A discrete-time queue with customers with geometric deadlines

Herwig Bruneel and Tom Maertens

SMACS Research Group
Department of Telecommunications and Information Processing
Ghent University - UGent
E-mail: {hb,tmaerten}@telin.UGent.be

Abstract

This paper studies a discrete-time queueing system where each customer has a maximum allowed sojourn time in the system, referred to as the "deadline" of the customer. More specifically, we model the deadlines of the consecutive customers as independent and geometrically distributed random variables. Customers enter the system according to a *general* independent arrival process, i.e., the numbers of arrivals during consecutive time slots are i.i.d. random variables with arbitrary distribution. Service times of the customers are deterministically equal to one slot each. For this queueing model, we are able to obtain exact formulas for such quantities as the generating function and the expected value of the system content, the mean customer delay and the deadline-expiration ratio. These formulas, however, contain infinite sums and infinite products, which implies that truncations are required to actually compute numerical values. Therefore, we also derive some easy-to-evaluate approximate results for the main performance measures, based on a *polynomial approximation* technique. We believe this technique, in its own right, is also one of the major (methodological) contributions of the paper.

Possible applications of this type of queueing model are numerous: the (variable) deadlines could model, for instance, the fact that customers may become impatient and leave the queue unserved if they have to wait too long in line, but they could also reflect the fact that the service of a customer is not useful anymore if it cannot be delivered soon enough, etc.

Key words: queueing; discrete-time; deadlines; closed-form results; polynomial approximation

1 Introduction

In a typical queueing model, customers present themselves near some service facility to receive some kind of service, and – if they cannot be served immediately upon arrival – wait patiently in a queue until the server is available for them. In some cases, however, customers may leave (or *abandon*) the queue unserved if their time in the queue becomes too big. Although sometimes referred to as "queues with abandonments" or "queues with reneging" in the literature, this type of queues is usually known as "queues with customer impatience".

The motivations for studying queues with customer impatience are legion. Early papers on the topic were written in the context of telephone traffic and call centers. One of the pioneering works is that of Palm [31]. He considered an unlimited M/M/n queue and assumed that each individual customer stays in the queue as long as his waiting time does not exceed an exponentially distributed impatience time (i.e., M/M/n+M, where the "+M" specifies the impatience law). This is the so-called *Erlang-A* model; it is the simplest model including abandonments. Amongst other results, he represented the steady-state distribution of the number of customers in the Erlang-A system, and some of its important performance measures, in terms of incomplete Gamma functions and the blocking probability in the Erlang-B (i.e., M/M/n/n) system. Several authors extended his results, in various directions and sometimes independently of each other. Readers are referred to the invited review paper of Gans et al. [16] and to the PhD thesis of Zeltyn [39] for extensive literature reviews (up to 2003-2004) and relevant results on call center research in general, including models of customers' impatience, and specifically on M/M/n queues with exponentially and generally distributed patient times, respectively. More recent references on customer impatience research in the context of call centers can be found in [22]. In [22], moreover, the authors propose an extension of the Erlang A model in which the possibility of balking (refusing to join the queue) is included. This simple extension makes the performance prediction by the queueing model much more accurate. Furthermore, they study a number of different service level definitions, including all those used in practice, and show how to explicitly compute their performance measures by using existing results on the *virtual waiting time*, i.e., the waiting time that a customer with infinite patience would experience.

Customer impatience (or, more general, abandonments), however, is also not to be ignored in, for example, real-time telecommunication applications (see, e.g., [7, 18, 26, 36]), inventory management (see, e.g., [2, 9, 19]), emergency situations, staffing decisions (see, e.g., [1, 12, 41]), parking policy (see, e.g., [27]), etc. Usually, it is the customer that takes the decision to abandon prematurely, e.g., because the customer (usually, a human being in this case) does not like to or cannot wait any longer. A call center is the most obvious example of this type of abandonments. On the other hand, also the system itself may decide to remove customers from the queue, e.g., if servicing those customers is deemed not to be useful any more after some time in the queue or if the customers are "expired" (e.g., perishable goods). The first situation may appear in audio or video streaming applications (see, e.g., [14, 28, 30]). In particular, when packets belonging to such applications would not arrive soon enough at their next destination if they have to wait any longer, they are removed from the buffer (see, e.g., [18, 29, 33]). The second situation can be of great importance in inventory management (see, e.g., [9, 19]).

There are many examples of perishable products such as food items, chemicals, pharmaceutical products, blood, etc. Understanding such systems and investigating the impact of the finiteness of product lifetimes on production and inventory control decisions is thus clearly necessary in a society in which waste is less and less accepted and in which extra costs (e.g., for cleaning up waste) are more and more avoided. For other examples and situations in which abandonments play an important role, we refer to [21] and [38].

There is clearly no shortage of *continuous-time* models to study queues with customer impatience (see, e.g., Zeltyn and Mandelbaum [40] and references therein for a good overview). In the present paper, however, we make a rare attempt to investigate a *discrete-time* queueing model with customer impatience, by means of a simple analytical model. Specifically, we study a GI/1/1 queue where the patience times (or "deadlines") of the customers are independent and geometrically distributed. We are able to obtain exact formulas for the probability generating function and the mean of the system content, the mean customer delay and the deadline-expiration ratio. These formulas, however, contain infinite sums and infinite products, and are thus not quite useful to see the impact of the various system parameters. Therefore, we also derive some easy-to-evaluate approximate results for the main performance measures. Jean-Marie and Hyon [20] consider the same model, but are interested in optimization rather than in in-depth structural analysis. They show that the optimal control of service in the GI/1/1+Geo queue is a threshold policy and they give the value of this threshold. In Kim et al. [24], furthermore, the authors study a discrete-time multi-server queue in which the customers arrive according to a simple Bernoulli process, in which the service times are geometrically distributed, and in which the customers wait for service for a limited time with a general distribution. They present exact expressions for the loss probability and the queue-length distribution. Van Velthoven et al. [34] derive an expression for the probability of abandonment in a Geo/Geo/1+GI queue and show that systems with a smaller patience distribution in the convex-ordering sense give rise to fewer abandonments (due to impatience), irrespective of whether customers become patient when entering the service facility. Finally, Wu et al. [37] combine the concept of customer impatience with the concepts of retrials and priorities in a discrete-time Geo/G/1 queue. They analyze the Markov chain underlying the considered queueing system and obtain the system-state distribution as well as the orbit-size and the system-size distributions in terms of their generating functions. Besides, they investigate a stochastic decomposition property and the corresponding continuous-time queueing system.

The contributions of the present paper concern the specific model that is considered and the methods that are used to obtain exact and approximate results for the main performance measures. As for the model, it is, as far as we know, the first attempt to perform a structural analysis of a *discrete-time* queue with customer impatience and a *general* independent arrival process. We are totally aware of the simplicity of the service process and of the abandonment process. However, since we want the focus of this paper to lie on the proposed approximation method, we have kept these processes as simple as possible. Of course, we will focus on generalizations of these processes in the future. It is commonly known from the continuous-time literature on queues with customer impatience that exact results of even simple models are often far from user-friendly and that more general models may soon become analytically intractable, so *approximations* have to be proposed to study these systems. We mention a

few of them; we refer to Xiong et al. [38] and Sakuma et al. [32] for more. Boxma and de Waal [8] were amongst the first who developed several approximations for the probability to abandon in the (continuous-time) M/G/n+G queue. These approximations, based on intuition and observations, use the exact results obtained for the M/M/n+M and M/M/n+D cases. Extensive tests of these approximations reveal a near-insensitivity of the overflow probability with respect to the service-time distribution, and - apart from a small traffic region - a rather weak sensitivity with respect to the patience distribution. In [38], the authors propose a methodology for approximating the mean waiting time by mapping a multi-server queue to a single server queue with an augmented service rate. The main objective of [32], finally, is to provide an approximation for the waiting-time distribution in an analytically tractable form. Their approximation is based on the tail asymptotics of this distribution, under the condition that the impatient time is an unbounded and asymptotically light-tailed random variable. In our paper, we propose to use a *polynomial (finite power-series)* approximation method to estimate the main performance measures. This technique is not completely new (see, e.g., Blanc [3] and Hooghiemstra et al [17]). However, whereas in these papers the power-series approximation is usually expressed in terms of powers of the load or the traffic intensity of the system, we use the patience distribution parameter to form the power series. In this way, we arrive at approximate, yet much simpler expressions than with the "exact" method. We show with some numerical examples that the approximate expressions are quite accurate. Moreover, they are much more suitable to study the influence of the system parameters. Therefore, we believe that this solution technique can be useful for more advanced queueing models with customer impatience. Note that the approach with power series in a parameter other than the load has also proven to be useful in other types of queueing models, e.g., a Generalized Processor Sharing model [35] and a model with train arrivals [11].

2 Mathematical model

We consider a discrete-time queueing system with one single server and an infinite waiting room. As in all discrete-time models, the time axis is divided into fixed-length intervals referred to as *slots*. New customers may enter the system at any given (continuous) point on the time axis, but services are synchronized to (i.e., can only start and end at) slot boundaries. We assume that the service of each customer requires exactly one slot.

The arrival process of new customers in the system is characterized by means of a sequence of i.i.d. nonnegative discrete random variables with common probability mass function (pmf) $a(n)$ and common probability generating function (pgf) $A(z)$, respectively. More specifically,

$$a(n) \triangleq \text{Prob}[\ n \text{ customer arrivals in one slot }] \quad , \quad n \geq 0 \ ,$$

$$A(z) \triangleq \sum_{n=0}^{\infty} a(n)\, z^n \ .$$

The mean number of customer arrivals per slot, in the sequel referred to as the *(mean) arrival*

rate, is given by

$$\lambda \triangleq A'(1) \ .$$

For ease of notation further in the paper, we also define the following arrival-process related quantities:

$$\alpha \triangleq A(0) \quad ; \quad \beta \triangleq 1 - A(0) \quad ; \quad \gamma \triangleq \lambda - 1 + A(0) \ . \tag{1}$$

As mentioned above, the special feature of the queueing model at hand is the fact that customers may leave the system before they have actually received service. Here we make a distinction between the *queue*, which collects the customers that are actually waiting for service, and the *system*, which encompasses all the customers, either waiting or being served. Customers that have entered the server, possibly after having spent some time in the queue, stay in the system until they have received service. However, no customer stays in the queue longer than a prescribed maximum time duration, referred to as the *deadline* of the customer. We assume that the deadlines of the customers may be different from one customer to another, but they are statistically independent and geometrically distributed with parameter σ, i.e., the pmf $s(n)$ and the pgf $S(z)$ of the deadlines are given by

$$s(n) \triangleq \text{Prob}[\text{ deadline equals } n \text{ slots }] = (1-\sigma)\sigma^{n-1} \ , \quad n \geq 1 \ ,$$

$$S(z) \triangleq \sum_{n=1}^{\infty} s(n)\, z^n = \frac{(1-\sigma)z}{1-\sigma z} \ .$$

The mean deadline is given by

$$D \triangleq S'(1) = \frac{1}{1-\sigma} \ . \tag{2}$$

The geometric nature of the deadlines implies that the probability that the deadline of any customer C, in the queue at the beginning of any slot S, expires at the end of the slot S does not depend on the amount of time customer C already spent waiting in the queue, and is simply given by $1-\sigma$. This property is crucial in the analysis of the system in the next sections.

The structure of the rest of this paper is as follows. In section 3, we analyze the queueing performance of the system, resulting in an exact yet complicated expression for the pgf of the number of customers in the system. Section 4 is devoted to an alternative approach in which we express the pgf of the system content in the form of (several) polynomials in the parameter σ. Here, we are able to obtain much simpler, but also approximate, results than in the previous section. From both types of results, we derive (complicated) nearly exact and (easier) approximate expressions for the mean system content, the mean customer delay and the deadline-expiration ratio, in section 5. In section 6, we compare the different approximations and illustrate our results by means of some numerical examples. Section 7 states some conclusions and indicates some possible future work.

3 Exact analysis of the system-content pgf

Let u_k denote the system content, i.e., the total number of customers present in the system, at the beginning of the k-th slot, and a_k the number of customers entering the system during

this slot. Then, the number of customers in the queue, i.e., the number of customers that can potentially abandon from the system prematurely, can be expressed as $(u_k-1)^+$, by introducing the notation $(\ldots)^+$ to indicate the quantity $\max(0,\ldots)$. It follows that the following recursive system equation can be established between u_k and u_{k+1}:

$$u_{k+1} = \sum_{i=1}^{(u_k-1)^+} c_{i,k} + a_k \ , \tag{3}$$

In the above equation, the $c_{i,k}$'s are a sequence of i.i.d. Bernoulli random variables with parameter σ, i.e., with common pgf

$$C(z) \triangleq 1 - \sigma + \sigma z \ . \tag{4}$$

Specifically, $c_{i,k}$ can be interpreted as the indicator function (taking values 1 or 0) of the event that the i-th customer in the queue at the beginning of slot k stays in the queue at the end of slot k.

