+\lambda_{\max} D_2}{\mu} \geq \frac{c}{c\mu - \lambda_{\max}}.$$

iv) $\underline{n \leq c-1}$

$$\begin{aligned}\Delta_z^* V_{m+1}(n,a) \leq & \int_0^\infty dv Q e^{Qv} dv \Big[\alpha \\ & +\alpha\lambda(a+v+z)D_1 \\ & +\alpha(\lambda(a+v+z)-\lambda(a+v))D_1 \\ & +\alpha n\mu D_1 \\ & +(1-\alpha(n+1)\mu)D_1\Big] \\ \leq & \; D_1+\alpha(1-(n+1)\mu D_1).\end{aligned}$$

Thus, the bound is established only if $D_1 \geq 1/\mu \geq c/(c\mu - \lambda_{\max})$. Consequently, the induction hypotheses (5.2) with Φ of (5.8) is verified for $k = m+1$. This completes the proof. □

Lemma 5.1 together with the error bound result of Lemma 2.2 leads to the following error bound results.

Theorem 5.1 Under the conditions of Lemma 5.1, for L and $\overline{L}$ the mean queue length of the $GI/M/c/N$ queue with arrival hazard rates λ and $\overline{\lambda}$, respectively, and with $\delta(t) = |\overline{\lambda}(t) - \lambda(t)|$, we have

$$|\overline{L} - L| \leq D_1 \int_{(n,a)} \delta(a) d\overline{\Pi}(n,a) + D_2 \int_{(n,a)} (n-c+1)\delta(a) d\overline{\Pi}(n,a). \tag{5.10}$$

Furthermore, if $\overline{\lambda}(t)$ is nondecreasing and if $\delta(t) \leq \delta_0$ for all $t \geq 0$, then

$$|\overline{L} - L| \leq \delta_0 \left(\overline{D}_1 + \overline{D}_2(L_q + \overline{\Pi}^c(c))\right). \tag{5.11}$$

As one may expect, the error bounds of (5.10) and (5.11) are not as sharp as in Section 3. The technical reason for this is that in the induction proof of Lemma 5.1 we have to omit the negative bias term that appears in the one before the last term in the right hand side of (5.6) if $c \geq 2$. This apparently degrades the upper bounds for the bias terms.

We finally note an error bound for the throughput. In this case, we can choose $\Phi(n) \equiv 1$. Hence, with the notation from Theorem 5.1 and Section 4, if $\lambda(t)$ is nondecreasing in t, then

$$|\overline{T} - T| \leq \int_{(n,a)} \delta(a) d\overline{\Pi}(n,a) = \frac{1}{m_{\overline{F}}} \int_0^\infty \delta(a)(1-\overline{F}(a)) da, \tag{5.12}$$

where $m_{\overline{F}}$ is the mean interarrival time of the perturbed system, where we use the fact that the marginal distribution of the attained arrival time in steady state is known to be the stationary forward recurrence time distribution. This simple bound seems intuitively obvious. Nevertheless, its formal verification was not obvious.

6. Some numerical results

In addition to the qualitative and asymptotic investigations of the quality of the error bounds as in Sections 3.2.1, 3.2.2 and 3.2.4 for the $GI/G/1/N$ case, and in Section 4.2 for the $M/G/c/N$ case, in this section also some numerical support will be provided. We aim to investigate the value of the error bound such as to see

- how it relates to the exact error in order,
- whether it can be of practical use.

To this end, we consider a perturbation of the service distribution G and for a few different situations of the arrival distribution F. Let F and G have hazard rates h_f and h_g, respectively. Assume that h_f and h_g are step functions as specified by, for different values (a, u_0, u_1):

$$h_f(t) = u_0 1(0 \leq t < a) + u_1 1(t \geq a),$$
$$h_g(t) = \sum_{i=0}^{3} v_i 1(b_i < t \leq b_{i+1}),$$

with $(b_0, b_1, b_2, b_3, b_4) = (0, 0.1, 0.3, 1, \infty)$ and $(v_0, v_1, v_2, v_3) = (8, 10, 15, 20)$, which implies that mean service rate $\mu \simeq 8.949$. We consider the perturbed hazard rate $\overline{h}_g$ given by

$$\overline{h}_g(t) = \sum_{i=0}^{3} (v_i + k\epsilon) 1(b_i < t \leq b_{i+1}),$$

for $\epsilon = 0.1$ and for each $k = 1, 2, 3, 4, 5$. Four different arrival distributions are considered for each of the above models. They are:

(i) $(a, u_0, u_1) = (0.3, 2, 3)$: $\lambda \simeq 2.448$,

(ii) $(a, u_0, u_1) = (0.3, 3, 4)$: $\lambda \simeq 3.339$,

(iii) $(a, u_0, u_1) = (0.3, 4, 5)$: $\lambda \simeq 4.256$,

(iv) $(a, u_0, u_1) = (0.3, 6, 4)$: $\lambda \simeq 5.542$.

For example, in the case of (i), we let $H_f = 3$, $H_g = 20 + 0.1k$, and $Q = H_f + H_g$. the mean service rates $\overline{\mu}$ for $\overline{h}_g$ and the corresponding values $\overline{h}_Q$ are for each $k = 1, 2, 3, 4, 5$:

$$\overline{\mu} \simeq 9.038, 9.126, 9.215, 9.304, 9.393,$$
$$\overline{h}_Q \simeq 8.303, 8.401, 8.499, 8.597, 8.695.$$

The cases of (i)-(iv) cover various traffic intensities ρ ranging from approximately 0.26 to 0.74. Furthermore, we consider two buffer sizes $N = 5$ and 10, and restrict our attention to the mean queue length L.

These models are simulated, which gives exact errors. The corresponding error bounds are computed by (3.21). The simulation program is coded by language C, and executed on a personal computer. Each simulation is executed up to 9×10^7 events, and the accuracy of simulation is about four digits. The results are shown in Table 6.1.

Table 6.1: Numerical results for the $GI/G/1/N$ queue, Analytic error bound as by (3.21) and exact error in parenthesis for the mean queue length L.

