Age-based Markovian approximation of the G/M/1 queue

Benjamin Legros

Version of Record: https://www.sciencedirect.com/science/article/pii/S0167637721001164
Manuscript_39bc9bf933bf2ba7e4796b6021ef28da

Age-based Markovian approximation of the G/M/1 queue

BenjaminLegros

Abstract

We extend the approach of [15] and [20] for the G/M/1 queue. The idea is to provide a Markovian approximation where a state represents the oldest customer's wait. This modeling is made possible by creating states with negative wait, representing an estimate of the time at which a new customer would arrive when the system is empty. We apply this method for performance evaluation and routing optimization. Finally, we further extend the model to the G/M/1+G queue.

Keywords: Approximation; G/M/1; Markov chain; Markov decision process; performance evaluation

1 Introduction

We propose a Markovian approximation of the G/M/1 queue, where a state of the system represents the age of the oldest customer. This provides an alternative way to determine the performance measures of G/M/1 related queues. It also allows investigating routing optimization issues via Markov Decision Processes (MDP) where decisions can be exercised as functions of the time spent in the system by a given customer. Time-based decisions can be more valuable than quantity-based ones when objective functions are non-linear in the time spent in the system. This is particularly the case if penalties have to be paid when reaching a certain time threshold in the system [19] or when percentile objectives are involved [17].

Specifically, we consider a first-come-first-served single server queue with infinite buffer capacity. The inter-arrival time is generally distributed and has a probability-density function, $f(t)$, for $t \geq 0$. The service time is exponentially distributed with service rate μ. We approximate the time spent by the oldest customer in the system - also called the First in Line (FIL) - by an Erlang distribution with rate γ and a random number of phases determined by the time at which the service ends. The idea was first proposed by [15] for an M/M/s queue and was later extended by [20] for an M/M/s+G queue. The aim of this note is to extend their model to the G/M/1 queue.

The representation in [15] and [20] cannot be applied to the G/M/1 queue. In these references, the state of the system x was either the waiting time of the oldest customer in the queue when $x > 0$ or else it determined the number of busy servers when $x \leq 0$. Therefore, for these references, the nature of the state description changed from a quantity when $x \leq 0$ to a time phase when $x > 0$. This was made possible due to the memoryless property of the inter-arrival time. With a generally distributed inter-arrival time, when the FIL leaves the system, we need to estimate the arrival phase of the next FIL. This one may already be

Benjamin Legros, Ecole de Management de Normandie, Laboratoire Métis, 64 Rue du Ranelagh, 75016 Paris, France, benjamin.legros@centraliens.net

present in the system or this customer may not have arrived yet. In the latter case, we need to determine how many phases will elapse before this customer arrives. To solve this problem, we propose an alternative definition of the system state where the time phase of the FIL x is also defined for $x \leq 0$ as a *negative wait* and we approximate the inter-arrival time by a phase-type distribution. In this way, in Section 2, we compute the transition probabilities and provide a one-dimensional Markovian approximation of the G/M/1 queue.

In Section 3, we use this Markovian representation to retrieve the performance measures of the G/M/1 queue and to optimize exclusion decisions. For performance evaluation, the Markov chain analysis of the G/M/1 queue also provides a relation between the expected idling and busy period durations which shows some connection with the Pollaczek-Khinchin formula. For exclusion optimization from a G/M/1 queue, we formulate a Markov decision process with the objective to obtain a trade-off between the rate of excluded customers and a time-cost function. We prove that the optimal exclusion policy is a time-out threshold policy and compute the optimal time-threshold for different distributions of the inter-arrival time. In particular, we observe that the optimal time-threshold may increase with the variability of the inter-arrival time.

In Section 4, we extend our modeling to the G/M/1+G queue. We assume that the patience time has a probability density function, denoted by $g(t)$, for $t \geq 0$. We provide a method to derive the transition probabilities leading the Markovian approximation of this queue to remain one-dimensional. Note that the proofs of the main results are provided in an online supplement.

The G/M/1 queue in the literature. The G/M/1 queue is one of the canonical models in queueing theory (e.g., see [13], Chapter 6). One way to derive the performance measures for this queue is to analyze the corresponding discrete time Markov chain at arrival instants. This approach is successful for computing the performance measures of G/M/1 related queues. For instance, [16] employed this method to analyze the finite buffer capacity G/M/1 queue, while [11] extended the analysis to the multi-server setting. Another example is [28] who investigated a queue for which the server stops serving the queue whenever the system becomes empty and resumes service when the number of waiting customers in the system reaches a certain threshold. Other approaches including martingale techniques, transform techniques, and sample-path arguments are presented in [1] to analyze the G/M/1 queue. As in this note, the authors analyzed the attained waiting time in the G/M/1 queue. [23] studied the busy period of the G/M/1 queue with restricted accessibility in connection with the M/G/1 queue using Laplace transforms. [12] employed the supplementary

variable technique [8, 24] where the remaining inter-arrival time plays the role of the supplementary variable to analyze the G/M/1 queue with a removable server. [7] computed the queue length probability-generating function when the server takes an exponential vacation each time the system empties. [4] developed level crossing arguments to investigate the G/M/1 queue with constant patience (i.e., the G/M/1+D queue). Our model is extended at the end of the note to account for customers' abandonment. Also using a sample path analysis and level crossing arguments, [22] analyzed the idle periods of the G/M/1 queue with a removable server. This note contributes to this literature by developing an alternative method to derive the performance measures of G/M/1 related queues and to perform policy optimization with time-based decisions.

