Submitted to *Operations Research*
manuscript (Please, provide the manuscript number!)

A Uniformization Approach for the Dynamic Control of Queueing Systems with Abandonments

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

Oualid Jouini
CentraleSupélec, Université Paris-Saclay, Laboratoire Genie Industriel, Grande Voie des Vignes, 92290 Chatenay-Malabry, France, oualid.jouini@centralesupelec.fr

Ger Koole
VU University Amsterdam, Department of Mathematics, De Boelelaan 1081a, 1081 HV Amsterdam, The Netherlands, ger.koole@vu.nl

We consider queueing systems with general abandonment. Abandonment times are approximated by a particular Cox distribution with all phase exponential rates being the same. We prove that this distribution arbitrarily closely approximate any non-negative distribution. By explicitly modeling the waiting time of the first customer in line, we obtain a natural bounded jump Markov process allowing for uniformization. This approach is useful to solve, via dynamic programming, various optimization problems where the objectives and/or constraints involve the distributions of the performance measures, not only their expected values. It is also useful for the performance analysis of queueing systems with general abandonment times.

Key words: queueing systems, Markov chains, dynamic programming, uniformization, scheduling, optimization, Markov decision process, Cox distribution, general abandonments

History: This paper was first submitted on July 8, 2014 and has been with the authors for 9 months for 3 revisions.

1. Introduction

Existing applications of Markov decision processes fail in addressing the dynamic control questions for queueing models with abandonments. One reason is that the available approaches require uniformization (Down et al. 2011), while in the considered queueing models, the jump rates are generally unbounded functions of actions and states. To overcome the limitations of the standard techniques, Bhulai et al. (2014) propose a method that modifies the system rates by linearly smoothing them. The value of this approximation is that the convexity properties of the operators in Markov decision problems are maintained on the boundaries whereas they are not with a simple truncation. However, this method only works with exponential patience distributions and control decisions that are based on the number of customers in the queue.

In numerous optimization problems, the objectives or the constraints are defined through the waiting time distribution and not its expected value (Legros 2016). A minimum service level of 80% of customers served in less than 20 seconds is common in call centers, also, a minimum service level of 90% of patients served in less than four hours is used in emergency departments. A percentile of the waiting time is in general preferred to its expectation because the former is perceived to be more informative (Bailey and Sweeney 2003). The expectation does not take into account for instance the variability of the waiting time. Such settings require the use of the customer actual waiting time as a decision variable.

In models that include customer abandonments, all existing methods fail when considering the actual waiting time as a decision variable (Legros et al. 2016). We propose here a non-standard definition for the system states that leads to a natural uniformized system with no rate modification, or state truncation. We explicitly model the waiting of the first customer in line (FIL) in the system state, instead of the traditional modeling using the number of customers. This idea has been first proposed by Koole et al. (2012) in order to analyze queueing systems with no abandonments. The approach consists of approximating the FIL waiting time using successive exponential phases, and reporting the waiting phase in the Markov process. The difficulty of applying the FIL method in the case of abandonments comes from the fact that the next customer first in line, if any, is no longer necessarily the customer that arrived after the customer who just left the queue. The former might actually have abandoned.

The contributions of this paper can be summarized as follows. We approximate the generally distributed abandonment times by a particular Cox distribution in the sense that we use the same exponential phase distribution as the waiting time approximation. It is referred to as a homogeneous Cox distribution. We prove that this distribution arbitrarily closely approximate any non-negative distribution. The explicit modeling of the waiting time of the FIL leads to a bounded jump Markov process allowing for uniformization. The proposed method is applicable to solve, via dynamic programming, various optimization problems where the objectives and/or constraints involve the distributions of the performance measures, not only their expected values. It can be also used to derive the performance measures of systems where the routing mechanisms are based on the actual waiting time. The limitations of the proposed method are: (i) Customer arrivals should follow a homogeneous Poisson process, (ii) routing decisions have to be taken only after entering the queue, and in the order of arrivals, (iii) structural results may not propagate under a value iteration step. A first illustration of the applicability of the method is given for the optimization of routing decisions in the canonical V-design queue with general abandonment times. A second illustration is given for the performance evaluation of queueing systems with general abandonment times.

The rest of the paper is organized as follows. In Section 2, we describe the FIL modeling and the Cox approximation for abandonment times. In Section 3, we study the convergence of this Cox distribution. In Section 4, we compute the transition probabilities in the FIL Markov chain. We next illustrate in Sections 5 and 6 the applicability of the proposed method for the optimization and the analysis of queues with abandonment. The paper ends with some concluding remarks.

2. Discretization of the First in Line Waiting Time

Consider a queueing system with one infinite first come, first served (FCFS) queue. Customers arrive according to a Poisson process with parameter λ. We let customers be impatient while waiting in the queue. Times before abandonment are i.i.d. and follow a general distribution. The service process is independent of the arrival process and no specific assumption is made on the service time distribution.