For all k, let $U_k(z)$ denote the pgf of u_k. Then, from equation (3) we can derive

$$U_{k+1}(z) = A(z) \cdot E\left[z^{\sum_{i=1}^{(u_k-1)^+} c_{i,k}}\right] \ , \tag{5}$$

with $E[\cdot]$ the expectation operator. Here, the second factor in the right hand side of (5) can be expanded further by means of the law of total probability, yielding

$$U_{k+1}(z) = A(z)\left[\text{Prob}[u_k = 0] + \sum_{j=1}^{\infty} \text{Prob}[u_k = j]\,[C(z)]^{j-1}\right] \ ,$$

which can be rewritten as

$$U_{k+1}(z) = A(z)\left[U_k(0) + \frac{U_k(C(z)) - U_k(0)}{C(z)}\right] \ . \tag{6}$$

Now, let us assume that the queueing system at hand is stable. In fact, it is not difficult to see that the system is always stable if the parameter σ is strictly less than 1, because in that case the deadlines are finite (with probability 1) and, hence, also the sojourn times of the customers in the system are necessarily finite. On the other hand, if $\sigma = 1$, the system reduces to a simple discrete-time buffer without deadlines, which is stable if and only if the mean number of customers entering the system per slot, given by λ, is strictly less than 1. We now let the time parameter k go to infinity in equation (6). Assuming the system reaches a steady state, then both functions $U_k(\cdot)$ and $U_{k+1}(\cdot)$ converge to a common limit function $U(\cdot)$, which denotes the pgf of the system content at the beginning of an arbitrary slot in steady state. As a result, equation (6) translates into

$$U(z) = F(z)\,[U(C(z)) + (C(z) - 1)U(0)] \ , \tag{7}$$

where

$$F(z) \triangleq \frac{A(z)}{C(z)} \ . \tag{8}$$

We are now faced with the problem of solving the (non-classical) functional equation (7). We note that very similar functional equations occur in the analysis of (continuous-time) queueing models with so-called "synchronized services" [13, 23]. These are queueing systems in which all customers present in the system are served simultaneously (from beginning to end) and leave the system (at exactly the same time instant) with some fixed probability at the end of each service cycle. Although such queueing models are quite different from our model, the formal resemblance of the resulting functional equations is striking. We refer to [13, 23] for a more formal justification of the solution method to be explained next.

First, however, we note that if $\sigma = 1$, the solution of (7) is very simple, because in that case $C(z) = z$ and (7) is, in fact, a simple linear equation for $U(z)$ with the well-known [10] solution

$$U(z)|_{\sigma=1} = \frac{(1-\lambda)(z-1)A(z)}{z - A(z)} \ , \tag{9}$$

which is valid for $\lambda < 1$. The corresponding probability of an empty buffer is then given by

$$U(0)|_{\sigma=1} = 1 - \lambda \tag{10}$$

and the mean system content is

$$E[u]\,|_{\sigma=1} = \lambda + \frac{A''(1)}{2(1-\lambda)} \ . \tag{11}$$

If $\sigma < 1$, however, the problem is much less trivial. One way to proceed is to use equation (7) recursively (similarly as in [13, 23]), as follows:

$$\begin{aligned}
U(z) =\ & F(z)\,[U(1-\sigma+\sigma z) + \sigma(z-1)U(0)] \\
=\ & F(z)F(1-\sigma+\sigma z)U(1-\sigma^2+\sigma^2 z) \\
& + F(z)F(1-\sigma+\sigma z)\sigma^2(z-1)U(0) + F(z)\sigma(z-1)U(0) \\
=\ & F(z)F(1-\sigma+\sigma z)F(1-\sigma^2+\sigma^2 z)U(1-\sigma^3+\sigma^3 z) \\
& + F(z)F(1-\sigma+\sigma z)F(1-\sigma^2+\sigma^2 z)\sigma^3(z-1)U(0) \\
& + F(z)F(1-\sigma+\sigma z)\sigma^2(z-1)U(0) + F(z)\sigma(z-1)U(0) \\
=\ & \cdots \\
=\ & U(1-\sigma^{m+1}+\sigma^{m+1}z)\prod_{i=0}^{m} F(1-\sigma^i+\sigma^i z) + U(0)(z-1)\sum_{i=0}^{m}\sigma^{i+1}\prod_{j=0}^{i} F(1-\sigma^j+\sigma^j z) \ ,
\end{aligned} \tag{12}$$

which is valid for all integers $m \geq 0$.

Taking the limit of the above result for $m \to \infty$ and using the fact that $\lim_{m\to\infty} \sigma^{m+1} = 0$ because $\sigma < 1$, we obtain

$$U(z) = \; U(1) \prod_{i=0}^{\infty} F(1-\sigma^i+\sigma^i z) + U(0)(z-1) \sum_{i=0}^{\infty} \sigma^{i+1} \prod_{j=0}^{i} F(1-\sigma^j+\sigma^j z) \; . \tag{13}$$

The quantity $U(1)$ in the right hand side of (13) is known to be $U(1) = 1$ (normalization of the system-content distribution). The remaining unknown $U(0)$ can now be computed by choosing $z = 0$ in (13) and solving the resulting equation for U(0), which leads to

$$U(0) = \frac{\prod_{i=0}^{\infty} F(1-\sigma^i)}{1 + \sum_{i=1}^{\infty} \sigma^i \prod_{j=0}^{i-1} F(1-\sigma^j)} \; . \tag{14}$$

An explicit expression for the pgf $U(z)$ is therefore given by

$$U(z) = \prod_{i=0}^{\infty} F(1-\sigma^i+\sigma^i z) + \frac{\prod_{i=0}^{\infty} F(1-\sigma^i)}{1 + \sum_{i=1}^{\infty} \sigma^i \prod_{j=0}^{i-1} F(1-\sigma^j)} (z-1) \sum_{i=0}^{\infty} \sigma^{i+1} \prod_{j=0}^{i} F(1-\sigma^j+\sigma^j z) \; . \tag{15}$$

We note that choosing $z = 0$ in (12) and solving the resulting equation for U(0), we can also easily derive an exact relation between the quantities $U(0)$ and $U(1-\sigma^{m+1})$, which will prove to be useful further on. The result is

$$U(0) = U(1-\sigma^{m+1}) \frac{\prod_{i=0}^{m} F(1-\sigma^i)}{1 + \sum_{i=1}^{m+1} \sigma^i \prod_{j=0}^{i-1} F(1-\sigma^j)} \; . \tag{16}$$

Equation (16) is valid for all $m \geq 0$; in the limit for $m \to \infty$ it reduces to (14).

The formulas (14) and (15) are exact, but not very transparent as to the dependence of $U(0)$ and $U(z)$ (and the performance measures that could be derived from them) on the various parameters of the system under study. Specifically, the main drawback of the formulas obtained so far is that they are expressed in terms of infinite sums and infinite products which may not, in all cases, allow the easy computation of numerical results. In an attempt to overcome these difficulties, we propose an alternative approximative approach in the next section.

4 Polynomial approximation of the system-content pgf

From the results in the previous section it is clear that the performance of the system depends both on the characteristics of the arrival process, i.e., the pgf $A(z)$, and the deadline parameter σ. In this section, we aim for an approximative representation of the pgf $U(z)$ (and the performance measures derived from it) in the form of a polynomial (finite power series) in the parameter σ. In particular, we approximate $U(z)$ as

$$U(z) \approx \sum_{i=0}^{L} \sigma^i V_i(z) \; , \tag{17}$$

where L is some finite integer and the functions $V_i(z)$ are independent of σ, but, of course, still dependent on the arrival characteristics. Instead of trying to solve the functional equation (7) for the pgf $U(z)$, we now focus on the derivation of the functions $V_i(z)$, for $0 \leq i \leq L$. Note that, owing to the fact that L is *finite*, the expression in (17), seen as a function of the parameter σ, always converges. It seems reasonable to expect that the accuracy of the polynomial approximation increases with L. Ideally, however, in order not to complicate the computations, a "low" value of L should suffice to obtain accurate results. In the sequel, we choose $L = 3$ and $L = 4$; both values are compared in section 6. It can be seen that our approach is somewhat similar to the well-known power series algorithm (PSA), discussed originally in [3, 17] and frequently used by Blanc in other papers (see, e.g., [4, 5, 6]). The difference is that in the PSA, as used in the papers just mentioned, *infinite* power series are used (which rises the issue of the convergence of these series) and the parameter of interest is not σ (as in our case), but the traffic intensity or the utilization factor of the queue at hand. Furthermore, the PSA is usually applied to express equilibrium probabilities of some Markov chain in the form of a power series, whereas we apply it to approximate a pgf.

In order to obtain equations for the $V_i(z)$'s, we first determine series expansions for all the quantities appearing in the crucial functional equation (7). The pgf $C(z)$ is given by

$$C(z) = 1 + \sigma(z-1) \ ,$$

i.e., linear in σ.

Next, the quantity $U(C(z))$ can also be expressed as a power series in the parameter σ, by means of a Taylor series expansion of the function U about the value 1 of its argument:

$$\begin{aligned} U(C(z)) &= U(1 + \sigma(z-1)) \\ &= \sum_{k=0}^{\infty} \frac{(\sigma(z-1))^k}{k!} U^{(k)}(1) \ , \end{aligned}$$

where $U^{(k)}$ indicates the k-th derivative of the function U. By introducing the expansion (17) and keeping powers of σ up to the fourth, we can transform this into

$$\begin{aligned} U(C(z)) \approx \ & 1 + \sigma(z-1)\left[V_0'(1) + \sigma V_1'(1) + \sigma^2 V_2'(1) + \sigma^3 V_3'(1)\right] \\ & + \frac{\sigma^2(z-1)^2}{2}\left[V_0''(1) + \sigma V_1''(1) + \sigma^2 V_2''(1)\right] \\ & + \frac{\sigma^3(z-1)^3}{6}[V_0'''(1) + \sigma V_1'''(1)] \\ & + \frac{\sigma^4(z-1)^4}{24} V_0^{(4)}(1) \ , \end{aligned}$$

which is valid if σ is "low enough". Here we have also used the normalization condition of the system-content distribution, $U(1) = 1$.