2 Markovian approximation of the G/M/1 queue

In this section, we develop an approximation of the G/M/1 queue. First in Section 2.1, we show how the inter-arrival time can be approximated by a phase-type distribution. Next in Section 2.2, we explain how the G/M/1 queue can be approximated by a continuous time Markov chain.

2.1 Phase-type approximation of the inter-arrival time

In this section, we show how the general inter-arrival time with probability-density function $f(t)$, for $t \geq 0$ can be approximated by a particular phase-type distribution which will allow us to construct the Markovian representation of the G/M/1 queue. Let us denote by X the random variable representing the inter-arrival time. In Proposition 1, we approximate X by a random sum of exponential time phases each with rate γ. The number of time phases, R, is random and depends on the distribution of X. We define r_n as $r_n = P(R = n)$, for $n \in \mathbb{Z}^+$.

Proposition 1. *For $n \in \mathbb{Z}^+$, we have*

$$r_n = \frac{\gamma^n}{n!} \int_{t=0}^{\infty} t^n e^{-\gamma t} f(t) \mathrm{d}t. \tag{1}$$

As γ tends to infinity, the phase-type distribution $\sum_{i=1}^{R} Y_i$ converges in distribution to X, where the Y_i's are i.i.d. exponential distributions with rate γ.

We mention that the definition of r_n in Proposition 1 can be replaced by an equivalent expression which also replicates the behavior of the inter-arrival time as γ tends to infinity. The alternative given

in Proposition 2 provides simpler expressions for the transition probabilities for distributions having non-continuous behavior like the deterministic or uniform ones. Moreover, when writing the balance equations to obtain the stationary probabilities, only a finite number of probabilities has to be involved for these distributions. Nevertheless, for other distributions like the exponential one, this alternative definition may lead to more complicated expressions to manipulate than r_n.

Proposition 2. *We define* $\overline{r_n}$ *for* $n \in \mathbb{N}$, $\gamma \in \mathbb{R}$, *with* $n/\gamma = \tau$ *and* $\tau > 0$ *as* $\overline{r_n} = \int_{n/\gamma}^{(n+1)/\gamma} f(t)\mathrm{d}t$. *We have* $\lim_{n,\gamma \longrightarrow \infty} \frac{\overline{r_n}}{r_n} = 1$.

Examples for r_n.

- For the exponential distribution with rate λ, we have $f(t) = \lambda e^{-\lambda t}$. This leads to $r_n = \left(\frac{\lambda}{\lambda+\gamma}\right)\left(\frac{\gamma}{\lambda+\gamma}\right)^n$. Therefore, the distribution of the number of phases between two arrivals follows a geometric distribution. This is not surprising as the geometric distribution is the only discrete distribution with the memoryless property. The expression of r_n in this case is also the same as that obtained in [15] and [20].

- An extension of the exponential case is the Erlang case. For an Erlang distribution with k phases and rate β per phase, we have $f(t) = \frac{\beta^k t^{k-1} e^{-\beta t}}{(k-1)!}$. We then deduce that $r_n = \binom{n+k-1}{n}\left(\frac{\gamma}{\gamma+\beta}\right)^n\left(\frac{\beta}{\gamma+\beta}\right)^k$, where $\binom{n+k-1}{n} = \frac{(n+k-1)!}{n!(k-1)!}$. The distribution of the number of phases between two arrivals follows a binomial distribution.

- The deterministic distribution with parameter τ is defined with the Dirac function: $f(t) = \delta_\tau(t)$. Thus, we have $r_n = \frac{(\gamma\tau)^n}{n!}e^{-\gamma\tau}$ which corresponds to a Poisson distribution. Knowing that the Poisson distribution can be viewed as a limit of the binomial distribution, this result was expected from the Erlang case as the deterministic distribution can also be viewed as a limit of an Erlang distribution. Alternatively, the deterministic case can be obtained in a faster way by choosing γ such that $\gamma\tau = k$, with $\overline{r_k} = 1$ and $\overline{r_n} = 0$, for $n \neq k$.

2.2 Continuous time Markov chain for the G/M/1 queue

We approximate the age of the oldest customer in the system (i.e., the FIL) by an Erlang distribution with rate γ per phase. This means that the actual time spent by the FIL in the system is translated into a number of time phases x, for $x > 0$. The transition rate from time phase x to time phase $x + 1$ is γ. At service completion, the FIL is removed from the system. Due to the first-come-first-served discipline,

the new FIL, if present in the system, spent less time in the system than the FIL who just left. In the approximated model, this creates a transition from state x to a state $x-n$, for $n \geq 0$, where the transition probability, $p_{x,x-n}$, is determined by the inter-arrival time distribution. We approximate the inter-arrival time by the phase-type distribution defined in Section 2.1. Therefore, we have $p_{x,x-n} = r_n$. Note that the inter-arrival time does not depend on the time phase of the FIL. This explains why $p_{x,x-n}$ does not depend on x.

This presentation is valid only if there is a new FIL present in the system (i.e., if $x - n > 0$). However, we know the distribution of the number of the time phase separating two consecutive arrivals through r_n, $n \geq 0$. Hence, we can estimate the number of phases after which the new FIL will arrive in the system. This allows us to extend the support of the time phase of the FIL, x, to $x \leq 0$. Specifically, $x > 0$ represents the number of time phase of the FIL while $x \leq 0$ indicates that the FIL will arrive at an empty system after $1 - x$ time phases. In this way, we represent the G/M/1 queue by a one-dimensional continuous time Markov chain. The only difference between $x > 0$ and $x \leq 0$ is that service completion cannot happen when $x \leq 0$ as there is no customer in service. In summary, the two transitions in the Markov chain are the following:

1. A phase increase with rate γ with $x \in \mathbb{Z}$, which changes the state to $x + 1$. The time phase of the FIL is increased by 1.