	N	L	$\delta_0 = 0.1$	0.2	0.3	0.4	0.5
(i)	5	0.344	0.034 (0.010)	0.066 (0.024)	0.098 (0.037)	0.128 (0.048)	0.158 (0.064)
	10	0.346	0.034 (0.012)	0.066 (0.025)	0.098 (0.037)	0.128 (0.048)	0.157 (0.060)
(ii)	5	0.538	0.040 (0.016)	0.077 (0.026)	0.114 (0.039)	0.148 (0.050)	0.181 (0.068)
	10	0.542	0.039 (0.015)	0.077 (0.028)	0.113 (0.038)	0.148 (0.052)	0.181 (0.065)
(iii)	5	0.778	0.049 (0.009)	0.095 (0.024)	0.139 (0.031)	0.180 (0.049)	0.219 (0.061)
	10	0.822	0.048 (0.018)	0.094 (0.038)	0.137 (0.043)	0.178 (0.046)	0.216 (0.071)
(iv)	5	1.283	0.065 (0.023)	0.125 (0.036)	0.179 (0.051)	0.230 (0.068)	0.277 (0.077)
	10	1.591	0.062 (0.019)	0.118 (0.058)	0.170 (0.068)	0.218 (0.088)	0.263 (0.101)

The following observations are made as based on the particular examples.

- Overall the analytic error bounds as according to (3.21) are in the order of no more than a factor of 3 to 4 times the exact numerical error.
- When $\delta_0 = \max_{t>0} |h_g(t) - \overline{h}_g(t)|$ increases, the exact errors grow linearly in the same linear order as expressed by (3.21).

Clearly, these observations are rather specific for the performance measure used, and the type of hazard functions. For example, with $(a, u_0, u_1) = (0.3, 7, 5)$ and hence a traffic intensity $\rho \simeq 0.74$, $N = 5$ and $\delta_0 = 0.10$, we find the probability of an empty system $\pi(0) = 0.305$ with exact error 0.006 (2%) and analytic error 0.081 (27%). However, even in this case the analytic error bound can still be practically useful as a secure sensitivity bound for a measure $\pi(0)$ that couldn't be obtained easily.

7. Concluding remarks

This concluding section contains an evaluation of the results, a brief discussion on the differences with stochastic comparison, and a brief discussion on limitations and possible extensions.

Evaluation of results In this paper, the Markov reward approach is applied for non-exponential service systems. Error bounds are derived for single and multi-server queues when the arrival or service distribution is perturbed. As side result, ordering results could be concluded also when systems are not ordered.

The results are essentially based on bounding so-called bias terms for specific performance measures of interest such as a mean queue length or steady state tail probability. The technical details of this step are rather complex and rely upon discrete-time induction. Once this step is established, the error bounds and comparison results can be concluded rather easily for different applications.

Both analytic and numerical results seem to support the results for practical purposes: To provide secure bounds and orders of magnitude.

Stochastic comparison The primary purpose of the Markov reward approach is to obtain error bound rather than just ordering results. This is the major advantage over the stochastic comparisons approach. As side results, it may also lead to ordering results.

As disadvantage though, the proofs by the Markov reward approach are generally less elegant and depend on the specific performance measures of interest. Most importantly, sample path arguments do not generally require exponentiality assumptions unless interchangeability arguments are used such as for finite queues. It has been this major disadvantage of exponentiality that the current paper has dealt with under the limitation of bounded hazard rates.

Limitations and extensions The error bounds will have some practical limitations. For one thing, either one of distributions Π or $\overline{\Pi}$ is needed to evaluate the error bounds (see (b) of Remark 2.1). Fortunately, in most cases, we only need characteristics determined by marginal distributions or moments. If these characteristics are unknown with respect to both of Π and $\overline{\Pi}$, one way to overcome this limitation is to use upper or lower bounds for the characteristics required. Clearly, this will degrade the quality of the error bounds.

Another limitation is the assumption that the hazard rates are bounded, which may be too restrictive in specific situations. This assumption might be relaxed in two ways: by an extended uniformization technique as developed in [20] or by a general framework such as GSMP (Generalized Semi-Markov Process). Both extensions, however, will still require a number of technical steps to be resolved. This seems to be a challenging problem.

Another challenging problem for future research is to extend the current approach to networks of nonexponential service systems. For the exponential service cases, the Markov reward approach has been successfully applied in a number of network situations (e.g., see [21, 24, 25]). But for perturbing the service distributions a more extended framework will be required in line with the present paper.

Acknowledgment

We are grateful to the anonymous referee for his careful reading and valuable comments. We appreciate the international exchange program of the Japan Society for the Promotion of Science (JSPS) and the Netherlands Organization for Scientific Research (N.W.O.), for the support by which this research could be realized.

References

[1] Baccelli, F., and Bremaud, P. (1994) *Elements of queueing theory.* Springer, Berlin/ Heidelberg.

[2] Daley, D.J. and Moran, P.A.P.(1968) Two-sided inequalities for waiting time and queue size in GI G 1. Theory Prob. Appl. 13, 356-358.

[3] D. Gross and C. M. Harris (1985) *Fundamentals of Queueing Theory.* 2nd Ed., John Wiley & Sons, New York.

[4] Jacobs, D.R. and Schach. S.(1972) Stochastic order relationships between $GI/G/k$ systems. Ann. Math. Statist.43, 1623-1632.

[5] Keilson, J. and Kester, A. (1977) Monotone matrices and monotone Markov processes, Stochastic Processes and Their Applications 5, 231-241.

[6] Kimura, T. (1994) Approximations for multi-server queues: system interpolations. Queueing Systems 17, 347-382.

[7] Massey, W. A. (1987) Strong orderings for Markov processes on partially ordered spaces, Math of Operations Research 12, 350-367.

[8] M. Miyazawa (1987) A generalized Pollaczek-Khinchine formula for the $GI/GI/1/k$ queue and its application to approximation. Stochastic Models, Vol. 3, 53-65.

[9] M. Miyazawa (1994) Rate conservation laws: a survey. Queueing Systems 15, 1-58.

[10] Miyazawa, M. and Van Dijk, N. M. (1997) A note on bounds and error bounds for non-exponential batch arrival systems. Probability in the Engineering and Informational Sciences 11, 189-201.

[11] Shaked, M. and Shanthikumar J. G. (1994) Stochastic Orders and Their Applications, Academic Press, San Diego.