In view of equation (8), the functional equation (7) can now be rewritten as

$$C(z)U(z) = A(z)\left[U(C(z)) + \sigma(z-1)U(0)\right] \ ,$$

or, expanding both sides of the equation in powers of σ,

$$\begin{aligned}
&[1+\sigma(z-1)][V_0(z)+\sigma V_1(z)+\sigma^2 V_2(z)+\sigma^3 V_3(z)+\sigma^4 V_4(z)] \\
&\quad\approx\ A(z)+\sigma(z-1)A(z)[V_0'(1)+\sigma V_1'(1)+\sigma^2 V_2'(1)+\sigma^3 V_3'(1)] \\
&\qquad+\frac{\sigma^2(z-1)^2}{2}A(z)[V_0''(1)+\sigma V_1''(1)+\sigma^2 V_2''(1)] \\
&\qquad+\frac{\sigma^3(z-1)^3}{6}A(z)[V_0'''(1)+\sigma V_1'''(1)] \\
&\qquad+\frac{\sigma^4(z-1)^4}{24}A(z)V_0^{(4)}(1) \\
&\qquad+\sigma(z-1)A(z)[V_0(0)+\sigma V_1(0)+\sigma^2 V_2(0)+\sigma^3 V_3(0)] \ . \qquad (18)
\end{aligned}$$

We can now identify the coefficients of equal powers of σ on both sides of the equation (18) to determine explicit expressions for the functions $V_i(z)$ for $i \geq 0$. For the coefficients of σ^0 we easily get

$$V_0(z) = A(z) \ . \qquad (19)$$

Next, for σ^1, we obtain

$$V_1(z) + (z-1)V_0(z) = (z-1)A(z)[V_0'(1) + V_0(0)] \ ,$$

where, from (19),

$$V_0'(1) = A'(1) = \lambda$$

$$V_0(0) = A(0) = \alpha \ ,$$

so that

$$V_1(z) = (z-1)A(z)[\lambda - 1 + A(0)] = \gamma(z-1)A(z) \ . \qquad (20)$$

Identifying coefficients of σ^2 leads to

$$V_2(z) + (z-1)V_1(z) = (z-1)A(z)[V_1'(1) + \frac{(z-1)}{2}V_0''(1) + V_1(0)] \ ,$$

where, from (19)-(20),

$$V_1'(1) = \gamma$$

$$V_0''(1) = A''(1)$$

$$V_1(0) = -\alpha\gamma \ ,$$

so that

$$V_2(z) = (z-1)A(z)\left\{\beta\gamma + (z-1)[-\gamma + \frac{A''(1)}{2}]\right\} \ . \tag{21}$$

Further, identification of the coefficients of σ^3 yields

$$V_3(z) + (z-1)V_2(z) = (z-1)A(z)[V_2'(1) + \frac{(z-1)}{2}V_1''(1) + \frac{(z-1)^2}{6}V_0'''(1) + V_2(0)] \ ,$$

where, from (19)-(21),

$$V_2'(1) = \beta\gamma$$

$$V_1''(1) = 2\lambda\gamma$$

$$V_0'''(1) = A'''(1)$$

$$V_2(0) = -\alpha\left\{(1+\beta)\gamma - \frac{A''(1)}{2}\right\} \ ,$$

so that

$$V_3(z) = (z-1)A(z)\Big\{(\beta^2-\alpha)\gamma + \alpha\frac{A''(1)}{2} + (z-1)\gamma^2$$

$$+ (z-1)^2[\gamma + \frac{A'''(1) - 3A''(1)}{6}]\Big\} \ . \tag{22}$$

Finally, for the coefficients of σ^4, we obtain

$$V_4(z) + (z-1)V_3(z) = (z-1)A(z)\Big\{V_3'(1) + \frac{z-1}{2}V_2''(1) + \frac{(z-1)^2}{6}V_1'''(1)$$

$$+ \frac{(z-1)^3}{24}V_0^{(4)}(1) + V_3(0)\Big\} \ ,$$

where, from (19)-(22),

$$V_3'(1) = (\beta^2-\alpha)\gamma + \alpha\frac{A''(1)}{2}$$

$$V_2''(1) = 2\gamma(\beta\lambda - 1) + A''(1)$$

$$V_1'''(1) = 3\gamma A''(1)$$

$$V_0^{(4)}(1) = A^{(4)}(1)$$

$$V_3(0) = -\alpha\Big\{(\beta^2+\beta-\gamma)\gamma - \beta\frac{A''(1)}{2} + \frac{A'''(1)}{6}\Big\} \ ,$$

so that

$$V_4(z) = (z-1)A(z)\Big\{\gamma[\beta^3 + \alpha(\gamma-\beta-1)] + \alpha(1+\beta)\frac{A''(1)}{2} - \alpha\frac{A'''(1)}{6}$$

$$+ (z-1)\beta\big[\gamma(\gamma-1) + \frac{A''(1)}{2}\big] + (z-1)^2\gamma\big[-\gamma + \frac{A''(1)}{2}\big]$$

$$+ (z-1)^3\big[-\gamma + \frac{A''(1)}{2} - \frac{A'''(1)}{6} + \frac{A^{(4)}(1)}{24}\big]\Big\} \tag{23}$$

Combining the results in equations (19), (20), (21), (22), and (23), we now dispose of the following approximate expressions ($\hat{U}(z)_3$ and $\hat{U}(z)_4$) for the pgf $U(z)$:

$$\hat{U}(z)_3 \triangleq V_0(z) + \sigma V_1(z) + \sigma^2 V_2(z) + \sigma^3 V_3(z) \quad , \quad \text{for } L = 3 \ ,$$

$$\hat{U}(z)_4 \triangleq V_0(z) + \sigma V_1(z) + \sigma^2 V_2(z) + \sigma^3 V_3(z) + \sigma^4 V_4(z) \quad , \quad \text{for } L = 4 \ .$$

In explicit form, these expressions can be written as

$$\hat{U}(z)_3 = A(z) + \sigma\gamma(z-1)A(z) + \sigma^2(z-1)A(z)\left\{\beta\gamma + (z-1)[-\gamma + \frac{A''(1)}{2}]\right\}$$

$$+ \sigma^3(z-1)A(z)\Big\{(\beta^2-\alpha)\gamma + \alpha\frac{A''(1)}{2} + (z-1)\gamma^2 + (z-1)^2[\gamma + \frac{A'''(1) - 3A''(1)}{6}]\Big\} \tag{24}$$

and

$$\hat{U}(z)_4 = \hat{U}(z)_3 + \sigma^4(z-1)A(z)\Big\{\gamma[\beta^3 + \alpha(\gamma-\beta-1)] + \alpha(1+\beta)\frac{A''(1)}{2} - \alpha\frac{A'''(1)}{6}$$

$$+ (z-1)\beta\big[\gamma(\gamma-1) + \frac{A''(1)}{2}\big] + (z-1)^2\gamma\big[-\gamma + \frac{A''(1)}{2}\big]$$

$$+ (z-1)^3\big[-\gamma + \frac{A''(1)}{2} - \frac{A'''(1)}{6} + \frac{A^{(4)}(1)}{24}\big]\Big\} \tag{25}$$

5 Derivation of performance metrics

In principle, various moments of the system-content distribution can be obtained by computing derivatives of the exact expression (15) for the pgf $U(z)$ at $z = 1$. The presence of (infinite) sums and products, however, does not always allow an easy computation of these moments.

To find the mean system content, for example, we first have to calculate $U'(z)$. From (15), we obtain that

$$\begin{aligned} U'(z) = &\sum_{i=0}^{\infty} \sigma^i F'(1-\sigma^i+\sigma^i z) \prod_{j=0,j\neq i}^{\infty} F(1-\sigma^j+\sigma^j z) \\ &+ \frac{\prod_{i=0}^{\infty} F(1-\sigma^i)}{1+\sum_{i=1}^{\infty}\sigma^i \prod_{j=0}^{i-1} F(1-\sigma^j)} \sum_{i=0}^{\infty} \sigma^{i+1} \Bigg\{ \prod_{j=0}^{i} F(1-\sigma^j+\sigma^j z) \\ &+ (z-1)\sum_{j=0}^{i} \sigma^j F'(1-\sigma^j+\sigma^j z) \prod_{k=0,k\neq j}^{i} F(1-\sigma^k+\sigma^k z) \Bigg\}. \end{aligned} \tag{26}$$

Choosing $z=1$ in this equation then yields

$$\begin{aligned} E[u] &= U'(1) \\ &= \sum_{i=0}^{\infty} \sigma^i F'(1) + \frac{\prod_{i=0}^{\infty} F(1-\sigma^i)}{1+\sum_{i=1}^{\infty}\sigma^i \prod_{j=0}^{i-1} F(1-\sigma^j)} \sum_{i=0}^{\infty} \sigma^{i+1} \\ &= \frac{\lambda-\sigma}{1-\sigma} + \frac{\prod_{i=0}^{\infty} F(1-\sigma^i)}{1+\sum_{i=1}^{\infty}\sigma^i \prod_{j=0}^{i-1} F(1-\sigma^j)} \frac{\sigma}{1-\sigma}, \end{aligned} \tag{27}$$

where we have assumed $\sigma < 1$ (in order for the geometric series in σ to converge) and used the definition of $F(z)$ in equation (8). Just like the expressions (14) and (15) for $U(0)$ and $U(z)$, this formula contains both an infinite sum and infinite products.

Now, rather than departing from the explicit expression (15) of $U(z)$, the moments of the system-content distribution can also be computed directly from the original functional equation (7). For instance, to find the mean system content, we take the first derivatives at $z=1$ of both sides of (7), i.e.,

$$U'(1) = F'(1) + U'(1)C'(1) + C'(1)U(0)$$

or

$$E[u] = \lambda - \sigma + E[u]\,\sigma + \sigma U(0) \ .$$

For $\sigma < 1$, this equation can be solved for $E[u]$, which easily leads to

$$E[u] = \frac{\lambda - \sigma(1-U(0))}{1-\sigma} \ . \tag{28}$$

Note that replacing $U(0)$ in equation (28) by expression (14) yields the same result as (27).

Next, by applying (the discrete-time version of) Little's theorem [25, 10, 15], the mean delay (system time) $E[d]$ of a customer can be obtained as

$$E[d] = \frac{E[u]}{\lambda} = \frac{\lambda - \sigma(1-U(0))}{\lambda(1-\sigma)} = D - \frac{(D-1)(1-U(0))}{\lambda} \ , \tag{29}$$

where D is the mean deadline of the customers, defined in (2). Equation (29) clearly illustrates that the mean delay of a customer cannot be higher than the mean deadline D, as expected intuitively.

Another quantity of interest in the context of a queueing system with deadlines is the fraction of customers that leave the queue unserved due to the expiration of their deadline before they can reach the server. We call this fraction the *deadline-expiration ratio* in the sequel. In order to compute this quantity from our earlier results, we make the following reasoning. From the fact that the system at hand is stable, it follows that the total mean number of customers leaving the system per time slot is identical to the mean arrival rate λ. Customers leave the system either because they have received service or because they abandon from the queue before reaching the server. A customer is served in a slot if and only if the system is non-empty at the beginning of that slot, i.e., with probability $1 - U(0)$. As the system only disposes of one single server, this is also equal to the mean number of customers that are served in a slot. It follows that the *deadline-expiration ratio* can be computed as

$$r_{ex} = \frac{\lambda - (1 - U(0))}{\lambda} \ , \tag{30}$$

where the numerator corresponds to the mean number of customers leaving the queue unserved per slot, i.e., the difference between the total mean number of departures in a slot (λ) and the mean number of customers receiving service per slot ($1 - U(0)$).

To summarize, as soon as the quantity $U(0)$ has been computed, the other performance measures $E[u]$, $E[d]$ and r_{ex} can be easily obtained from (28), (29) and (30).

5.1 Truncation approximation

As mentioned earlier, the exact expression (14) for $U(0)$ contains both an infinite sum and infinite products. One pragmatic way to circumvent these difficulties is to truncate the infinite sum and products in the expression for $U(0)$ as follows:

$$U(0) \approx U(0)_K \triangleq \frac{\prod_{i=0}^{K-1} F(1-\sigma^i)}{1 + \sum_{i=1}^{K-1} \sigma^i \prod_{j=0}^{i-1} F(1-\sigma^j)} \ , \tag{31}$$

where the integer K is such that $1 - \sigma^K$ is "close enough" to 1. (In practice, a value $K = 2000$ proves to be sufficient for most values of $\sigma < 1$.)