2. A service completion with rate μ while the system is not empty (i.e., $x > 0$), which changes the state to $x - n$ with probability r_n for $n \geq 0$, that is, the new FIL is in time phase $x - n$.

Remark: In the approximated models of [15] and [20], states with negative time phase do not need to be defined as the inter-arrival time is exponentially distributed. Due to the memoryless property of the exponential distribution, a single state representing the empty system could be defined instead of providing the number of phases before which the new FIL will come.

Applicability and limitations. This approximation provides a Markovian representation of the G/M/1 queue where the state description is a discretized version of the oldest customer's wait. This representation is useful for performance evaluation when the service process depends on customers' wait. For example, the model studied in Section 3 in [5] where the service speed depends on the wait of the FIL could also be investigated with our approach. We could also study models with parallel queues [2, 6, 9] where a server prioritizes the service in one queue in function of the oldest customer's wait in each queue. In addition, our approach allows solving routing optimization problem using Markov decision processes. As a state of

the system represents either the time spent by the oldest customer or the remaining idling period, we can investigate problems with objectives involving the distributions of idling time or time spent in the system. For instance in call centers, it is common to have an objective of 80% of calls served in less than 20 seconds [10]. This type of objective cannot be captured with classical Markovian representation where a state of the system represents the number of customers.

However, our approach has limitations. First, it is limited to a single server analysis. The case with an infinite amount of servers could be treated as an abandonment feature but the case with a finite number of servers remains to be determined. One difficulty lies in relating the time spent by the oldest customer in the system and the number of busy servers. Second, the first-come-first-served discipline is essential to apply our approach. With other disciplines, a service completion would not necessarily lead to the first customer in the system having a lower time phase. Finally, the arrival process should not be state-dependent. For instance, we cannot investigate a finite capacity system with this approach.

3 Applications of the method

In this section, we show how the approximated model developed in Section 2 can be used for performance evaluation and policy optimization. In Section 3.1, we show how to retrieve the performance measures of the G/M/1 queue. Next in Section 3.2, we develop a Markov decision process approach for exclusion optimization from a G/M/1 queue.

3.1 Performance evaluation of the G/M/1 queue

We denote by π_x the stationary probability of being in state x, for $x \in \mathbb{Z}$. The system balance equations are as follows:

$$(\gamma + \mu)\pi_x = \gamma\pi_{x-1} + \mu\sum_{k=0}^{\infty} r_k \pi_{x+k}, \text{ for } x > 0, \text{ and,} \tag{2}$$

$$\gamma\pi_{-x} = \gamma\pi_{-x-1} + \mu\sum_{k=1}^{\infty} r_{x+k}\pi_k, \text{ for } x \geq 0. \tag{3}$$

In Theorem 1, we give the solutions of (2) and (3).

Theorem 1. *Under the stability condition $\mu E(A) > 1$, we have*

$$\pi_{-x} = \pi_0 \left(1 - \frac{\mu}{\gamma} A^*(\gamma(1-\sigma)) S_{x-1}(\sigma^{-1}) + \frac{\mu}{\gamma} \sum_{k=0}^{x-1} r_k S_{x-1-k}(\sigma^{-1}) \right), \text{ for } x \geq 0, \text{ and,} \tag{4}$$

$$\pi_x = \pi_0 \sigma^x, \text{ for } x \geq 0, \text{ with} \tag{5}$$

$$\pi_0 = \frac{1-\sigma}{\sigma \mu E(A)}, \text{ and} \tag{6}$$

where σ is the solution of

$$(\gamma + \mu)\sigma = \gamma + \mu\sigma A^*(\gamma(1-\sigma)), \tag{7}$$

not equal to 1, where $A^*(z)$ is the Laplace transform of function f, at point z, and $S_n(z) = \sum_{k=0}^{n} z^k = \frac{1-z^{n+1}}{1-z}$, for $z \neq 1$.

In Proposition 3, we deduce the probability of having an empty system P_0, the expected time spent by an arbitrary customer in the system $E(T)$, and the probability of staying in the system longer than $y > 0$ for an arbitrary customer $P(T > y)$. We also prove that these performance measures tend to the exact ones as γ tend to infinity. It is interesting to note that the approximated model directly gives the exact expression of P_0.

Proposition 3. *Under the stability condition $\mu E(A) > 1$, we have*

$$P_0 = 1 - \frac{1}{\mu E(A)}, \quad E(T) = \frac{1}{\gamma(1-\sigma)}, \text{ and, } P(T > y) = e^{-\gamma y(1-\sigma)}, \text{ for } y > 0. \tag{8}$$

These performance measures tend to the exact ones as γ tends to infinity.