[12] Sonderman, D.(1978) Comparison Results for Stochastic Processes Arising in Queueing Systems. Ph.D. Dissertation. Yale University.

[13] Sonderman, D. (1979a) Comparing multi-server queues with finite waiting rooms, I: Same number of servers. Adv. Appl. Prob. 11, 439-447.

[14] Sonderman, D. (1979b) Comparing multi-server queues with finite waiting rooms, II: Different number of servers. Adv. Appl. Prob. 11, 448-455.

[15] Stidham, S. (1970) On the Optimality of single-server queueing systems. Opns. Res. 18, 708-732.

[16] Stoyan, D. (1977) Bounds and approximations in queueing through monotonicity and continuity. Opns Res.25, 851-863.

[17] Stoyan, D. (1983) Comparison Method for Queues and Other Stochastic Models. Wiley, New York.

[18] Tijms, H. C. (1986) Stochastic Modeling and Analysis, a computational approach. Wiley, Chichester.

[19] Van Dijk, N.M. (1989) A note on extended uniformization for non-exponential stochastic networks. J. Appl. Prob. 28, 955-961.

[20] Van Dijk, N.M. (1991) The importance of bias-terms for error bounds and comparison results, in Numerical Solutions of Markov Chains (ed. W.J. Stewart), Marcel Dekker.

[21] Van Dijk, N.M. (1998) Bounds and error bounds for queueing networks, Annals of Operations Research 79, 295-319.

[22] Van Dijk, N.M. and Korezlioglu, H (1992): On Product Form Approximations for Communication Networks with Losses: Error bounds. Annals Operations Research 35, 60-94.

[23] Van Dijk, N.M. and Puterman, N.L. (1988) Perturbation theory for Markov reward processes with applications to queueing systems, Adv. Appl. Prob. 20, 79-89.

[24] Van Dijk, N.M. and Taylor, P. G. (1998) Strong stochastic bounds for the stationary distribution of a class of multicomponent performability models, Operations Research 46, 665-674.

[25] Van Dijk, N.M. and Van de Wal, J. (1989) Simple bounds and monotonicity results for multi-server exponential tandem queues, Queueing Systems 4, 1-16.

[26] Whitt, W. (1978) Approximations of Dynamic Programs I, Math. Operations Res. 3, 231-243.

[27] Wolff, R. W. (1989) Stochastic Modeling and the Theory of Queues. Prentice-Hall, New Jersey.

[28] Yu, O.S. (1974) Stochastic bounds for heterogeneous-server queues with Erlang service-times. J. Appl. Prob. 11, 785-796.

Appendix A

We prove (3.13) for $h(t) = h_1 + (h_2 - h_1)1(t \geq a)$ with constants $h_2 > h_1 > 0$ and $a > 0$. Since

$$1 - G(t) = \begin{cases} e^{-h_1 t}, & 0 \leq t < a, \\ e^{(h_2-h_1)a-h_2 t}, & a \leq t, \end{cases}$$

it is easy to see that

$$h_Q = h_1 + (h_2 - h_1)e^{-Qa}, \qquad \mu = \frac{h_1 h_2}{h_2 - (h_2 - h_1)e^{-h_1 a}}.$$

Hence, (3.13) is equivalent to

$$\left(1 + \left(\frac{h_2}{h_1} - 1\right) e^{-Qa}\right) \left(1 + \left(\frac{h_1}{h_2} - 1\right) e^{-h_1 a}\right) \leq 1\,.$$

This is further equivalent to

$$\eta(a) \equiv \left(\frac{h_2}{h_1} + \frac{h_1}{h_2} - 2\right) e^{-Qa} - \left(\frac{h_2}{h_1} - 1\right) e^{-(Q-h_1)a} - \frac{h_1}{h_2} + 1 \geq 0, \qquad a \geq 0\,,$$

for each h_1, h_2 and Q such that $0 < h_1 < h_2 < Q$. Clearly $f(0) = 0$, and

$$\begin{aligned} \eta'(a) &= \left(\frac{h_2 - h_1}{h_1}(Q - h_1)e^{h_1 a} - \frac{(h_2 - h_1)^2}{h_1 h_2} Q\right) e^{-Qa} \\ &= \frac{h_2 - h_1}{h_1} \left((Q - h_1)e^{h_1 a} - \frac{(h_2 - h_1)}{h_2} Q\right) e^{-Qa} > 0, \quad a \geq 0, \end{aligned}$$

since $h_2(Q - h_1) > (h_2 - h_1)Q$. Hence, $f(a) \geq 0$ for all $a \geq 0$. Thus we get (3.13).

Appendix B (Quality of (3.21))

Consider quality of the error bound (3.21) for the $M/G/1/\infty$ queue. In this case, the mean queue length L is well known as the Pollaczek-Khinchine formula. Let $H(t) = \int_0^t h(v)dv$. Then, under the perturbation that $\overline{h}(t) = h(t) + \delta_0$, the asymptotic ratio is computed as

$$\begin{aligned} \lim_{\delta_0 \downarrow 0} \frac{\overline{L} - L}{\delta_0} &= \frac{\partial}{\partial \delta_0} \left(\frac{\lambda^2 \int_0^\infty t e^{-H(t)-\delta_0 t}}{1 - \lambda \int_0^\infty e^{-H(t)-\delta_0 t} dt} + \lambda \int_0^\infty e^{-H(t)-\delta_0 t} dt \right) \Bigg|_{\delta_0 = 0} \\ &= -\frac{\lambda^2 E(S^3)(1-\rho)/3 + \lambda^3 E^2(S^2)/4}{(1-\rho)^2} - \lambda E(S^2)/2, \end{aligned}$$

where S is a random variable which represents the service time. In general, it is too complicated to compare this with the error bound in (3.21). So, we just consider the case that the service distribution is the 2nd order Erlang distribution, i.e., its density g and hazard rate h are given by

$$g(t) = (2\mu)^2 t e^{-2\mu t}, \qquad h(t) = \frac{(2\mu)^2 t}{1 + 2\mu t}.$$

In this case, $E(S^2) = 3/2\mu^2$, $E(S^3) = 3/\mu^3$ and

$$L = \frac{\rho(1 - \frac{1}{4}\rho)}{1 - \rho}.$$

Using the notation $\rho = \lambda/\mu$, the asymptotic ratio is thus computed as

$$\begin{aligned} \lim_{\delta_0 \downarrow 0} \frac{1}{\delta_0} \frac{|\overline{L} - L|}{L} &= \frac{\rho}{\mu(1-\rho)^2} \left(\frac{3}{4} + \frac{1}{4}\rho - \frac{7}{16}\rho^2 \right) \frac{1 - \rho}{\rho(1 - \frac{1}{4}\rho)} \\ &= \frac{1}{\mu - \lambda} \frac{3 + \rho - \frac{7}{4}\rho^2}{4 - \rho}. \end{aligned}$$