Replacing the quantity $U(0)$ by the truncation approximation $U(0)_K$ in (28) yields a first approximation for the mean system content $E[u]$. We refer to this approximation as $E[u]_K$, i.e.,

$$E[u]_K \triangleq \frac{\lambda - \sigma(1 - U(0)_K)}{1 - \sigma} \ . \tag{32}$$

In a similar way, we can substitute $U(0)_K$ for $U(0)$ in (29) and (30) to obtain approximations for $E[d]$ and r_{ex}; they are denoted by $E[d]_K$ and $r_{ex,K}$, respectively:

$$E[d]_K \triangleq \frac{\lambda - \sigma(1 - U(0)_K)}{\lambda(1 - \sigma)} \ , \tag{33}$$

$$r_{ex,K} \triangleq \frac{\lambda - (1 - U(0)_K)}{\lambda} \ . \tag{34}$$

5.2 Polynomial approximation

As an alternative to truncation, we can also use the simpler polynomial approximations (24) and (25) of the pgf $U(z)$ to derive explicit closed-form expressions for various performance measures of the queueing system at hand, in terms of the basic system parameters, i.e., the pgf $A(z)$ (and the related quantities α, β and γ) of the arrival process and the probability σ which characterizes the deadline distribution. First, we derive from (24) and (25) approximations $\hat{U}(0)_3$ and $\hat{U}(0)_4$ for the probability of an empty system:

$$\hat{U}(0)_3 \triangleq V_0(0) + \sigma V_1(0) + \sigma^2 V_2(0) + \sigma^3 V_3(0) \ ,$$

$$\hat{U}(0)_4 \triangleq V_0(0) + \sigma V_1(0) + \sigma^2 V_2(0) + \sigma^3 V_3(0) + \sigma^4 V_4(0) \ ,$$

which can be expressed more explicitly as

$$\begin{aligned}
\hat{U}(0)_3 &= \alpha - \alpha\gamma\sigma - \alpha\left\{(1+\beta)\gamma - \frac{A''(1)}{2}\right\}\sigma^2 - \alpha\left\{(\beta^2+\beta-\gamma)\gamma - \beta\frac{A''(1)}{2} + \frac{A'''(1)}{6}\right\}\sigma^3 \ , \\
\hat{U}(0)_4 &= \hat{U}(0)_3 - \alpha\left\{\beta\gamma(\beta^2 - 2\gamma + 1 + \beta) + [\gamma - \beta(1+\beta)]\frac{A''(1)}{2} + \beta\frac{A'''(1)}{6} - \frac{A^{(4)}(1)}{24}\right\}\sigma^4 \ .
\end{aligned} \tag{35}$$

Another, somewhat more involved, way to arrive at approximations for $U(0)$ is to apply the polynomial approximations (24) and (25) to compute approximate expressions for the quantity $U(1-\sigma^{m+1})$ and then use these in equation (16) to obtain estimations of $U(0)$. This results in two additional approximation formulas $\hat{U}(0)_{3,m}$ and $\hat{U}(0)_{4,m}$ for the probability of an empty system:

$$\begin{aligned}
\hat{U}(0)_{3,m} &\triangleq \hat{U}(1-\sigma^{m+1})_3 \frac{\prod_{i=0}^{m} F(1-\sigma^i)}{1 + \sum_{i=1}^{m+1} \sigma^i \prod_{j=0}^{i-1} F(1-\sigma^j)} \ , \\
\hat{U}(0)_{4,m} &\triangleq \hat{U}(1-\sigma^{m+1})_4 \frac{\prod_{i=0}^{m} F(1-\sigma^i)}{1 + \sum_{i=1}^{m+1} \sigma^i \prod_{j=0}^{i-1} F(1-\sigma^j)} \ ,
\end{aligned} \tag{36}$$

where

$$\hat{U}(1-\sigma^{m+1})_3 \triangleq V_0(1-\sigma^{m+1}) + \sigma V_1(1-\sigma^{m+1}) + \sigma^2 V_2(1-\sigma^{m+1}) + \sigma^3 V_3(1-\sigma^{m+1}) \ ,$$

$$\hat{U}(1-\sigma^{m+1})_4 \triangleq \hat{U}(1-\sigma^{m+1})_3 + \sigma^4 V_4(1-\sigma^{m+1}) \ .$$

Next, we use two different approaches to derive approximate expressions for the mean system content $E[u]$, referred to as $E[\hat{u}]_{\text{dir}}$ and $E[\hat{u}]_{\text{ind}}$, respectively. The first (so-called "direct")

approximation is obtained directly from the approximate expressions (24) or (25) of $U(z)$:

$$E[\hat{u}_3]_{\text{dir}} \triangleq \hat{U}'(1)_3 = V_0'(1) + \sigma V_1'(1) + \sigma^2 V_2'(1) + \sigma^3 V_3'(1) \ ,$$

$$E[\hat{u}_4]_{\text{dir}} \triangleq \hat{U}'(1)_4 = V_0'(1) + \sigma V_1'(1) + \sigma^2 V_2'(1) + \sigma^3 V_3'(1) + \sigma^4 V_4'(1) \ ,$$

or, explicitly,

$$\begin{aligned} E[\hat{u}_3]_{\text{dir}} &= \lambda + \gamma\sigma + \beta\gamma\sigma^2 + \left\{(\beta^2 - \alpha)\gamma + \alpha\frac{A''(1)}{2}\right\}\sigma^3 \ , \\ E[\hat{u}_4]_{\text{dir}} &= E[\hat{u}_3]_{\text{dir}} + \left\{\gamma[\beta^3 + \alpha(\gamma - 1 - \beta)] + \alpha(1+\beta)\frac{A''(1)}{2} - \alpha\frac{A'''(1)}{6}\right\}\sigma^4 \ . \end{aligned} \tag{37}$$

The second (so-called "indirect") approximation is obtained by substituting the polynomial approximations $\hat{U}(0)_3$ and $\hat{U}(0)_4$ and $\hat{U}(0)_{3,m}$ and $\hat{U}(0)_{4,m}$ for $U(0)$ in the exact equation (28):

$$\begin{aligned} E[\hat{u}_3]_{\text{ind}} &\triangleq \frac{\lambda - \sigma(1 - \hat{U}(0)_3)}{1 - \sigma} \quad , \quad E[\hat{u}_4]_{\text{ind}} \triangleq \frac{\lambda - \sigma(1 - \hat{U}(0)_4)}{1 - \sigma} \ , \\ E[\hat{u}_{3,m}]_{\text{ind}} &\triangleq \frac{\lambda - \sigma(1 - \hat{U}(0)_{3,m})}{1 - \sigma} \quad , \quad E[\hat{u}_{4,m}]_{\text{ind}} \triangleq \frac{\lambda - \sigma(1 - \hat{U}(0)_{4,m})}{1 - \sigma} \ . \end{aligned} \tag{38}$$

Corresponding approximations for the mean customer delay are given by

$$\begin{aligned} E\left[\hat{d}_3\right]_{\text{dir}} &\triangleq \frac{E[\hat{u}_3]_{\text{dir}}}{\lambda} \quad , \quad E\left[\hat{d}_4\right]_{\text{dir}} \triangleq \frac{E[\hat{u}_3]_{\text{dir}}}{\lambda} \quad , \\ E\left[\hat{d}_3\right]_{\text{ind}} &\triangleq \frac{E[\hat{u}_3]_{\text{ind}}}{\lambda} \quad , \quad E\left[\hat{d}_4\right]_{\text{ind}} \triangleq \frac{E[\hat{u}_4]_{\text{ind}}}{\lambda} \ , \\ E\left[\hat{d}_{3,m}\right]_{\text{ind}} &\triangleq \frac{E[\hat{u}_{3,m}]_{\text{ind}}}{\lambda} \quad , \quad E\left[\hat{d}_{4,m}\right]_{\text{ind}} \triangleq \frac{E[\hat{u}_{4,m}]_{\text{ind}}}{\lambda} \ . \end{aligned} \tag{39}$$

In view of equation (30), approximations for the deadline-expiration ratio can be computed as

$$\begin{aligned} \hat{r}_{ex,3} &= \frac{\lambda - (1 - \hat{U}(0)_3)}{\lambda} \quad , \quad \hat{r}_{ex,4} = \frac{\lambda - (1 - \hat{U}(0)_4)}{\lambda} \ , \\ \hat{r}_{ex,3,m} &= \frac{\lambda - (1 - \hat{U}(0)_{3,m})}{\lambda} \quad , \quad \hat{r}_{ex,4,m} = \frac{\lambda - (1 - \hat{U}(0)_{4,m})}{\lambda} \ . \end{aligned} \tag{40}$$

6 Discussion of results and numerical examples

In this section, we discuss the results obtained in the previous sections, both from a qualitative perspective and by means of some numerical examples. In particular, we also validate the approximate polynomial results against more accurate results, obtained by truncation of infinite sums and products. We examine the case of Poisson arrivals in subsection 6.1. Some other choices of the arrival distribution are discussed in subsection 6.2.

6.1 Poisson arrivals

In case of Poisson arrivals with mean λ, the pmf $a(n)$ and the pgf $A(z)$ are given by

$$\begin{aligned} a(n) &= e^{-\lambda}\frac{\lambda^n}{n!} \ , \quad n \geq 0 \ , \\ A(z) &= e^{\lambda(z-1)} \ . \end{aligned} \tag{41}$$

It then easily follows that $\alpha = e^{-\lambda}$, $\beta = 1-e^{-\lambda}$, $\gamma = \lambda - 1 + e^{-\lambda}$, and the consecutive derivatives of $A(z)$ for $z = 1$ are

$$A'(1) = \lambda \ , \ A''(1) = \lambda^2 \ , \ A'''(1) = \lambda^3 \ , \ A^{(4)}(1) = \lambda^4 \ . \tag{42}$$

6.1.1 Probability of empty system

In Fig. 1, we have plotted the "exact" result $U(0)_K$ (with $K = 2000$) and the (most simple) polynomial approximations $\hat{U}(0)_3$ and $\hat{U}(0)_4$ for the probability of an empty system, according to formulas (31) and (35) respectively, versus the arrival rate λ, for various values of the deadline-distribution parameter σ. (The curve for $\sigma = 1$ was obtained from the exact result for this case in equation (10).) The figure shows that the probability of an empty system decreases when the arrival rate increases, as expected. It also shows that this probability decreases more slowly when the deadlines get smaller, i.e., when σ decreases, which can be attributed to the fact that more customers leave the queue prematurely. Note that, when $\sigma < 1$, owing to the finite deadlines of the customers, the queue remains empty for $\lambda > 1$ with a positive probability, even though in this case more customers arrive per slot than the server can handle. The figure finally also illustrates the accuracy of the polynomial approximations $\hat{U}(0)_3$ and $\hat{U}(0)_4$, which is very good for all values of λ, as long as σ is not too high (say, $\sigma < 0.75$). For low values of λ (say, $\lambda < 0.5$) as well as for high values of λ (say, $\lambda > 1.5$), the σ-polynomials are very accurate for all σ. The figure also demonstrates that, as expected, keeping powers of σ up to the fourth yields more accurate results than keeping powers only up to the third, but the improvement is rather small. Therefore, we are convinced that, at least in case of Poisson arrivals, choosing $L = 3$ or $L = 4$ is really enough to obtain fairly accurate results from fairly simple formulas. (In the sequel, we will mainly consider the value $L = 4$.)

Very similar conclusions can be drawn from Fig. 2, where we have plotted $U(0)_K$ (with $K = 2000$) and $\hat{U}(0)_3$ and $\hat{U}(0)_4$ versus the deadline-distribution parameter σ, for various values of the arrival rate λ. This figure illustrates very clearly that the probability of an empty

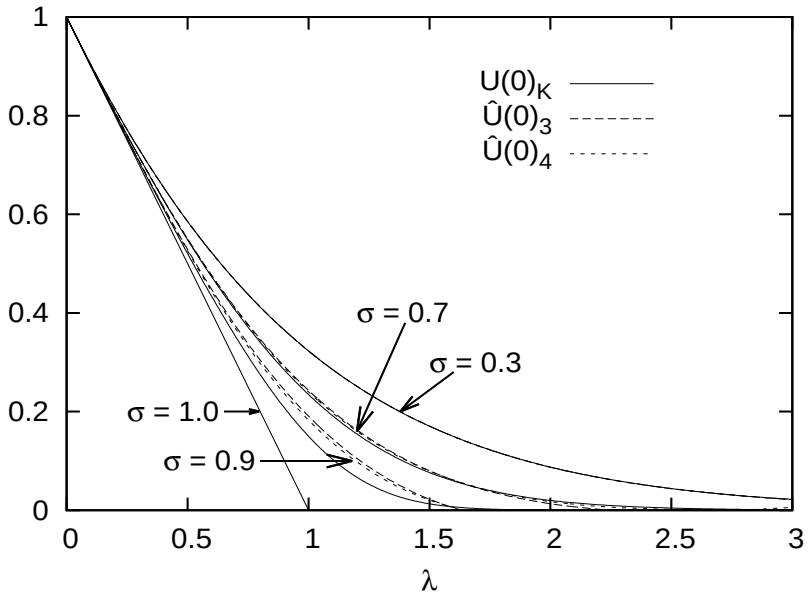

Figure 1: Probability of empty system: truncation approximation $U(0)_K$ and polynomial approximations $\hat{U}(0)_3$ and $\hat{U}(0)_4$, versus mean arrival rate λ, for Poisson arrivals and various values of the deadline-distribution parameter σ