We now investigate the relation between the idle and busy periods. To this end, we introduce the function $P(z) = \sum_{x=-\infty}^{+\infty} \pi_x e^{izx}$, where i is the complex number such that $i^2 = -1$. The function $P(z)$ is decomposed into $P(z) = P^+(z) + P^-(z)$, where $P^+(z) = \sum_{x=1}^{\infty} \pi_x e^{izx}$ and $P^-(z) = \sum_{x=0}^{\infty} \pi_{-x} e^{-izx}$. Using the functions $P^+(z)$ and $P^-(z)$, we relate the moments of the idling duration with those of the busy period. In what follows, we show how these functions can be used to relate the expected idling duration with the expected duration of the busy period. From (2) and (3), we deduce that

$$(\gamma + \mu)P^+(z) + \gamma P^-(z) = \gamma e^{iz}(P^+(z) + P^-(z)) + \mu P^+(z)R(z),$$

where $R(z) = \sum_{x=0}^{\infty} r_x e^{-izx}$. This equation can be rewritten as

$$P^-(z) = \frac{\mu(R(z)-1) - \gamma(1-e^{iz})}{\gamma(1-e^{iz})} P^+(z). \tag{9}$$

From this expression, after applying L'Hôpital's rule twice, we obtain $\frac{\partial P^-(z)}{\partial z}\Big|_{z=0}$. Since the expected duration of a phase time is $\frac{1}{\gamma}$, and the probability of an empty system is $1 - \frac{1}{\mu E(A)}$, the expected time for the next arrival given an empty system is $E(I) = \frac{1}{-i} \frac{\partial P^-(z)}{\partial z}\Big|_{z=0} \frac{1}{\gamma\left(1-\frac{1}{\mu E(A)}\right)}$. This leads to $E(I) = -\frac{1}{\gamma(1-\sigma)} + \frac{1}{\gamma\left(1-\frac{1}{\mu E(A)}\right)} + \frac{1}{2}\frac{E(A^2)}{E(A)\left(1-\frac{1}{\mu E(A)}\right)}$. When γ tends to infinity, we relate $E(I)$ and $E(T)$ via

$$E(I) + E(T) = \frac{E(A)\left(1+cv^2\right)}{2\left(1-\frac{1}{\mu E(A)}\right)}, \tag{10}$$

where cv is the ratio between the standard deviation and the mean of the inter-arrival time. This result shows the equivalent of the Pollaczek-Khinchin formula for the G/M/1 queue.

3.2 Exclusion optimization from a G/M/1 queue

In this section, we investigate a problem of exclusion optimization from a G/M/1 queue. For this problem, a controller can decide to exclude the oldest customer in the system at any point in time. The objective is to minimize the long-run cost function defined as a linear combination of the rate of excluded customers and a time-based performance measure like the expected time spent in the system or a percentile of the time spent in the system. Exclusion optimization in this context differs from the rejection policies studied in the academic literature where customers are rejected upon arrival [3, 14, 26, 27]. Rejecting customers at arrival is optimal when the time-based performance measure is the expected wait or the expected time spent in the system by both served and rejected customers. However, when considering the expected time spent in the system by served customers only or a wait percentile, rejecting customers after letting them wait is preferred as compared to rejection at arrival [18, 21].

To solve the optimization problem, we formulate a Markov decision process and next use the value iteration technique to prove the form of the optimal policy. For the Markovian process defined in Section 2, the maximal event rate $\mu + \gamma$ is bounded. Therefore, we apply the uniformization technique [25]. By replacing the transition rates μ and γ by the corresponding transition probabilities $\frac{\mu}{\mu+\gamma}$ and $\frac{\gamma}{\mu+\gamma}$, our continuous time MDP can be investigated as a discrete time one. In states $x > 0$ after a $\gamma-$transition, the controller can decide to reject the FIL from the system. In such case, a penalty γP is counted to capture the

rate of excluded customers. In addition, a time-cost function $c(x)$ is defined for states $x > 0$. For instance with $c(x) = c\frac{x}{\gamma}$, the cost function translates the expected time spent in the system as the expected time phase is $\frac{1}{\gamma}$. With $c(x) = c\frac{(x-m)^+}{\gamma}$ or $c(x) = c\mathbb{1}_{x \geq m}$ with $\frac{m}{\gamma} = \tau$, the cost function expresses the expected excess $E((T-\tau)^+)$ or the percentile of the wait $P(T > \tau)$. Note that we could distinguish whether a customer is served or excluded in the cost function by having a cost $c_1(x)$ after a $\mu-$transition and $c_2(x)$ after a $\gamma-$transition in case of customer exclusion.

We then define the dynamic programming value function $V_k(x)$ over k steps for $x \in \mathbb{Z}$ by $V_0(x) = 0$,

$$V_{k+1}(x) = \frac{\gamma}{\gamma+\mu} V_k(x+1) + \frac{\mu}{\gamma+\mu} V_k(x) \text{ for } x \leq 0, \text{ and} \tag{11}$$

$$V_{k+1}(x) = c(x) + \frac{\gamma}{\gamma+\mu} \min(V_k(x+1), F(V_k(x)) + \gamma P) + \frac{\mu}{\gamma+\mu} F(V_k(x)) \text{ for } x > 0, \tag{12}$$

where the operator F applied on a function $f(x)$ for $x > 0$ results in $F(f(x)) = \sum_{n=0}^{\infty} r_n f(x-n)$. As k tends to infinity, the difference $V_{k+1} - V_k$ converges to the long-run optimal cost and allows us to identify the optimal policy [25].