On the other hand, choosing $Q = (\rho + 2)\mu$,

$$\begin{aligned} h_Q &= \int_0^\infty \frac{(2\mu)^2 t}{1 + 2\mu t} (\rho + 2)\mu e^{-(\rho+2)\mu t} dt \\ &= 4(\rho + 2)\mu \int_0^\infty \frac{x}{1 + 2x} e^{-(\rho+2)x} dx. \end{aligned}$$

Since h_Q is decreasing in ρ, we compute the maximum ρ such that $h_Q - \rho > 0$, which is about $\rho_0 \simeq 0.694527$ (computed by Mathematica 4.0). Then, we have

$$h_Q > \rho_0 \mu, \qquad \text{for} \qquad \rho < \rho_0$$

where the upper lower bound is attained as ρ goes to ρ_0, while h_Q goes up to about 0.807305 as ρ goes down to 0. Hence, the asymptotic ratio for the error bound in (3.21) is bounded by

$$\begin{aligned} \lim_{\delta_0 \downarrow 0} \overline{C} \left(1 + \frac{1 - \pi(0)}{L} \right) &< \lim_{\delta_0 \downarrow 0} \frac{1}{\rho_0 \mu + \delta_0 - \lambda} \left(1 + \frac{4 - 4\rho}{4 - \rho} \right) \\ &= \frac{1}{\rho_0 \mu - \lambda} \left(1 + \frac{4 - 4\rho}{4 - \rho} \right). \end{aligned}$$

Though this error bound is not as good as in the $M/M/1/\infty$ case, it still remains a constant for $\rho < \rho_0$.

Appendix C (Proof of Lemma 4.1)

Before proving Lemma 4.1, we note the following fact, which will be used in an induction procedure.

Lemma C.1 If (4.2)-(4.4) hold for all possible $\boldsymbol{s}$, w, s and $c = n$ for each fixed k, then

$$
\begin{aligned}
{}^1\boldsymbol{\Delta}_i^k(\boldsymbol{s},0) - {}^2\boldsymbol{\Delta}_i^k(\boldsymbol{s},w)[t] &\geq 0, && \text{(C.1)}\\
{}^3\boldsymbol{\Delta}_i^k(\boldsymbol{s},w)[t] - {}^2\boldsymbol{\Delta}_i^k(\boldsymbol{s},w)[t] &\geq 0. && \text{(C.2)}
\end{aligned}
$$

Proof. From (4.2)-(4.4), we have

$$
\begin{aligned}
&{}^1\boldsymbol{\Delta}_i^k(\boldsymbol{s},0) - {}^2\boldsymbol{\Delta}_i^k(\boldsymbol{s},w)[t] = V_k(\boldsymbol{s} \ominus \boldsymbol{u}_i \oplus \boldsymbol{u}_i(t), w) - V_k(\boldsymbol{s} \ominus \boldsymbol{u}_i, 0)\\
&\quad = V_k(\boldsymbol{s} \ominus \boldsymbol{u}_i \oplus \boldsymbol{u}_i(t), w) - V_k(\boldsymbol{s} \ominus \boldsymbol{u}_i \oplus \boldsymbol{u}_i(0), w-1)\\
&\qquad\quad + V_k(\boldsymbol{s} \ominus \boldsymbol{u}_i \oplus \boldsymbol{u}_i(0), w-1) - V_k(\boldsymbol{s} \ominus \boldsymbol{u}_i, 0)\\
&\quad \ldots\\
&\quad = {}^3\boldsymbol{\Delta}_i^k(\boldsymbol{s} \ominus \boldsymbol{u}_i \oplus \boldsymbol{u}_i(t), w)[0]\\
&\qquad\quad + \sum_{\ell=1}^{w-1} {}^3\boldsymbol{\Delta}_i^k(\boldsymbol{s} \ominus \boldsymbol{u}_i \oplus \boldsymbol{u}_i(0), \ell)[0] + {}^1\boldsymbol{\Delta}_i^k(\boldsymbol{s} \ominus \boldsymbol{u}_i \oplus \boldsymbol{u}_i(0), 0) \geq 0\,,
\end{aligned}
$$

and, similarly,

$$
{}^3\boldsymbol{\Delta}_i^k(\boldsymbol{s},w)[t] - {}^2\boldsymbol{\Delta}_i^k(\boldsymbol{s},w)[t] = {}^3\boldsymbol{\Delta}_i^k(\boldsymbol{s} \ominus \boldsymbol{u}_i \oplus \boldsymbol{u}_i(t), w)[t] \geq 0\,. \qquad \square
$$

(*Proof of Lemma 4.1* The lemma will be proved by induction on k. Clearly, (4.2)-(4.4) hold for $k = 0$ as $V_0(\cdot) \equiv 0$. Assume that they holds for $k = m$. Then we need to verify (4.2)-(4.4) for $k = m + 1$. We will do so for (4.2), (4.3) and (4.4) separately, under i), ii) and iii) below. Herein, we let $\alpha = 1/Q$ and we note in advance that this α is kept in the expression below, while it could be canceled with Q, both for its probabilistic interpretation and a corresponding usage later on.