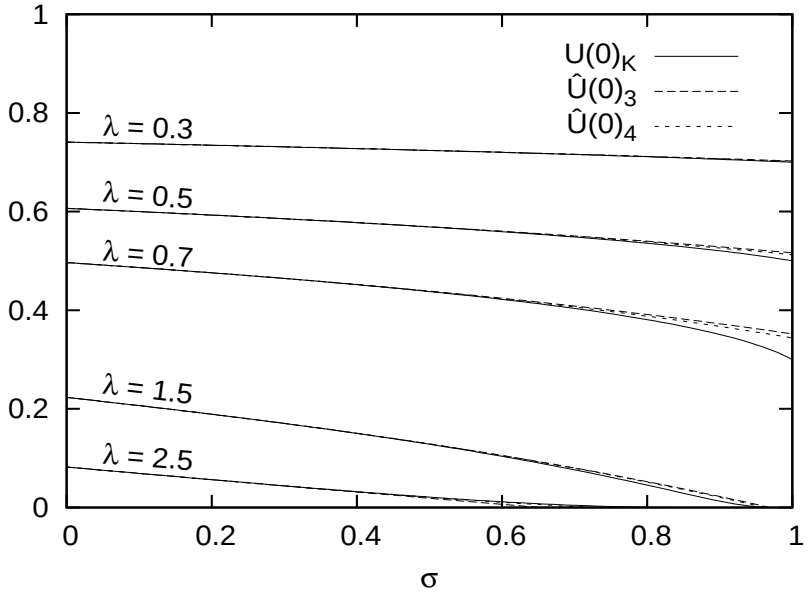

Figure 2: Probability of empty system: truncation approximation $U(0)_K$ and polynomial approximations $\hat{U}(0)_3$ and $\hat{U}(0)_4$, versus deadline-distribution parameter σ, for Poisson arrivals and various values of the mean arrival rate λ

system decreases when the deadlines get longer (i.e., for higher values of σ), because the number of customers that leave the queue prematurely goes down in these circumstances. The accuracy

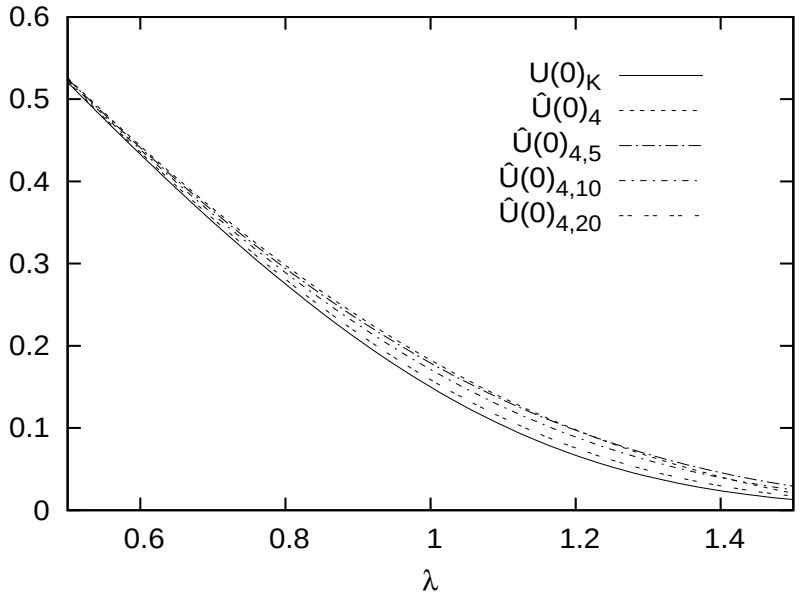

Figure 3: Probability of empty system: truncation approximation $U(0)_K$, simple polynomial approximation $\hat{U}(0)_4$ and more involved polynomial approximation $\hat{U}(0)_{4,m}$, versus mean arrival rate λ for Poisson arrivals, for $\sigma = 0.9$ and various values of m

of the polynomial approximation for $\sigma < 0.75$ and also for low and high values of λ is again very striking.

In figures 3 and 4, we focus on the "critical" parameter values where the polynomial approximations $\hat{U}(0)_3$ and $\hat{U}(0)_4$ were somewhat less accurate, i.e., $\sigma > 0.75$ and $0.5 < \lambda < 1.5$. Specifically, Fig. 3 reconsiders the special case where $\sigma = 0.9$, in the interval $0.5 < \lambda < 1.5$; the figure does not just show the (so far best) approximation $\hat{U}(0)_4$, but also the somewhat more involved approximation $\hat{U}(0)_{4,m}$, for various values of m. The results clearly illustrate that the accuracy of the approximations increases with m. Very similar conclusions can be drawn from Fig. 4, where we have depicted the approximations $\hat{U}(0)_4$ and $\hat{U}(0)_{4,m}$, for various m, versus σ for two "critical" values of λ, i.e., $\lambda = 0.7$ and $\lambda = 1.2$.

We may conclude from the above results that, in case of Poisson arrivals, $U(0)$ can usually be quite well approximated by the simple estimation $\hat{U}(0)_4$. If a higher precision is required, we advise to use the somewhat more complicated formula $\hat{U}(0)_{4,m}$. In theory, this approximation becomes "exact" when $m \to \infty$ (or $m = K$); in practice, a value around $m = 20$ seems sufficient for good accuracy, even for "critical" parameter values, as long as σ is not too close to its "forbidden" value $\sigma = 1$, for which the system reduces to a queue without deadlines.

6.1.2 Mean system content

Fig. 5 shows the "exact" results $E[u]_K$ (again, for $K = 2000$), and the four simplest power-series approximations $E[\hat{u}_3]_{\text{dir}}$, $E[\hat{u}_4]_{\text{dir}}$, $E[\hat{u}_3]_{\text{ind}}$ and $E[\hat{u}_4]_{\text{ind}}$, for the mean system content, versus the mean arrival rate λ, for various values of the deadline-distribution parameter σ. As

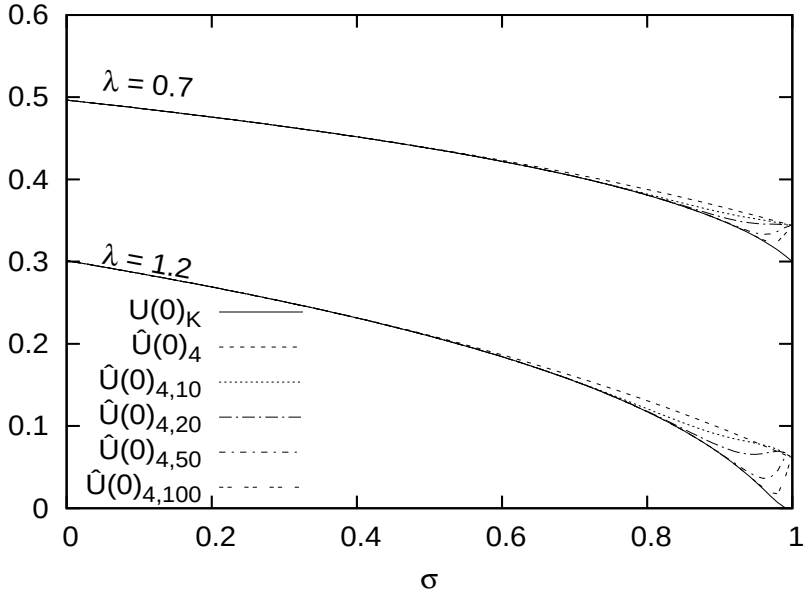

Figure 4: Probability of empty system: truncation approximation $U(0)_K$, simple polynomial approximation $\hat{U}(0)_4$ and more involved polynomial approximation $\hat{U}(0)_{4,m}$, versus deadline-distribution parameter σ, for Poisson arrivals with mean arrival rate $\lambda = 0.7$ and $\lambda = 1.2$, and various values of m

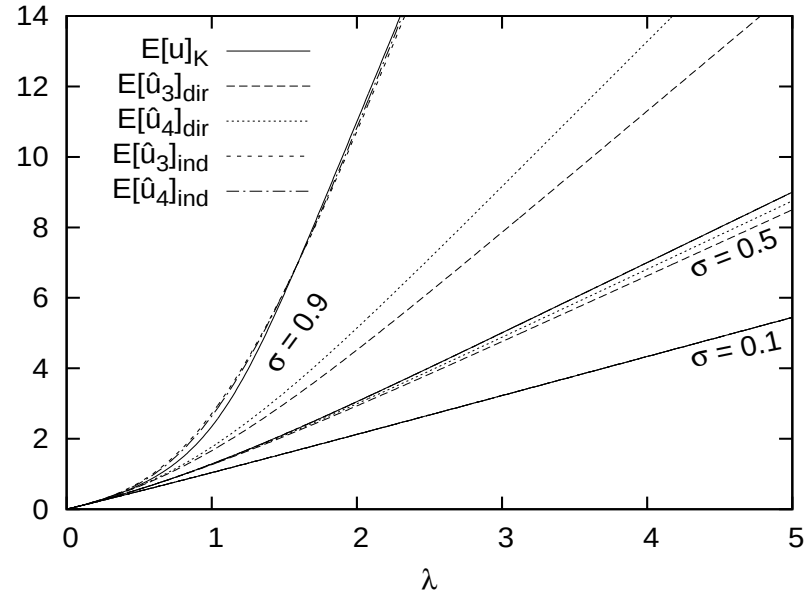

Figure 5: Mean system content: truncation approximation $E[u]_K$, direct polynomial approximations $E[\hat{u}_3]_{\text{dir}}$ and $E[\hat{u}_4]_{\text{dir}}$ and indirect polynomial approximations$E[\hat{u}_3]_{\text{ind}}$ and $E[\hat{u}_4]_{\text{ind}}$, versus mean arrival rate λ, for Poisson arrivals and various values of the deadline-distribution parameter σ

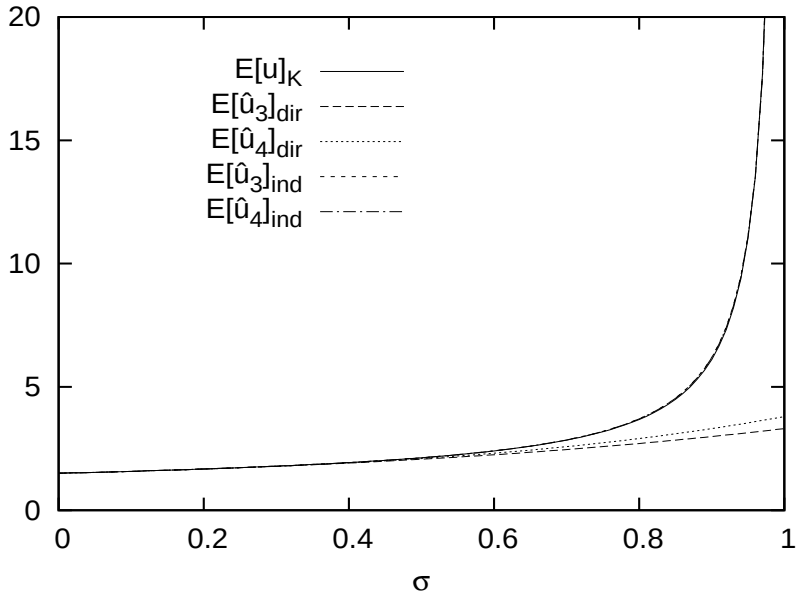

Figure 6: Mean system content: truncation approximation $E[u]_K$, direct polynomial approximations $E[\hat{u}_3]_{\rm dir}$ and $E[\hat{u}_4]_{\rm dir}$ and indirect polynomial approximations$E[\hat{u}_3]_{\rm ind}$ and $E[\hat{u}_4]_{\rm ind}$, versus the deadline-distribution parameter σ, for Poisson arrivals with $\lambda = 1.5$

expected, all the curves increase with λ. The figure also makes clear that, for a given arrival rate λ, the mean system content increases when the deadlines become longer, i.e., when the parameter σ takes higher values. Again, we note that the system remains stable for $\lambda > 1$, due to the finite length of the deadlines (if $\sigma < 1$). We also observe again that keeping powers of σ up to the fourth is always better than up to the third.

Fig. 5 suggests that the second ("indirect") polynomial approximation is always more accurate than the first ("direct") one, and is very close to the exact results for all values of σ. In fact, this is true for most values of λ, except for "small" λ, as is clearly illustrated in Figs. 6 and 7, where we have plotted the "exact" and approximate values of the mean system content as functions of the deadline parameter σ, for two different values of λ (1.5 and 0.3 respectively). For high values of λ (Fig. 6), the second approximation is clearly the best, for all values of σ. For smaller values of λ (Fig. 7), $E[\hat{u}]_{\rm dir}$ and $E[\hat{u}]_{\rm ind}$ perform equally well (irrespective of the value of L), as long as $\sigma < 0.85$. Only for high σ, the direct approximation performs better than the indirect one: the latter goes to infinity for $\sigma \to 1$, whereas the former remains finite for $\sigma \to 1$ when $\lambda < 1$ (as it should). Fig. 8 shows that replacing $E[\hat{u}_4]_{\rm ind}$ by the somewhat more complicated approximation $E[\hat{u}_{4,m}]_{\rm ind}$ can mitigate the reduced accuracy of the indirect approximation in case $\lambda < 1$ up to very high values of $\sigma < 1$. For all the above reasons, we only use the second ("indirect") polynomial approximation in the sequel.