We are interested in the long-run optimal decision in each state. The classical result in such one-dimensional control problem is to prove that the optimal policy is of threshold type. With a minimizing operator, this usually consists of showing by induction that the value function is convex. Here, the transition structure leads to a different structural property. For a given induction step k, proving a threshold policy consists of showing that if it is optimal to reject a customer from phase x, then it is also optimal to reject a customer from phase $x+1$. Therefore, if $V_k(x+1) \geq F(V_k(x)) + \gamma P$ (rejection from state x), then we should have $V_k(x+2) \geq F(V_k(x+1)) + \gamma P$ (rejection from state $x+1$). This implication holds if $V_k(x+2) + F(V_k(x)) - F(V_k(x+1)) - V_k(x+1) \geq 0$, for $x > 0$. We then define Property (13) for a function $z(x)$ as

$$z(x+2) + F(z(x)) - F(z(x+1)) - z(x+1) \geq 0, \tag{13}$$

for $x \geq 0$. Note that Property (13) is less restrictive than the convexity property. If a function is convex then this function satisfies (13).

In Theorem 2, we prove the threshold form of the optimal policy by showing the propagation of (13) by induction on k for a process that behaves asymptotically like V_k when k tends to infinity. To prove this result, we assume that Property (13) holds for $c(x)$. Numerically, we observe that the threshold form of

the optimal policy remains when $c(x)$ is increasing in x without satisfying (13) (for instance when $c(x)$ is a step function).

Theorem 2. *Assuming that (13) holds for $c(x)$, the optimal long-run exclusion policy is of threshold type. This means that there exists a time phase $n > 1$ such that a customer is excluded from the system after a $\gamma-$transition from time phase $n-1$.*

Numerical illustration. In Table 1, we illustrate how the optimal exclusion policy can be computed using (11) and (12) with $c(x) = \frac{x}{\gamma}$. We consider the deterministic, the exponential and the hyper-exponential distributions where the expected inter-arrival time is equal to 1. The hyper-exponential distribution is defined with two rates, 5 and 5/9, and a probability of 50% to have an inter-arrival time exponentially distributed with rate 5. For the computation, the state space needs to be bounded. We introduce a bound D such that the state space is defined for $-D+1 \leq x \leq D$. At state $x = D$, we force customer's exclusion after a $\gamma-$transition and count an exclusion penalty. The parameter D is chosen such that increasing D does not affect the optimal policy. In our examples, we vary γ from 1 to 80 and select $D = 1000$. For each distribution, we estimate the optimal cost g^* by recursively evaluating V_k until $\max_{-D+1 \leq x \leq D} ||V_{k+2}(x) - V_{k+1}(x)| - |V_{k+1}(x) - V_k(x)|| \leq 10^{-6}$. For each distribution, we also indicate the optimal threshold state n^* such that a $\gamma-$transition from state $n^* - 1$ results in the FIL being excluded. The optimal time t^* at which exclusion should be operated is estimated via $t^* = \frac{n^*}{\gamma}$.

Table 1: Evaluation of the optimal policy ($\mu = 1$, $E(A) = 1$, $c(x) = \frac{x}{\gamma}$, $P = 10$, $D = 1000$)

	Deterministic			Exponential			Hyper-exponential		
γ	n^*	t^*	g^*	n^*	t^*	g^*	n^*	t^*	g^*
1	2	2.000	2.0000	2	2.000	2.0000	2	2.000	2.0000
5	13	2.600	2.5868	15	3.000	3.1270	16	3.200	3.5861
10	28	2.800	2.6544	33	3.300	3.3411	37	3.700	3.9200
20	57	2.850	2.6831	69	3.450	3.4581	78	3.900	4.1072
30	85	2.833	2.6914	104	3.467	3.4987	120	4.000	4.1731
40	114	2.850	2.6953	140	3.500	3.5193	162	4.050	4.2067
50	143	2.860	2.6975	176	3.520	3.5318	203	4.060	4.2262
60	172	2.867	2.6988	212	3.533	3.5402	245	4.083	4.2271
70	200	2.857	2.7000	248	3.543	3.5459	287	4.100	4.2285
80	229	2.863	2.7000	284	3.550	3.5496	328	4.100	4.2252

For each distribution, we observe that t^* tends to increase with γ. This can be explained by two reasons. First, in our approximation the elapsing of time spent by the FIL in the system is represented by an Erlang distribution. As γ increases, the variability of this distribution reduces. Thus, there is a better control of the time spent by customers in the system when γ increases. This allows the controller to let customers stay longer. The second reason is that the cost of early exclusion tends to increase with γ. For instance with

our parameter values and a rejection threshold $n^* = 2$, the system cost is $\frac{\frac{1}{\gamma}+\frac{\gamma^2 P}{1+\gamma}}{2+\gamma}$. This function increases with γ.

As expected, the optimal cost g^* increases with the variability of the inter-arrival time since the system's congestion increases with the variability of the inter-arrival time. We could expect the exclusion time t^* to reduce with the variability of the inter-arrival time as a way to reduce the system's congestion. However, the opposite phenomenon is observed. When the variability of the inter-arrival time increases, customers tend to arrive in batch. Thus, a low exclusion time leads to more exclusion if the variability of the inter-arrival time is high. To avoid a too high rate of excluded customers, it is then preferred to increase the exclusion time (at the cost of increasing the expected time spent in the system).