i) By comparing (2.3) in states $(n, a, \boldsymbol{s})$ and $(n, a, \boldsymbol{s} \ominus \boldsymbol{u}_i)$ with $n \leq c$, which correspond with $(\boldsymbol{s}, 0)$ and $(\boldsymbol{s} \ominus \boldsymbol{u}_i, 0)$ in the notation of Section 4, while substituting P of Section 2.1, we find

$$
\begin{aligned}
V_{m+1}(\boldsymbol{s},0) = \int_0^\infty dv Q e^{-vQ} \Big\{ &\alpha \sum_j h(s_j + v) + \alpha\lambda 1(n < c) V_m(\boldsymbol{s}_v \oplus \boldsymbol{u}_{n+1}(0), 0)\\
&+ \alpha\lambda 1(n = c) V_m(\boldsymbol{s}_v, 1)\\
&+ \alpha \sum_j h(s_j + v) V_m(\boldsymbol{s}_v \ominus \boldsymbol{u}_j, 0)\\
&+ \Big[1 - \alpha\lambda - \alpha \sum_j h(s_j + v)\Big] V_m(\boldsymbol{s}_v, 0) \Big\} && \text{(C.3)}
\end{aligned}
$$

$$
\begin{aligned}
V_{m+1}(\boldsymbol{s} \ominus \boldsymbol{u}_i, 0) = \int_0^\infty dv Q e^{-vQ} \Big\{ &\alpha \sum_{j \neq i} h(s_j + v) + \alpha\lambda V_m(\boldsymbol{s}_v \ominus \boldsymbol{u}_i \oplus \boldsymbol{u}_n(0), 0)\\
&+ \alpha \sum_{j \neq i} h(s_j + v) V_m(\boldsymbol{s}_v \ominus (\boldsymbol{u}_i \vee \boldsymbol{u}_j), 0)\\
&+ \Big[1 - \alpha\lambda - \alpha \sum_{j \neq i} h(s_j + v)\Big] V_m(\boldsymbol{s}_v \ominus \boldsymbol{u}_i, 0) \Big\}\,. && \text{(C.4)}
\end{aligned}
$$

To compare the subtraction of (C.4) from (C.3) in a transition wise manner, the following steps are convenient. In (C.4), write;

$$\begin{aligned}\alpha\lambda V_m(\boldsymbol{s}_v \ominus \boldsymbol{u}_i \oplus \boldsymbol{u}_n(0), w) &= \alpha\lambda 1(n < c) V_m(\boldsymbol{s}_v \ominus \boldsymbol{u}_i \oplus \boldsymbol{u}_n(0), w) \\ &\quad +\alpha\lambda 1(n = c) V_m(\boldsymbol{s}_v \ominus \boldsymbol{u}_i \oplus \boldsymbol{u}_n(0), w)\,.\end{aligned}$$

Artificially add but also subtract the term also in (C.4):

$$\alpha h(s_i + v) V_m(\boldsymbol{s}_v \ominus \boldsymbol{u}_i)\,.$$

Then, after these steps and subtracting (C.4) from (C.3), the following expression is obtained, where the one but last term is indeed equal to 0 but kept in for clarity as well as an argument below.

$$\begin{aligned}V_{m+1}(\boldsymbol{s}, 0) - V_{m+1}(\boldsymbol{s} \ominus \boldsymbol{u}_i, 0) = \int_0^\infty & dv Q e^{-vQ} \Big\{ \alpha h(s_i + v) \\ & +\alpha\lambda 1(n < c) \Big[V_m(\boldsymbol{s}_v \oplus \boldsymbol{u}_{n+1}(0), 0) - V_m(\boldsymbol{s}_v \ominus \boldsymbol{u}_i \oplus \boldsymbol{u}_n(0), 0) \Big] \\ & +\alpha\lambda 1(n = c) \Big[V_m(\boldsymbol{s}_v, 1) - V_m(\boldsymbol{s}_v \ominus \boldsymbol{u}_i \oplus \boldsymbol{u}_n(0), 0) \Big] \\ & +\alpha \sum_{j \neq i} h(s_j + v) \Big[V_m(\boldsymbol{s}_v \ominus \boldsymbol{u}_j, 0) - V_m(\boldsymbol{s}_v \ominus (\boldsymbol{u}_i \vee \boldsymbol{u}_j), 0) \Big] \\ & +\alpha h(s_i + v) \Big[V_m(\boldsymbol{s}_v \ominus \boldsymbol{u}_i, 0) - V_m(\boldsymbol{s}_v \ominus \boldsymbol{u}_i, 0) \Big] \\ +\Big[1 - \alpha\lambda - \alpha \sum_{j \neq i} h(s_j + v) - \alpha h(s_i + v) \Big] & \Big[V_m(\boldsymbol{s}_v, 0) - V_m(\boldsymbol{s}_v \ominus \boldsymbol{u}_i, 0) \Big] \Big\} \qquad \text{(C.5)}\end{aligned}$$

Now note that for the term with coefficient $\alpha\lambda 1(n = c)$, we could write:

$$V_m(\boldsymbol{s}_v, 1) - V_m(\boldsymbol{s}_v \ominus \boldsymbol{u}_i \oplus \boldsymbol{u}_n(0), 0) = V_m(\boldsymbol{s}_v, 1) - V_m(\boldsymbol{s}_v \ominus \boldsymbol{u}_i \oplus \boldsymbol{u}_i(0), 0)\,,$$

as the actual stochastic behavior and expected rewards do not depend on the actual positioning of the jobs but only their attained times. By recalling the notation as per (4.2) and (4.4), we have so obtained:

$$\begin{aligned}{}^1\boldsymbol{\Delta}_i^{m+1}(\boldsymbol{s}, 0) = \int_0^\infty & dv Q e^{-vQ} \Big\{ \alpha h(s_i + v) \\ & +\alpha\lambda 1(n < c)\, {}^1\boldsymbol{\Delta}_i^m(\boldsymbol{s}_v \oplus \boldsymbol{u}_{n+1}(0), 0) \\ & +\alpha\lambda 1(n = c)\, {}^3\boldsymbol{\Delta}_i^m(\boldsymbol{s}_v, 0)[0] \\ & +\alpha \sum_{j \neq i} h(s_j + v)\, {}^1\boldsymbol{\Delta}_i^m(\boldsymbol{s}_v \ominus \boldsymbol{u}_j, 0) \\ & +\alpha h(s_i + v) \times 0 \\ & +\Big[1 - \alpha\lambda - \alpha \sum_j h(s_j + v) \Big]\, {}^1\boldsymbol{\Delta}_i^m(\boldsymbol{s}_v, 0) \Big\} \qquad \text{(C.6)}\end{aligned}$$