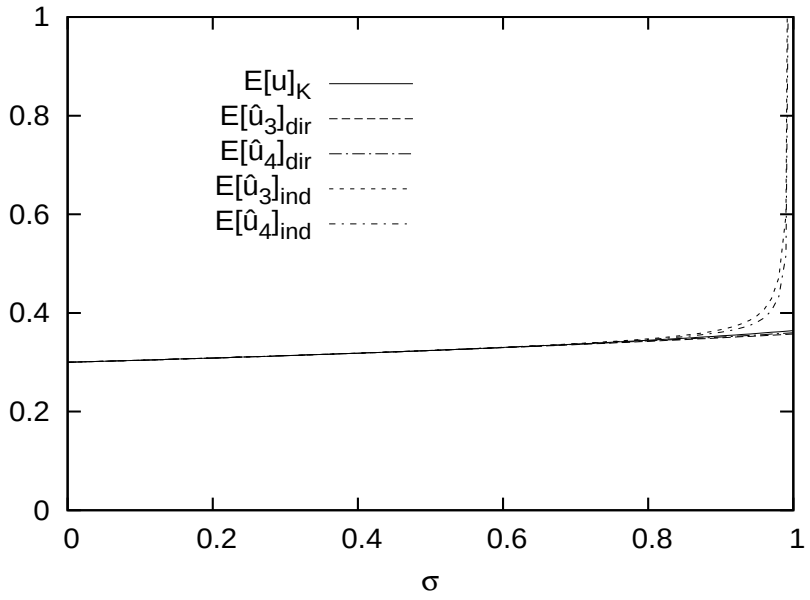

Figure 7: Mean system content: truncation approximation $E[u]_K$, direct polynomial approximations $E[\hat{u}_3]_{\text{dir}}$ and $E[\hat{u}_4]_{\text{dir}}$ and indirect polynomial approximations$E[\hat{u}_3]_{\text{ind}}$ and $E[\hat{u}_4]_{\text{ind}}$, versus the deadline-distribution parameter σ, for Poisson arrivals with $\lambda = 0.3$

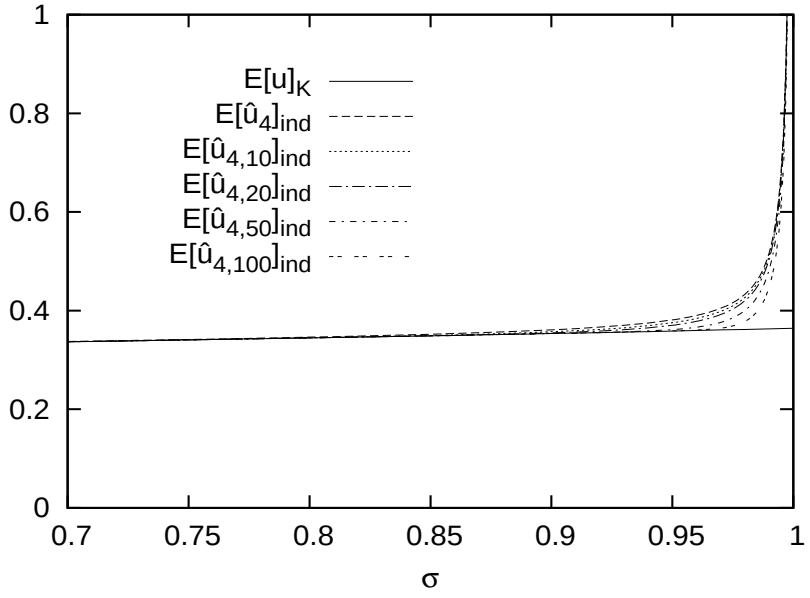

Figure 8: Mean system content: truncation approximation $E[u]_K$, simple polynomial approximation $E[\hat{u}_4]_{\text{ind}}$ and more involved polynomial approximation $E[\hat{u}_{4,m}]_{\text{ind}}$, versus the deadline-distribution parameter σ, for Poisson arrivals with $\lambda = 0.3$ and various values of m

6.1.3 Mean customer delay

The approximate mean customer delays $E\left[\hat{d}_3\right]_{\text{ind}}$ and $E\left[\hat{d}_4\right]_{\text{ind}}$ are compared with the "exact" values $E[d]_K$ in Fig. 9. Again, the accuracy of the approximation turns out to be quite good

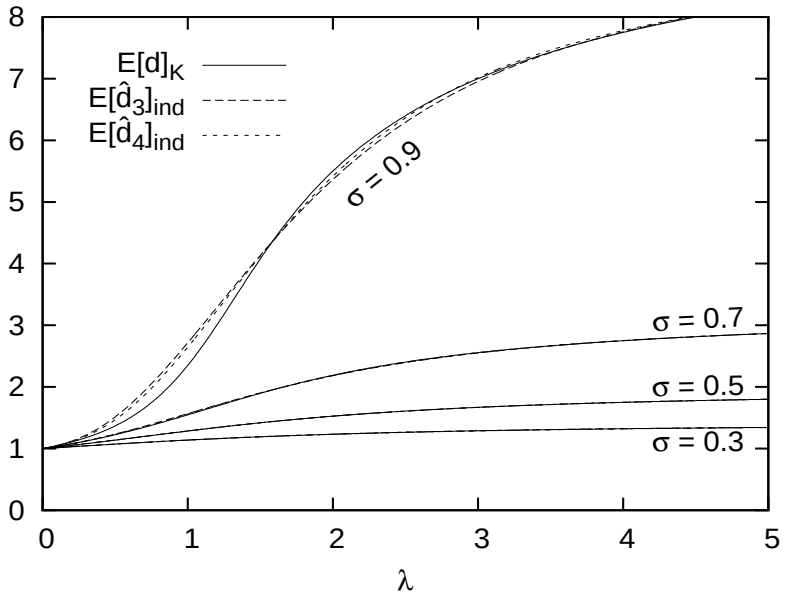

Figure 9: Mean customer delay: truncation approximation $E[d]_K$ and simple polynomial approximations $E\left[\hat{d}_3\right]_{\text{ind}}$ and $E\left[\hat{d}_4\right]_{\text{ind}}$, versus mean arrival rate λ, for Poisson arrivals and various values of the deadline-distribution parameter σ

for all λ, as long as σ is not too close to 1. As expected, the mean delay increases with the arrival rate λ. Also, the mean delay becomes smaller (for all relevant values of λ) when the mean deadline of the customers decreases. Specifically, for the four curves in Fig. 9, we have that the mean deadline $D = 1/(1-\sigma)$ takes values 10, 3.33, 2 and 1.42 for $\sigma = 0.9$, $\sigma = 0.7$, $\sigma = 0.5$, and $\sigma = 0.3$ respectively. It is very clear that the curves for the mean customer delay stay well below these mean deadlines, for all values of the arrival rate λ. Fig. 10 reconsiders the case where $\sigma = 0.9$; it illustrates that the accuracy of the indirect approximation, which in general can already be characterized as "quite good", can be further improved by using the somewhat more involved approximation $E\left[\hat{d}_{4,m}\right]_{\text{ind}}$ instead of $E\left[\hat{d}_4\right]_{\text{ind}}$.

6.1.4 Deadline-expiration ratio

Finally, some results for the deadline-expiration ratio are shown in Fig. 11. As the deadline-expiration ratio is computed directly from the probability of an empty system (see equations (30) and (40)), we expect the power-series approximation $\hat{r}_{ex,4}$ to be accurate as long as σ is not too high. This is indeed confirmed by the results in Fig. 11, which shows that even for a σ-value as high as 0.9 the polynomial approximation is very good. In Fig. 12, we illustrate that the accuracy can (again) be further improved by replacing $\hat{r}_{ex,4}$ by the somewhat more involved approximation $\hat{r}_{ex,4,m}$. Furthermore, the figure reveals that, for a given deadline-distribution parameter $\sigma < 1$, the fraction of customers that leave the queue unserved grows steadily with the arrival rate λ. An intuitive explanation of this observation is not so obvious, in view of the

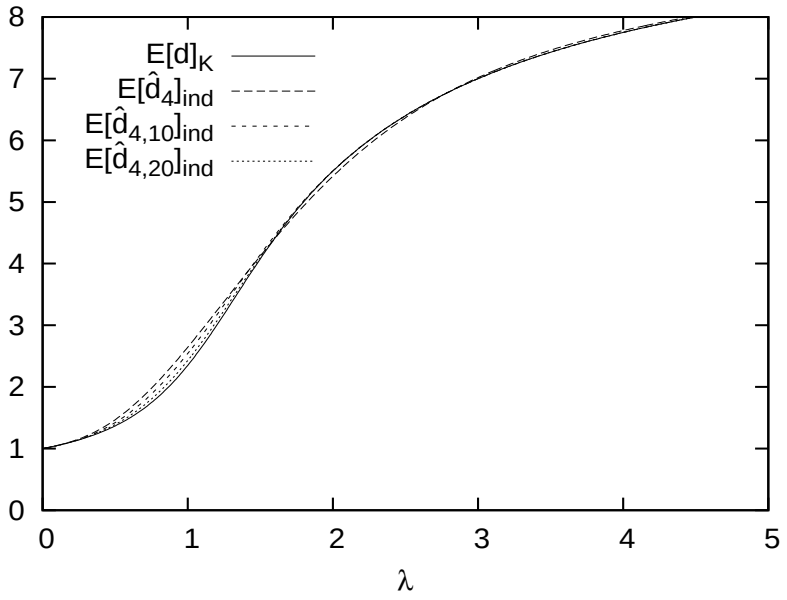

Figure 10: Mean customer delay: truncation approximation $E[d]_K$, simple polynomial approximation $E\left[\hat{d}_4\right]_{\text{ind}}$ and more involved polynomial approximation $E\left[\hat{d}_{4,m}\right]_{\text{ind}}$, versus mean arrival rate λ, for Poisson arrivals and $\sigma = 0.9$ and various values of m

fact that in a system without deadlines (i.e., $\sigma = 1$) the deadline-expiration ratio is constant and equal to zero for all values of λ, either smaller than 1 (stable system) or larger than 1 (instable system). In a system with deadlines ($\sigma < 1$), we expect the deadline-expiration ratio to increase with λ when $\lambda > 1$, because in this case, on average, at least $\lambda - 1$ customers per slot (i.e., at least a fraction $(\lambda - 1)/\lambda$) leave the system prematurely, since the server cannot handle more than 1 customer per slot. Perhaps more surprisingly, according to Fig. 11, the deadline-expiration ratio also grows with λ in the region $\lambda < 1$, which means that the fraction of customers that do get served before they leave the system decreases when λ increases. A possible explanation for this behavior lies in the fact that for increasing λ the length of the queue grows and customers are more likely to reach their deadline while waiting.

As a general conclusion on this series of numerical results, we may say that, in case of Poisson arrivals, even the simplest polynomial approximations (indexed with subscript 3 or 4) perform quite well as long as σ does not approach its "forbidden" value $\sigma = 1$. For high values of σ, the more involved approximations (indexed with subscript (4,m)), are preferable. Actually, this is the best we could hope for from the very beginning, because the whole polynomial approximation approach is based on the assumption that "high powers of σ become negligible". This, of course, requires σ not to be too close to 1.