4 Extension of the model for the G/M/1+G queue

We now present how the model developed in the previous sections can be extended to the G/M/1+G queue. We denote by t_x the probability of abandoning the queue before x phases of wait. Probability $1-t_x$ represents the probability of surviving after x phases of wait. We denote the random variable representing the patience time by Z. We can thus write $t_x = P(Z < Y_1 + Y_2 + \ldots + Y_{x+1}) = \int_{t=0}^{\infty} \sum_{i=0}^{x} e^{-\gamma t} \frac{(\gamma t)^i}{i!} g(t) \mathrm{d}t$. Note that this expression is only valid if $x \geq 1$. For $x < 1$, abandonment cannot occur as customers have not yet arrived. As in Section 1, we are interested in determining the transition probabilities $p_{x,x-k}$, for $x \geq 1$, and $k \geq 0$.

Consider a customer at phase $x > 0$. Probability $p_{x,x}$ is the probability that there is at least one customer present at time phase x when the FIL leaves the system. The number of arrivals at the same time phase, N_0, is geometrically distributed with parameter r_0: $P(N_0 = n_0) = r_0^{n_0}(1-r_0)$. The probability that time phase x is empty is the probability that all customers who arrive at this time phase abandon the system. Thus, we have $1 - p_{x,x} = \sum_{n_0=0}^{\infty} (1-r_0) r_0^{n_0} t_x^{n_0} = \frac{1-r_0}{1-r_0 t_x}$, and we deduce that $p_{x,x} = \frac{r_0(1-t_x)}{1-r_0 t_x}$.

We now look at computing $p_{x,x-k}$, for $x > 0$ and $k < x$. We denote by $N_0, N_1, \ldots, N_k$ the number of arrivals at phases $x, x-1, \ldots, x-k$, respectively. The expression of $P(N_0 = n_0, N_1 = n_1, \ldots, N_k = n_k)$ depends on whether $n_1, n_2, \ldots, n_k$ are strictly positive or equal to zero. If $n_j > 0$, then the term $r_0^{n_j - 1}$ becomes part of the expression as some customers are in phase $x-j$. If we have $n_{j+1} = n_{j+2} = \ldots = n_{m-1} = 0$, with $m > j$, $n_j > 0$ and $n_m > 0$, then the term r_{m-j} is generated. If we have $n_{j+1} = n_{j+2} = \ldots = n_k = 0$, with $n_j > 0$ then the term $1 - r_0 - r_1 - \ldots r_{k-j}$ is involved. In order to have a common expression for all possible expressions of $P(N_0 = n_0, N_1 = n_1, \ldots, N_k = n_k)$, we introduce the functions $\delta(j)$ and $d(j)$

defined by

$$\delta(j) = \begin{cases} 1, \text{ if } j > 0, \text{ and,} \\ 0, \text{ if } j = 0, \end{cases} \tag{14}$$

and

$$d(j) = \begin{cases} \max\{i \in \{1, \ldots, j-1\} : \delta(n_i) = 1\}, \text{ if } j > 1, \text{ and,} \\ 0, \text{ if } j = 1, \end{cases} \tag{15}$$

with the convention $\max\{i \in \{1, \ldots, j-1\} : \delta(n_i) = 1\} = 0$, if $\delta(n_1) = \delta(n_2) = \ldots = \delta(n_{j-1}) = 0$. We thus write

$$P(N_0 = n_0, N_1 = n_1, \ldots, N_k = n_k) = r_0^{n_0} \left(\prod_{j=1}^{k} \left(r_{j-d(j)} \cdot r_0^{n_j - 1} \right)^{\delta(n_j)} \right) \left(1 - \sum_{j=0}^{k-d(k+1)} r_j \right). \tag{16}$$

For a given time phase $x - j$, the probability that n_j customers would have abandoned the system is $t_{x-j}^{n_j}$, if $x - j > 0$. If $x - j \leq 0$, the only possibility to have an empty phase $x - j$ is not to have any arrival at phase $x - j$. Thus, we have

$$1 - p_{x,x} - \ldots - p_{x,x-k} = \sum_{n_0, n_1, \ldots, n_k = 0}^{\infty} \left(r_0^{n_0} t_x^{n_0} \left(\prod_{j=1}^{k} \left(r_{j-d(j)} \cdot r_0^{n_j - 1} t_{x-j}^{n_j} \right)^{\delta(n_j)} \right) \left(1 - \sum_{j=0}^{k-d(k+1)} r_j \right) \right),$$

if $k < x$, and

$$1 - p_{x,x} - \ldots - p_{x,x-k} = \sum_{n_0, n_1, \ldots, n_{x-1} = 0}^{\infty} \left(r_0^{n_0} t_x^{n_0} \left(\prod_{j=1}^{x-1} \left(r_{j-d(j)} \cdot r_0^{n_j - 1} t_{x-j}^{n_j} \right)^{\delta(n_j)} \right) \left(1 - \sum_{j=0}^{k-d(x)} r_j \right) \right),$$

if $k \geq x$. The expression of $p_{x,x-k}$ can then be deduced explicitly from these expressions.

The direct computation of the above expression may be difficult. To facilitate the computation, we provide an alternative way of computing $p_{x,x-k}$ in Proposition 4.