Now we can directly substitute the induction hypothesis (4.2) and (4.4) for $k = m$ to conclude the lower estimate ${}^1\boldsymbol{\Delta}_i^{m+1}(\boldsymbol{s}_v, 0) \geq 0$. Similarly, to conclude the upper estimate

1 in (4.2) for $k = m+1$, substitute the upper estimates from (4.2) and (4.4) for $k = m$ and note that all coefficients represent probabilities and thus sum up to 1. Furthermore, the first additional term $\alpha h(s_i + v)$ is compensated by the 0-term with coefficient $\alpha h(s_i + v)$. Accordingly, we concludes: ${}^1\boldsymbol{\Delta}_i^{m+1}(\boldsymbol{s}_v, 0) \leq 1$.

ii) By similar steps, by writing out (2.3) in states $(\boldsymbol{s}, w)$ and $\boldsymbol{s} \ominus \boldsymbol{u}_i \oplus \boldsymbol{u}_i(t), w)$, we have

$$\begin{aligned}
V_{m+1}(\boldsymbol{s}, w) = \int_0^\infty dvQe^{-vQ}\Big\{&\alpha \sum_j h(s_j + v) \\
&+\alpha\lambda 1(n < c, w = 0)V_m(\boldsymbol{s}_v \oplus \boldsymbol{u}_{n+1}(0), w) \\
&+\alpha\lambda 1(n = c, c + w < N)V_m(\boldsymbol{s}_v, w + 1) \\
&+\alpha\lambda 1(n = c, c + w = N)V_m(\boldsymbol{s}_v, w) \\
&+\alpha \sum_j h(s_j + v)1(w = 0)V_m(\boldsymbol{s}_v \ominus \boldsymbol{u}_j, 0) \\
&+\alpha \sum_j h(s_j + v)1(w > 0)V_m(\boldsymbol{s}_v \ominus \boldsymbol{u}_j \oplus \boldsymbol{u}_j(0), w - 1) \\
&+\Big[1 - \alpha\lambda - \alpha \sum_j h(s_j + v)\Big] V_m(\boldsymbol{s}_v, w)\Big\}. \qquad \text{(C.7)}
\end{aligned}$$

$$\begin{aligned}
V_{m+1}(\boldsymbol{s} \ominus \boldsymbol{u}_i \oplus \boldsymbol{u}_i(t), w) = \int_0^\infty dvQe^{-vQ}\Big\{&\alpha \sum_{j\neq i} h(s_j + v) + \alpha h(t + v) \\
&+\alpha\lambda 1(n < c, w = 0)V_m(\boldsymbol{s}_v \ominus \boldsymbol{u}_i \oplus \boldsymbol{u}_i(t + v) \oplus \boldsymbol{u}_{n+1}(0), w) \\
&+\alpha\lambda 1(n = c, c + w < N)V_m(\boldsymbol{s}_v \ominus \boldsymbol{u}_i \oplus \boldsymbol{u}_i(t + v), w + 1) \\
&+\alpha\lambda 1(n = c, c + w = N)V_m(\boldsymbol{s}_v \ominus \boldsymbol{u}_i \oplus \boldsymbol{u}_i(t + v), w)
\end{aligned}$$

$$\begin{aligned}
&+\alpha \sum_{j\neq i} h(s_j + v)1(w = 0)V_m(\boldsymbol{s}_v \ominus \boldsymbol{u}_i \oplus \boldsymbol{u}_i(t + v) \ominus \boldsymbol{u}_j, 0) \\
&+\alpha h(t + v)1(w = 0)V_m(\boldsymbol{s}_v \ominus \boldsymbol{u}_i, 0) \\
&+\alpha \sum_{j\neq i} h(s_j + v)1(w > 0)V_m(\boldsymbol{s}_v \ominus \boldsymbol{u}_i \oplus \boldsymbol{u}_i(t + v) \ominus \boldsymbol{u}_j \oplus \boldsymbol{u}_j(0), w - 1) \\
&+\alpha h(t + v)1(w > 0)V_m(\boldsymbol{s}_v \ominus \boldsymbol{u}_i \oplus \boldsymbol{u}_i(0), w - 1) \\
&+\Big[1 - \alpha\lambda - \alpha \sum_{j\neq i} h(s_j + v) - \alpha h(t + v)\Big] V_m(\boldsymbol{s}_v \ominus \boldsymbol{u}_i \oplus \boldsymbol{u}_i(t), w)\Big\} \text{ .(C.8)}
\end{aligned}$$

Hence, taking the difference of (C.7) and (C.8), we find the following result after recalling the notation for ${}^2\boldsymbol{\Delta}_i^m$.

$$\begin{aligned}
{}^2\boldsymbol{\Delta}_i^{m+1}(\boldsymbol{s}, w)[t] = \int_0^\infty dvQe^{-vQ}\Big\{&\alpha[h(s_i + v) - h(t + v)] \\
&+\alpha\lambda 1(n < c)\, {}^2\boldsymbol{\Delta}_i^m(\boldsymbol{s}_v \oplus \boldsymbol{u}_{n+1}, 0)[t + v] \\
&+\alpha\lambda 1(n = c, n + w < N)\, {}^2\boldsymbol{\Delta}_i^m(\boldsymbol{s}_v, w + 1)[t + v] \\
&+\alpha\lambda 1(n = c, n + w = N)\, {}^2\boldsymbol{\Delta}_i^m(\boldsymbol{s}_v, w)[t + v] \\
&+\alpha \sum_{j\neq i} h(s_j + v)1(w = 0)\, {}^2\boldsymbol{\Delta}_i^m(\boldsymbol{s}_v \ominus \boldsymbol{u}_j, 0)[t + v]
\end{aligned}$$

$$\begin{aligned}
&+\alpha \sum_{j\neq i} h(s_j+v)1(w>0)\,{}^{2}\boldsymbol{\Delta}_i^m(\boldsymbol{s}_v \ominus \boldsymbol{u}_j \oplus \boldsymbol{u}_j(0), w-1)[t+v] \\
&+\alpha(h(s_i+v)-h(t+v))1(w=0)V_m(\boldsymbol{s}_v\ominus\boldsymbol{u}_i,0) \\
&+\alpha(h(s_i+v)-h(t+v))1(w>0)V_m(\boldsymbol{s}_v\ominus\boldsymbol{u}_i\oplus\boldsymbol{u}_i(0),w-1) \\
&+\Big[1-\alpha\lambda-\alpha\sum_{j} h(s_j+v)\Big]\,{}^{2}\boldsymbol{\Delta}_i^m(\boldsymbol{s}_v,w)[t]\Big\} \\
&-\alpha(h(s_i+v)-h(t+v))V_m(\boldsymbol{s}_v\ominus\boldsymbol{u}_i\oplus\boldsymbol{u}_i(t+v),w)\,. \qquad \text{(C.9)}
\end{aligned}$$