6.2 Non-Poisson arrivals

The analysis performed in this paper is valid for all possible choices of the arrival pgf $A(z)$. In particular, it is by no means confined to Poisson arrivals only. In fact, from the point of

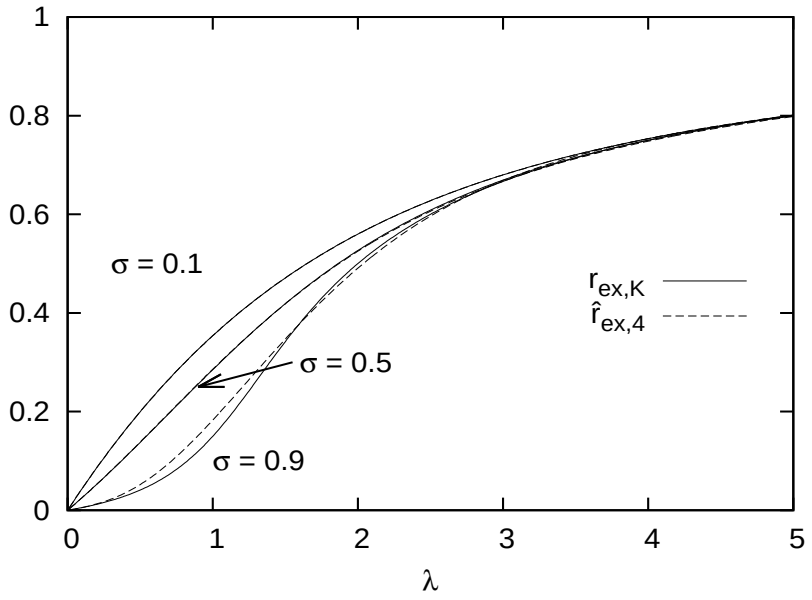

Figure 11: Deadline-expiration ratio: truncation approximation $\hat{r}_{ex,K}$ and simple polynomial approximation $\hat{r}_{ex,4}$, versus mean arrival rate λ, for Poisson arrivals and various values of the deadline-distribution parameter σ

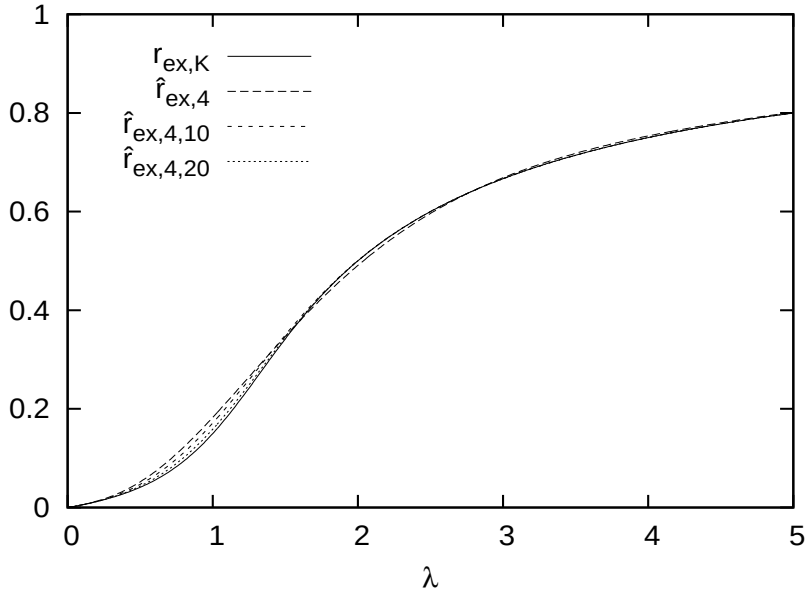

Figure 12: Deadline-expiration ratio: truncation approximation $\hat{r}_{ex,K}$, simple polynomial approximation $\hat{r}_{ex,4}$ and more involved polynomial approximation $\hat{r}_{ex,4,m}$, versus mean arrival rate λ, for Poisson arrivals and $\sigma = 0.9$ and various values of m

view of analytical simplicity, the Poisson case is even a more complicated case than many other choices due to the transcendental nature of the Poisson pgf $e^{\lambda(z-1)}$. In this subsection we briefly

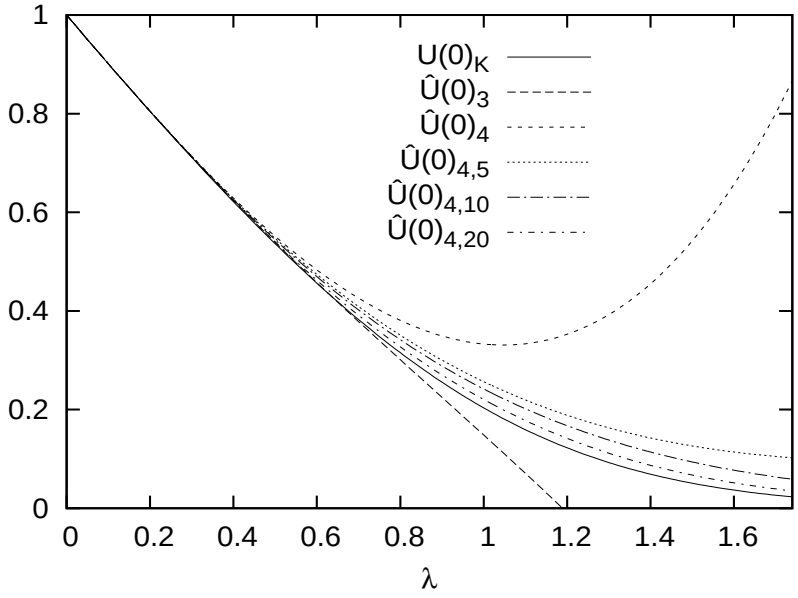

Figure 13: Probability of empty system: truncation approximation $U(0)_K$, simple polynomial approximations $\hat{U}(0)_3$ and $\hat{U}(0)_4$, and more involved polynomial approximations $\hat{U}(0)_{4,m}$, versus mean arrival rate λ, for geometric arrivals and $\sigma = 0.9$ and various values of m

comment on two other possible choices for $A(z)$, i.e., geometric arrivals and batch-Bernoulli arrivals respectively.

6.2.1 Geometric arrivals

In case of geometric arrivals with mean λ, the pmf $a(n)$ and the pgf $A(z)$ are given by

$$\begin{aligned} a(n) &= \frac{1}{1+\lambda}\left(\frac{\lambda}{1+\lambda}\right)^n \quad , \quad n \geq 0 \ , \\ A(z) &= \frac{1}{1+\lambda-\lambda z} \ . \end{aligned} \tag{43}$$

It then easily follows that $\alpha = \frac{1}{1+\lambda}$, $\beta = \frac{\lambda}{1+\lambda}$, $\gamma = \frac{\lambda^2}{1+\lambda}$, and the consecutive derivatives of $A(z)$ for $z = 1$ are

$$A'(1) = \lambda \ , \ A''(1) = 2\lambda^2 \ , \ A'''(1) = 6\lambda^3 \ , \ A^{(4)}(1) = 24\lambda^4 \ . \tag{44}$$

Some numerical results for geometric arrivals are presented in Figs. 13 and 14, assuming a "challenging" value for the deadline distribution parameter, i.e., $\sigma = 0.9$. Fig. 13 depicts the "exact" probability of empty system $U(0)_K$ along with its approximations $\hat{U}(0)_3$ and $\hat{U}(0)_4$ and $\hat{U}(0)_{4,m}$ for various values of m. The figure shows that the most simple polynomial approximations $\hat{U}(0)_3$ and $\hat{U}(0)_4$ are acceptable as long as λ stays below a value of about 0.5, but they

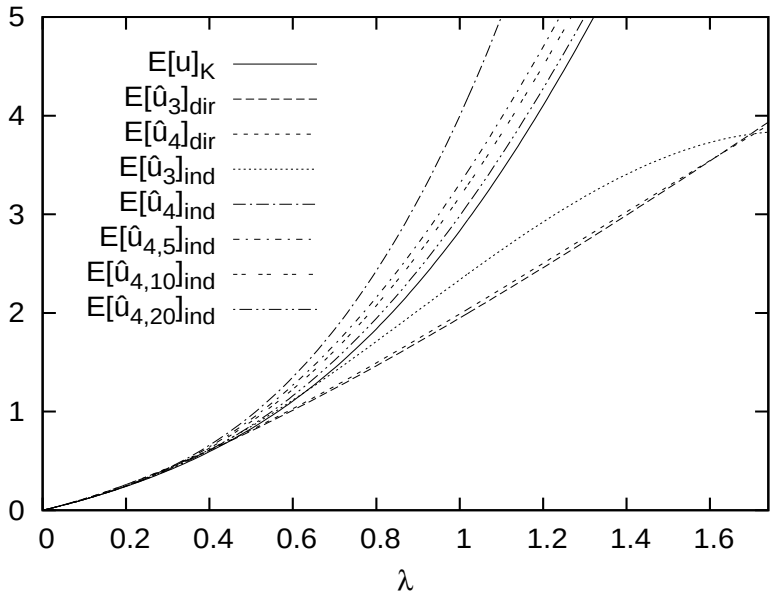

Figure 14: Mean system content: truncation approximation $E[u]_K$, direct polynomial approximations $E[\hat{u}_3]_{\rm dir}$ and $E[\hat{u}_4]_{\rm dir}$, indirect polynomial approximations $E[\hat{u}_3]_{\rm ind}$ and $E[\hat{u}_4]_{\rm ind}$, and more involved indirect polynomial approximation $E[\hat{u}_{4,m}]_{\rm ind}$, versus mean arrival rate λ, for geometric arrivals and $\sigma = 0.9$ and various values of m

fail completely for larger values of λ. This was not the case for Poisson arrivals - see Figs. 1 and 3. However, as soon as the more involved approximations of type $\hat{U}(0)_{4,m}$ are used, the approximations become increasingly close as m increases. As in the case of Poisson arrivals, a value of $m = 20$ seems to be sufficient for good accuracy. Very similar conclusions can be obtained from Fig. 14 with respect to the mean system content. Again, the simplest polynomial approximations $E[\hat{u}]_{\rm dir}$ and $E[\hat{u}]_{\rm ind}$ are not very convincing in case of geometric arrivals, but the more involved approximations, indexed with subscript $(4, m)$, are quite accurate; again, a good value of m seems to be 20.

6.2.2 Batch-Bernoulli arrivals

In this subsection, we assume that customers enter the system pairwise, i.e., two by two, according to a so-called batch-Bernoulli distribution with mean λ. This means that the pmf $a(n)$ and the pgf $A(z)$ are given by

$$\begin{aligned} &a(0) = 1 - \frac{\lambda}{2} \quad , \quad a(2) = \frac{\lambda}{2} \quad , \quad a(n) = 0 \text{ , if } n \neq 0 \text{ and } n \neq 2 \text{ ,} \\ &A(z) = 1 - \frac{\lambda}{2} + \frac{\lambda}{2} z^2 \ . \end{aligned} \tag{45}$$

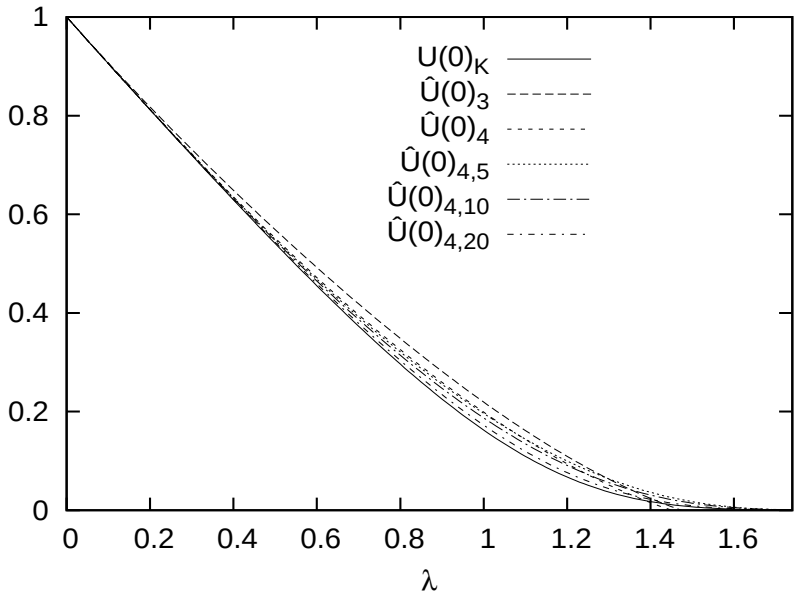

Figure 15: Probability of empty system: truncation approximation $U(0)_K$, simple polynomial approximations $\hat{U}(0)_3$ and $\hat{U}(0)_4$, and more involved polynomial approximations $\hat{U}(0)_{4,m}$, versus mean arrival rate λ, for batch-Bernoulli arrivals and $\sigma = 0.9$ and various values of m

It is clear that, in this case, $\lambda \leq 2$, $\alpha = 1 - \frac{\lambda}{2}$, $\beta = \frac{\lambda}{2}$, $\gamma = \frac{\lambda}{2}$, and the consecutive derivatives of $A(z)$ for $z = 1$ are

$$A'(1) = \lambda \ , \ A''(1) = \lambda \ , \ A'''(1) = 0 \ , \ A^{(4)}(1) = 0 \ . \tag{46}$$

The batch-Bernoulli case is illustrated by means of some numerical results in Figs. 15 and 16, again assuming $\sigma = 0.9$. Fig. 15 depicts the "exact" probability of empty system $U(0)_K$ along with its approximations $\hat{U}(0)_3$ and $\hat{U}(0)_4$ and $\hat{U}(0)_{4,m}$ for various values of m. The accuracy of the various approximations seems to be very similar as in the Poisson case: the most simple approximations $\hat{U}(0)_3$ and $\hat{U}(0)_4$ are already quite good for all values of λ, and can be even improved by using the more involved approximations of type $\hat{U}(0)_{4,m}$. Very similar conclusions can be obtained from Fig. 16 with respect to the mean system content. Again, among the simplest polynomial approximations $E[\hat{u}]_{\text{dir}}$ and $E[\hat{u}]_{\text{ind}}$, the indirect version is to be preferred, but the more involved approximations, indexed with subscript $(4, m)$, are more accurate.