Proposition 4. *For $x > 0$, we have*

$$p_{x,x-k} = \frac{T_x}{t_x} \sum_{b_1,b_2\ldots,b_{k-1}\in\{0,1\}} \frac{(1-(1-r_0)T_{x-k})\, a_{b_1,b_2,\ldots,b_{k-1}} \cdot r_{k-\max\{i\in\{1,k-1\}:b_i=1\}}}{1-r_0-\ldots-r_{k-1-\max\{i\in\{1,k-1\}:b_i=1\}}} T_{x-1}^{b_1} T_{x-2}^{b_2} \ldots T_{x-(k-1)}^{b_{k-1}}, \tag{17}$$

for $k < x$, and,

$$p_{x,x-k} = \frac{T_x}{t_x} \sum_{\substack{b_1,b_2\ldots,b_{x-1}\in\{0,1\},\\ b_x=b_{x+1}=\ldots=b_{k-1}=0}} \frac{a_{b_1,b_2,\ldots,b_{k-1}} \cdot r_{k-\max\{i\in\{1,k-1\}/b_i=1\}}}{1-r_0-\ldots-r_{k-1-\max\{i\in\{1,k-1\}/b_i=1\}}} T_{x-1}^{b_1} T_{x-2}^{b_2} \ldots T_1^{b_{x-1}}, \text{ for } k \geq x, \tag{18}$$

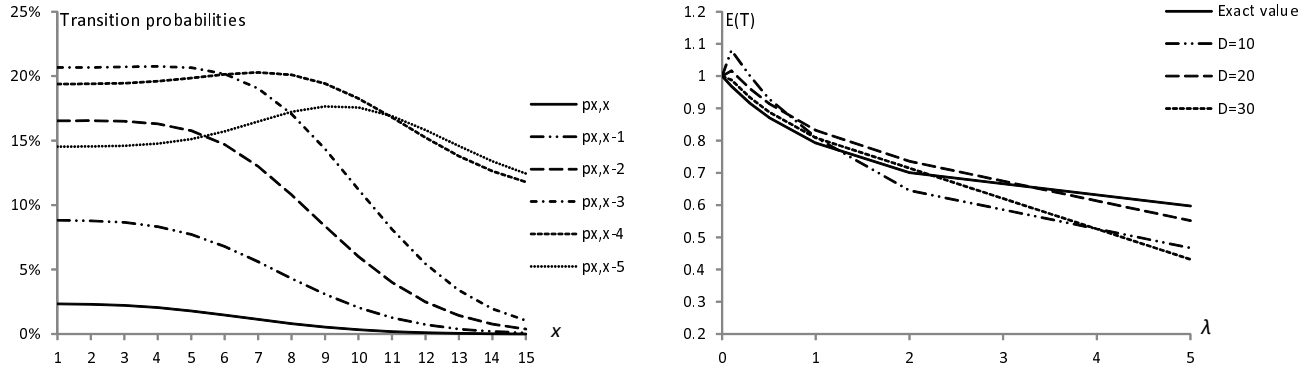
with $T_{x-j} = \frac{t_{x-j}}{1-r_0 t_{x-j}}$ for $0 \leq j < x$, and

$$a_{b_1,b_2,\ldots,b_k,1} = a_{b_1,b_2,\ldots,b_k} \frac{(1-r_0) r_{k+1-\max\{i\in\{1,k\}:b_i=1\}}}{1-r_0-\ldots-r_{k-\max\{i\in\{1,k\}:b_i=1\}}}, \text{ and,}$$
$$a_{b_1,b_2,\ldots,b_k,0} = a_{b_1,b_2,\ldots,b_k} \frac{1-r_0-\ldots-r_{k+1-\max\{i\in\{1,k\}:b_i=1\}}}{1-r_0-\ldots-r_{k-\max\{i\in\{1,k\}:b_i=1\}}},$$

with the convention $\max\{i \in \{1,k\} : b_i = 1\} = 0$, *if* $b_1 = b_2 = \ldots = b_k = 0$, $a_1 = r_1(1-r_0)$, *and* $a_0 = 1 - r_0 - r_1$.

Numerical illustration. In Figure 1(a), we give the transition probabilities $p_{x,x-k}$ for $x = 1, 2, \ldots, 15$, $k = 0, 1, \ldots, 5$ and $\gamma = 5$ in the case of a deterministic inter-arrival time with $E(A) = 0.75$ and a deterministic abandonment with expected patience time $E(B) = 1.5$. We observe that for small values of x, $p_{x,x-k}$ is close to r_k which corresponds to the transition probability without abandonment. When the FIL has a small time phase, abandonment cannot have a significant effect since most customers stayed in the system a shorter duration than their patience time. For high values of x and small values of k, the transition probability $p_{x,x-k}$ tends to zero. In this case, abandonment has removed customers with long wait from the system.

Although appealing for performance evaluation and policy optimization, the computation of the transition probabilities $p_{x,x-k}$ using (17) requires the summation of 2^{k-1} terms which corresponds to an exponential complexity. This precludes implementing this approach when γ is high and when the state space is large. One direction for future research is to approximate the transition probabilities $p_{x,x-k}$ to reduce the complexity of the computation. For instance, when γ increases we can show that r_k and T_k tend to zero



(a) Transition probabilities deterministic case ($E(A) = 0.75$, $E(B) = 1.5$, $\gamma = 5$)

(b) Approximation of $E(T)$ in the M/M/1+M queue ($\beta = 2$, $\mu = 1$, $D = 10\gamma$)