Now first note that the terms with coefficient $\alpha[h(s_i+v)-h(t+v)]$ can be written as:

$$\begin{aligned}
\alpha[h(s_i+v)-h(t+v)]\Big\{&-1(w=0)\,{}^{1}\boldsymbol{\Delta}_i^m(\boldsymbol{s}_v\ominus\boldsymbol{u}_i\oplus\boldsymbol{u}_i(t+v),0) \\
&-1(w>0)\,{}^{3}\boldsymbol{\Delta}_i^m(\boldsymbol{s}_v\ominus\boldsymbol{u}_i\oplus\boldsymbol{u}_i(t+v),w)[0]\Big\}\,,
\end{aligned}$$

which by virtue of the induction hypothesis (4.2) and (4.4) for $k=m$ can thus by estimated from below by $-\alpha[h(s_i+v)-h(t+v)]$ and from above by 0. Consequently, by taking into account the first additional term $-\alpha[h(s_i+v)-h(t+v)]$, by substituting the lower estimates 0 from (4.2) -(4.4) for $k=m$, we conclude: ${}^{2}\boldsymbol{\Delta}_i^{m+1}(\boldsymbol{s},w)[t]\geq 0$. Conversely, by deleting the one but last nonpositive term and noting that its probability coefficient is equal to the additional first term $-\alpha[h(s_i+v)-h(t+v)]$, substituting the upper estimates 1 from (4.3) and recalling thus all coefficients sum up to 1, we also conclude: ${}^{2}\boldsymbol{\Delta}_i^{m+1}(\boldsymbol{s},w)[t]\leq 1$.

iii) For $0<w\leq N$, we get the following expression from (2.3) similarly to (C.8).

$$\begin{aligned}
V_{m+1}(\boldsymbol{s}\ominus\boldsymbol{u}_i\oplus\boldsymbol{u}_i(t),w-1) = \int_0^\infty dvQe^{-vQ}\Big\{&\alpha\sum_{j\neq i}h(s_j+v)+\alpha h(t+v) \\
&+\alpha\lambda 1(n=c,c+w\leq N)V_m(\boldsymbol{s}_v\ominus\boldsymbol{u}_i\oplus\boldsymbol{u}_i(t+v),w) \\
&+\alpha\sum_{j\neq i}h(s_j+v)1(w=1)V_m(\boldsymbol{s}_v\ominus\boldsymbol{u}_i\oplus\boldsymbol{u}_i(t+v)\ominus\boldsymbol{u}_j,0) \\
&+\alpha h(t+v)1(w=1)V_m(\boldsymbol{s}_v\ominus\boldsymbol{u}_i,0) \\
&+\alpha\sum_{j\neq i}h(s_j+v)1(w>1)V_m(\boldsymbol{s}_v\ominus\boldsymbol{u}_i\oplus\boldsymbol{u}_i(t+v)\ominus\boldsymbol{u}_j\oplus\boldsymbol{u}_j(0),w-2) \\
&+\alpha h(t+v)1(w>1)V_m(\boldsymbol{s}_v\ominus\boldsymbol{u}_i\oplus\boldsymbol{u}_i(0),w-2) \\
&+\Big[1-\alpha\lambda-\alpha\sum_{j\neq i}h(s_j+v)-\alpha h(t+v)\Big]V_m(\boldsymbol{s}_v\ominus\boldsymbol{u}_i\oplus\boldsymbol{u}_i(t),w-1)\Big\}. \quad \text{(C.10)}
\end{aligned}$$

Taking the difference of (C.7) and (C.10) and appropriately arranging some terms similar to (C.9), we obtain, for $0 < w \leq N$ and $c = n$,

$$\begin{aligned}
{}^{3}\boldsymbol{\Delta}_i^{m+1}(\boldsymbol{s}, w)[t] = \int_0^\infty & dvQe^{-vQ}\Big\{\alpha[h(s_i+v) - h(t+v)] \\
& +\alpha\lambda 1(n+w<N)\,{}^{3}\boldsymbol{\Delta}_i^{m}(\boldsymbol{s}_v, w+1)[t+v] \\
& +\alpha\lambda 1(n+w=N){}^{2}\boldsymbol{\Delta}_i^{m}(\boldsymbol{s}_v, w+1)[t+v] \\
& +\alpha\sum_{j\neq i} h(s_j+v)1(w=1)\Big[V_m(\boldsymbol{s}_v \ominus \boldsymbol{u}_j \oplus \boldsymbol{u}_j(0), 0)) \\
& \qquad\qquad\qquad -V_m(\boldsymbol{s}_v \ominus \boldsymbol{u}_i \oplus \boldsymbol{u}_i(t+v) \ominus \boldsymbol{u}_j, 0)\Big] \\
& +\alpha\sum_{j\neq i} h(s_j+v)1(w>1)\,{}^{3}\boldsymbol{\Delta}_i^{m}(\boldsymbol{s}_v \ominus \boldsymbol{u}_j \oplus \boldsymbol{u}_j(0), w-1)[t+v] \\
& +\alpha h(s_i+v)1(w=1)V_m(\boldsymbol{s}_v \ominus \boldsymbol{u}_i \oplus \boldsymbol{u}_i(0), 0) \\
& -\alpha h(t+v)1(w=1)V_m(\boldsymbol{s}_v \ominus \boldsymbol{u}_i, 0) \\
& +\alpha h(s_i+v)1(w>1)V_m(\boldsymbol{s}_v \ominus \boldsymbol{u}_i \oplus \boldsymbol{u}_i(0), w-1) \\
& -\alpha h(t+v)1(w>1)V_m(\boldsymbol{s}_v \ominus \boldsymbol{u}_i \oplus \boldsymbol{u}_i(0), w-2) \\
& +\Big[1-\alpha\lambda-\alpha\sum_j h(s_j+v)\Big]\,{}^{3}\boldsymbol{\Delta}_i^{m}(\boldsymbol{s}_v, w)[t+v]\Big\} \\
& +\alpha h(t+v)V_m(\boldsymbol{s}_v \ominus \boldsymbol{u}_i \oplus \boldsymbol{u}_i(t+v), w-1) \\
& -\alpha h(s_i+v)V_m(\boldsymbol{s}_v \ominus \boldsymbol{u}_i \oplus \boldsymbol{u}_i(t+v), w-1)\,. \qquad \text{(C.11)}
\end{aligned}$$