6.2.3 Comparison

Having compared the numerical results that we have obtained for Poisson, geometric and batch-Bernoulli arrivals, as described above, along with many other numerical experiments not reported in this paper, we have come to the conviction that the most simple polynomial approximations (indexed with subscripts 3 or 4) are quite accurate for most values of the deadline-distribution parameter σ, as long as the successive moments of the arrival distribution, or, alternatively, the successive derivatives of the pgf $A(z)$ at $z = 1$, "do not become too

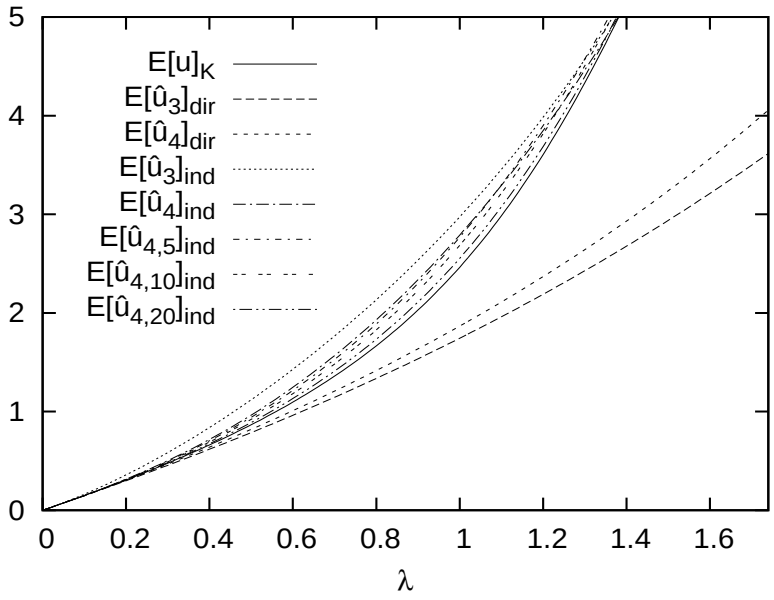

Figure 16: Mean system content: truncation approximation $E[u]_K$, direct polynomial approximations $E[\hat{u}_3]_{\text{dir}}$ and $E[\hat{u}_4]_{\text{dir}}$, indirect polynomial approximations $E[\hat{u}_3]_{\text{ind}}$ and $E[\hat{u}_4]_{\text{ind}}$, and more involved indirect polynomial approximation $E[\hat{u}_{4,m}]_{\text{ind}}$, versus mean arrival rate λ, for batch-Bernoulli arrivals and $\sigma = 0.9$ and various values of m

big" for $\lambda \approx 1$. An intuitive justification of this conjecture is based on the observation that the quantities $A^{(k)}(1)$ appear in the coefficients of the consecutive powers σ^k in the formulas for $\hat{U}(0)_3$ and $\hat{U}(0)_4$ as displayed in equations (35). In the case of Poisson arrivals (equation (42)), the $A^{(k)}(1)$s remain constant for $\lambda = 1$, for geometric arrivals (equation (44)), they increase with k and in the batch-Bernoulli case (equation (46)), they reduce to zero as k increases.

7 Conclusions and future work

This paper has examined a relatively simple model for a *discrete-time* single-server queueing system with *general* independent arrivals, in which customers are subjected to *deadlines*. We have been able to derive nearly exact but complicated formulas, as well as simpler approximate formulas for the main performance measures of the system. From the methodological point of view, we believe that one of the main contributions of our paper lies in the polynomial approximation method that we have developed in sections 4 and 5, a technique that may be useful in the solution of other queueing models that lead to hard-to-solve functional equations such as equation (7). In terms of numerical results, we have been able to explain most of the observed dependencies between performance measures and system parameters intuitively.

The main restriction of this work seems to be the assumption that the service times of the customers are *deterministically* equal to one slot each (although an extension to geometric service times seems straightforward) and that the deadlines of the customers are *geometrically distributed*. Future work will focus on generalizations of these assumptions.

Acknowledgement

This research has been co-funded by the Interuniversity Attraction Poles Programme initiated by the Belgian Science Policy Office.

References

[1] M. Armony and A. Mandelbaum. Routing and staffing in large-scale service systems: The case of homogeneous impatient customers and heterogeneous servers. *Operations Research*, 59(1):50–65, 2011.

[2] S. Benjaafar, J. Gayon, and S. Tepe. Optimal control of a production-inventory system with customer impatience. *Operations Research Letters*, 38(4):267–272, 2010.

[3] J. Blanc. On a numerical method for calculating state probabilities for queueing systems with more than one waiting line. *Journal of Computational and Applied Mathematics*, 20:119–125, 1987.

[4] J. Blanc. A numerical approach to cyclic-service queueing models. *Queueing Systems: Theory and Applications*, 6(1):173–188, 1990.

[5] J. Blanc. The power-series algorithm applied to the shortest-queue model. *Operations Research*, 40(1):157–167, 1992.

[6] J. Blanc. The power-series algorithm for polling systems with time limits. *Probability in the Engineering and Informational Sciences*, 12(2):221–237, 1998.

[7] T. Bonald and J. Roberts. Performance modeling of elastic traffic in overload. In *Proceedings of the 2001 ACM SIGMETRICS International Conference on Measurement and Modeling of Computer Systems*, pages 342–343, 2001.

[8] O. Boxma and P. de Waal. Multiserver queues with impatient customers. In *Proceedings of the 14th International Teletraffic Congress (ITC14)*, pages 743–756, 1994.

[9] R. Broekmeulen and K. van Donselaar. A heuristic to manage perishable inventory with batch ordering, positive lead-times, and time varying demand. *Computers & Operations Research*, 36(11):3013–3018, 2009.

[10] H. Bruneel and B. Kim. *Discrete-time models for communication systems including ATM.* Kluwer Academic, Boston, USA, 1993.

[11] K. De Turck, D. Fiems, S. Wittevrongel, and H. Bruneel. A Taylor series expansions approach to queues with train arrivals. In *Proceedings of the 5th International ICST Conference on Performance Evaluation Methodologies and Tools*, pages 447–455, 2011.

[12] M. Defraeye and I. Van Nieuwenhuyse. Controlling excessive waiting times in small service systems with time-varying demand: An extension of the ISA algorithm. *Decision Support Systems*, 54(4):1558–1567, 2013.

[13] A. Economou, S. Kapodistria, and J. Resing. The single server queue with synchronized services. *Stochastic Models*, 26(4):617–648, 2010.

[14] M. Feldman and J. Naor. Non-preemptive buffer management for latency sensitive packets. In *Proceedings of the 29th IEEE Conference on Computer Communications (IEEE INFOCOM 2010)*, pages 186–190, 2010.

[15] D. Fiems and H. Bruneel. A note on the discretization of Little's result. *Operations Research Letters*, 30(1):17–18, 2002.

[16] N. Gans, G. Koole, and A. Mandelbaum. Telephone call centers: A tutorial and literature review. Invited review paper. *Manufacturing and Service Operations Management*, 5(2):79–141, 2003.

[17] G. Hooghiemstra, M. Keane, and S. van de Ree. Power series for stationary distributions of coupled processor models. *SIAM Journal of Applied Mathematics*, 48(5):1159–1166, 1988.

[18] P. Hurley, J.-Y. Le Boudec, and M. Kara. ABE: Providing a low-delay service within best effort. *IEEE Network*, 15(3):60–69, 2001.

[19] S. Ioannidis, O. Jouini, A. Economopoulos, and V. Kouikoglou. Control policies for single-stage production systems with perishable inventory and customer impatience. *Annals of Operations Research*, 209(1):115–138, 2013.

[20] A. Jean-Marie and E. Hyon. Scheduling services in a queuing system with impatience and setup costs. *Computer Journal*, 55(5):553–563, 2012.

[21] O. Jouini. Analysis of a last come first served queueing system with customer abandonment. *Computers & Operations Research*, 39(12):3040–3045, 2012.

[22] O. Jouini, G. Koole, and A. Roubos. Performance indicators for call centers with impatient customers. *IIE Transactions*, 45(3):341–354, 2013.

[23] S. Kapodistria. The M/M/1 queue with synchronized abandonments. *Queueing Systems: Theory and Applications*, 68(1):79–109, 2011.

[24] J. Kim, B. Kim, and J. Kang. Discrete-time multiserver queue with impatient customers. *Electronics Letters*, 49(1):38–39, 2013.

[25] L. Kleinrock. *Queueing systems, part I.* Wiley, New York, USA, 1975.

[26] S. Krishnan and R. Sitaraman. Video stream quality impacts viewer behavior: Inferring causality using quasi-experimental designs. In *Proceedings of the 2012 ACM Internet Measurement Conference*, pages 211–224, 2012.

[27] R. Larson and K. Sasanuma. Congestion pricing: A parking queue model. *Journal of Industrial and Systems Engineering*, 4(1):1–17, 2010.

[28] F. Li. Competitive scheduling of packets with hard deadlines in a finite capacity queue. In *Proceedings of the 28th IEEE Conference on Computer Communications (IEEE INFOCOM 2009)*, pages 1062–1070, 2009.

[29] M. Li, T. Lin, and S. Cheng. Arrival process-controlled adaptive media playout with multiple thresholds for video streaming. *Multimedia Systems*, 18(5):391–407, 2012.

[30] R. Li and A. Eryilmaz. Scheduling for end-to-end deadline-constrained traffic with reliability requirements in multihop networks. *IEEE-ACM Transactions on Networking*, 20(5):1649–1662, 2012.

[31] C. Palm. Research on telephone traffic carried by full availability groups. *Tele*, 1:1–107, 1957.

[32] Y. Sakuma, A. Inoie, K. Kawanishi, and M. Miyazawa. Tail asymptotics for waiting time distribution of an M/M/s queue with general impatient time. *Journal of Industrial and Management Optimization*, 7(3):593–606, 2011.

[33] B. Steyaert, K. Laevens, D. De Vleeschauwer, and H. Bruneel. Analysis and design of a playout buffer for VBR streaming video. *Annals of Operations Research*, 162(1):159–169, 2008.

[34] J. Van Velthoven, B. Van Houdt, and C. Blondia. On the probability of abandonment in queues with limited sojourn and waiting times. *Operations Research Letters*, 34(3):333–338, May 2006.

[35] J. Walraevens, J. van Leeuwaarden, and O. Boxma. Power series approximations for two-class generalized processor sharing systems. *Queueing Systems: Theory and Applications*, 66(2):107–130, 2010.

[36] C.-C. Wu, K.-T. Chen, C.-Y. Huang, and C.-L. Lei. An empirical evaluation of VoIP playout buffer dimensioning in Skype, Google talk, and MSN Messenger. In *Proceedings of the 18th International Workshop on Network and Operating Systems Support for Digital Audio and Video*, pages 97–102, 2009.

[37] J. Wu, J. Wang, and Z. Liu. A discrete-time Geo/G/1 retrial queue with preferred and impatient customers. *Applied Mathematical Modelling*, 37(4):2552–2561, 2013.

[38] W. Xiong and T. Altiok. An approximation for multi-server queues with deterministic reneging times. *Annals of Operations Research*, 172(1):143–151, 2009.

[39] S. Zeltyn. *Call centers with impatient customers: Exact analysis and many-server asymptotics of the M/M/n+G queue*. PhD thesis, Technion, 2004.

[40] S. Zeltyn and A. Mandelbaum. Call centers with impatient customers: Many-server asymptotics of the M/M/n + G queue. *Queueing Systems: Theory and Applications*, 51(3-4):361–402, 2005.

[41] B. Zhang, J. van Leeuwaarden, and B. Zwart. Staffing call centers with impatient customers: Refinements to many-server asymptotics. *Operations Research*, 60(2):461–474, 2012.