Figure 1: Numerical illustration

for $k \geq 0$. This suggests to approximate the transition probabilities $p_{x,x-k}$ by

$$1 - p_{x,x} - p_{x,x-1} - \ldots - p_{x,x-k} \simeq \frac{T_x}{t_x} \left(R_k + \sum_{i=1}^{k} r_i T_{x-i} R_{x-i} \right),$$

where $R_j = 1 - r_0 - \ldots - r_j$ for $j \geq 0$ with the convention $T_j = 0$ if $j \leq 0$. In this approximation, the terms involving products of the type $r_k r_j T_{x-k} T_{x-j}$ are removed. Thus, instead of having an exponential complexity, this approximation provides a linear one. In Figure 1(b) we evaluate the expected time spent in the system $E(T)$ for an M/M/1+M queue from a Markov chain analysis using this approximation as a function of the arrival rate λ. In this example, the abandonment time is exponentially distributed with rate $\beta = 2$. We truncate the state space with D such that the admissible states are $x = -D+1, -D+2, \ldots, D$ and we relate D and γ via $\frac{D}{\gamma} = 10$ (i.e., the maximal expected time spent in the system is 10 time units). We observe that the approximation provides a good estimation of $E(T)$ when the arrival rate is low as it corresponds to a situation with a small proportion of abandonment. When λ increases the approximation diverges from the exact model. In the approximation, the transition probabilities are overestimated which reduces $E(T)$ for large λ as compared to the exact value. Having a large value of D may also increase the inaccuracy of the approximation when λ is large as more transitions are involved.

References

[1] I. Adan, O. Boxma, and D. Perry. The G/M/1 queue revisited. *Mathematical Methods of Operations Research*, 62(3):437–452, 2005.

[2] H. Ahn and M. Lewis. Flexible server allocation and customer routing policies for two parallel queues

when service rates are not additive. *Operations Research*, 61(2):344–358, 2013.

[3] O.Z. Akşin, F. De Véricourt, and F. Karaesmen. Call center outsourcing contract analysis and choice. *Management Science*, 54(2):354–368, 2008.

[4] J. Bae and S. Kim. The stationary workload of the G/M/1 queue with impatient customers. *Queueing Systems*, 64(3):253–265, 2010.

[5] R. Bekker, G. Koole, B. Nielsen, and T. Nielsen. Queues with waiting time dependent service. *Queueing Systems*, 68(1):61–78, 2011.

[6] C. Buyukkoc, P. Varaiya, and J. Walrand. The cμ rule revisited. *Advances in Applied Probability*, 17(1):237–238, 1985.

[7] K. Chae, S. Lee, and H. Lee. On stochastic decomposition in the GI/M/1 queue with single exponential vacation. *Operations Research Letters*, 34(6):706–712, 2006.

[8] J. Cohen and A. Browne. *The single server queue*, volume 8. North-Holland Amsterdam, 1982.

[9] C. Derman, G. Lieberman, and S. Ross. On the optimal assignment of servers and a repairman. *Journal of Applied Probability*, pages 577–581, 1980.

[10] N. Gans, G. Koole, and A. Mandelbaum. Telephone call centers: Tutorial, review, and research prospects. *Manufacturing & Service Operations Management*, 5(2):79–141, 2003.

[11] P. Hokstad. The G/M/m queue with finite waiting room. *Journal of Applied Probability*, 12(4):779–792, 1975.

[12] J. Ke and K. Wang. A recursive method for the N policy G/M/1 queueing system with finite capacity. *European Journal of Operational Research*, 142(3):577–594, 2002.

[13] L. Kleinrock. *Queueing Systems, Theory*, volume I. A Wiley-Interscience Publication, 1975.

[14] Y. Koçağa and A. Ward. Admission control for a multi-server queue with abandonment. *Queueing Systems*, 65(3):275–323, 2010.

[15] G. Koole, B. Nielsen, and T. Nielsen. First in line waiting times as a tool for analysing queueing systems. *Operations Research*, 60(5):1258–1266, 2012.

[16] G. Laslett. Characterising the finite capacity GI/M/1 queue with renewal output. *Management Science*, 22(1):106–110, 1975.

[17] B. Legros. Unintended consequences of optimizing a queue discipline for a service level defined by a percentile of the waiting time. *Operations Research Letters*, 44(6):839–845, 2016.

[18] B. Legros. Late-rejection, a strategy to perform an overflow policy. *European Journal of Operational Research*, 281(1):66–76, 2020.

[19] B. Legros, Y. Bouchery, and J. Fransoo. A time-based policy for empty container management by consignees. *Production and Operations Management*, 28(6):1503–1527, 2019.

[20] B. Legros, O. Jouini, and G. Koole. A uniformization approach for the dynamic control of queueing systems with abandonments. *Operations Research*, 66(1):200–209, 2018.

[21] B. Legros, O. Jouini, and G. Koole. Should we wait before outsourcing? Analysis of a revenue-generating blended contact center. *Manufacturing & Service Operations Management*, 2021.

[22] A. Löpker and D. Perry. The idle period of the finite G/M/1 queue with an interpretation in risk theory. *Queueing Systems*, 64(4):395–407, 2010.

[23] D. Perry, W. Stadje, and S. Zacks. Busy period analysis for M/G/1 and G/M/1 type queues with restricted accessibility. *Operations Research Letters*, 27(4):163–174, 2000.

[24] N. Prabhu. Queues and inventories., 1965.

[25] M. Puterman. *Markov Decision Processes*. John Wiley and Sons, 1994.

[26] Z. Ren and Y. Zhou. Call center outsourcing: Coordinating staffing level and service quality. *Management Science*, 54(2):369–383, 2008.

[27] J. Schrieck, Z. Akşin, and P. Chevalier. Peakedness-based staffing for call center outsourcing. *Production and Operations Management*, 23(3):504–524, 2014.

[28] Z. Zhang and N. Tian. The N threshold policy for the GI/M/1 queue. *Operations Research Letters*, 32(1):77–84, 2004.