First, collecting the terms with coefficient $\alpha h(s_i+v)$ in (C.11), we have

$$\begin{aligned}
-\alpha h(s_i+v) & \left(V_m(\boldsymbol{s}_v \ominus \boldsymbol{u}_i \oplus \boldsymbol{u}_i(0), w-1) - V_m(\boldsymbol{s}_v \ominus \boldsymbol{u}_i \oplus \boldsymbol{u}_i(t+v), w-1)\right) \\
&= -\alpha h(s_i+v)\,{}^{2}\boldsymbol{\Delta}_i^{m}(\boldsymbol{s}_v \ominus \boldsymbol{u}_i \oplus \boldsymbol{u}_i(t+v), w-1)[0]\ ,
\end{aligned}$$

and, similarly, the terms with coefficient $\alpha h(s_i+v)$ becomes

$$\begin{aligned}
+\alpha h(t+v)\Big(& 1(w=1)\,{}^{1}\boldsymbol{\Delta}_i^{m}(\boldsymbol{s}_v \ominus \boldsymbol{u}_i \oplus \boldsymbol{u}_i(t+v), 0) \\
& +1(w>1)\,{}^{3}\boldsymbol{\Delta}_i^{m}(\boldsymbol{s}_v \ominus \boldsymbol{u}_i \oplus \boldsymbol{u}_i(t+v), w-1)[0]\Big)\ .
\end{aligned}$$

Hence, from (C.1)) and (C.2) of Lemma C.1, the sum of these two collected terms is bounded from below by

$$\begin{aligned}
-\alpha(h(s_i+v) - h(t+v))\,{}^{2}\boldsymbol{\Delta}_i^{m}(\boldsymbol{s}_v \ominus \boldsymbol{u}_i \oplus \boldsymbol{u}_i(t+v), w-1)[0] \\
\geq -\alpha(h(s_i+v) - h(t+v))\,, \qquad \text{(C.12)}
\end{aligned}$$

where the last inequality follows from the induction hypothesis (4.3) for $k = m$. On the other hand, using (4.2) and (4.4), the same sum is bounded from above by

$$\alpha h(t+v)\,.$$

Secondly, for the term with $i < j$ and coefficient $\alpha h(s_j + v)1(w = 1)$ with $\boldsymbol{y}_v = \boldsymbol{s}_v \ominus (\boldsymbol{u}_i \vee \boldsymbol{u}_j)$, for $s \leq s_i$ we can write:

$$\begin{aligned}
&V_m(\boldsymbol{s}_v \ominus \boldsymbol{u}_j \oplus \boldsymbol{u}_j(0), 0)) - V_m(\boldsymbol{s}_v \ominus \boldsymbol{u}_i \oplus \boldsymbol{u}_i(t+v) \ominus \boldsymbol{u}_j, 0) \\
&\qquad = V_m(\boldsymbol{y}_v \oplus \boldsymbol{u}_i(s_i + v) \oplus \boldsymbol{u}_j(0), 0)) - V_m(\boldsymbol{y}_v \oplus \boldsymbol{u}_i(t+v), 0)\,. \qquad \text{(C.13)}
\end{aligned}$$

Clearly, by applying the induction hypotheses (4.2) and (4.3) for $k = m$, this difference is directly estimated from below by 0. To estimate this difference from above, now note that

$$V_m(\boldsymbol{y}_v \oplus \boldsymbol{u}_i(0) \oplus \boldsymbol{u}_j(s_i + v), 0)\,,$$

since the actual position of the jobs does not influence the stochastic behavior and thus the expected number of completions. Hence, we can also rewrite this difference from (C.13) as:

$$\begin{aligned}
&V_m(\boldsymbol{y}_v \oplus \boldsymbol{u}_i(s_i + v) \oplus \boldsymbol{u}_j(0), 0) - V_m(\boldsymbol{y}_v \oplus \boldsymbol{u}_i(t+v), 0) \\
&\qquad = V_m(\boldsymbol{y}_v \oplus \boldsymbol{u}_i(s_i + v) \oplus \boldsymbol{u}_j(0), 0) - V_m(\boldsymbol{y}_v \oplus \boldsymbol{u}_i(0), 0) \\
&\qquad\quad + V_m(\boldsymbol{y}_v \oplus \boldsymbol{u}_i(0), 0) - V_m(\boldsymbol{y}_v \oplus \boldsymbol{u}_i(t+v), 0) \\
&\qquad = {}^1\boldsymbol{\Delta}_i^m(\boldsymbol{y}_v \oplus \boldsymbol{u}_i(s_i + v) \oplus \boldsymbol{u}_j(0), 0) - {}^2\boldsymbol{\Delta}_i^m(\boldsymbol{y}_v \oplus \boldsymbol{u}_i(t+v), 0)[0]\,.
\end{aligned}$$

By substituting the induction hypotheses (4.2) and (4.3) for $k = m$, the latter difference is bounded from above by 1 and thus also the difference in (C.4) by 0 and 1.

Finally, by combining (C.11)-(C.13) and collecting these arguments, by using the induction hypotheses (4.2) -(4.4) for $k = m$ in relation (C.11), by similar arguments as before, we have also shown (4.4) for $k = m + 1$. Induction now completes the proof of Lemma 4.1. □