We use a non-traditional approach for the definition of the system states, as proposed in Koole et al. (2012). We define a continuous time Markov chain in which we approximate the waiting time of the customer first in line (FIL) by a succession of exponential phases with rate γ per phase. The total number of phases of waiting time required is not known beforehand. This is determined by service completion times and the FIL abandonment time. The system states are defined by the waiting time phase denoted by i ($i > 0$) of the customer FIL, if any. State 0 represents an empty queue. The transition rate from the waiting phase i to $i+1$ is γ, for $i > 0$. The transition rate from state 0 to state 1 is λ.

Once the current FIL leaves the queue from state i ($i > 0$) to start service or because she abandons, the next state is $i-h$, $i > 0$ and $0 \leq h \leq i$. The difficulty here is to find the transition probability, because the next customer first in line, if any, is no longer necessarily the customer that arrived after the FIL who just left. The former might actually have abandoned. We will provide this transition probability in Section 4.

We approximate times before abandonment by a particular Cox random variable denoted by $X_{\gamma,D}$, with D phases ($D > 0$) where all phases durations are independent and exponentially distributed with the same rate γ. It is referred to as a homogeneous γ-Cox random variable. We denote by b the probability for an arbitrary customer to accept waiting, $P(X_{\gamma,D} > 0) = b$. The probability for a given customer to move from phase i to $i+1$ is r_i, $r_i \in [0,1]$ and $1 \leq i \leq D-1$. After phase i, a customer abandons (leaves the system) with probability $1 - r_i$. After phase D, a customer is forced to abandon (rejected by the system), $r_i = 0$ for $i \geq D$. For modeling purposes clearly the subsequent impatience phases run simultaneously with the waiting time phases, hence with the same exponential parameter γ. This homogeneous γ-Cox distribution is depicted in Figure 1. In the following section, we study the convergence of $X_{\gamma,D}$ to any non-negative valued distribution as D and γ tend to infinity.

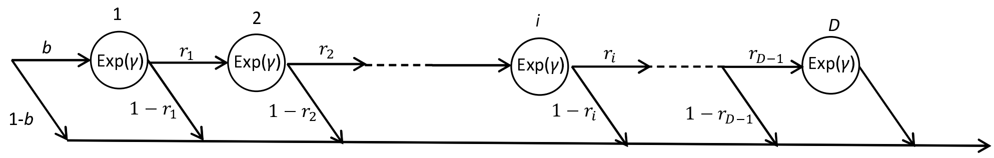


Figure 1 The Homogeneous γ-Cox Distribution for Abandonment Times.

3. Convergence of the Homogeneous γ-Cox Distribution

The truncation of the state space introduces the risk of having a large probability of abandonment at the truncated state, particularly if $\gamma >> D$. Therefore, we consider the iterated limit of the homogeneous γ-Cox random variable $X_{\gamma,D}$ by letting first D and next γ tend to infinity. In Theorem 1, we prove the convergence in distribution of $X_{\gamma,D}$ to any non-negative random variable. We then provide the parameters for $X_{\gamma,D}$ that allow its convergence to some classical random variables. Finally, in Proposition 1, we prove that $X_{\gamma,D}$ does not converge in stronger convergence senses.

Theorem 1. *Let X be a non-negative random variable. There exists parameters of the homogeneous γ-Cox random variable, $X_{\gamma,D}$, such that $X_{\gamma,D}$ converges in distribution to X in the sense*

$$\lim_{\gamma\to\infty}\left(\lim_{D\to\infty} P(X_{\gamma,D} < t)\right) = P(X < t),$$

for any $t \geq 0$.

Proof of Theorem 1. The proof is divided into three steps. In the first step, we prove the existence and uniqueness of $\lim_{D\to\infty} P(X_{\gamma,D} < t)$. The corresponding random variable is denoted by X_γ, $\lim_{D\to\infty} X_{\gamma,D} = X_{\gamma,\infty} = X_\gamma$. In the second step, we prove that the homogeneous γ-Cox random variable X_γ can arbitrarily closely approximate in distribution (ACAD) any Cox random variable, denoted by Z_ϵ, with phase parameters being all different. In the third step, we prove that a Cox random variable, Z_ϵ, with phase parameters being all different, ACAD any Cox random variable (with arbitrarily chosen phase parameters, i.e., all different or not), denoted by Z. The result then follows from Schaßberger's 1973 book where it is proven that Cox distributions are dense in the field of all non-negative distributions (Schaßberger 1973).

Step 1: We prove the existence and uniqueness of the limit of $X_{\gamma,D}$ as D tends to ∞. Let us denote by $E_{k,\gamma}$ the Erlang random variable with k phases and rate γ per phase, $k \geq 1$. The cumulative distribution function (cdf) of $X_{\gamma,D}$ is $1 - P(X_{\gamma,D} > t)$ with

$$\begin{aligned} P(X_{\gamma,D} > t) &= b(1-r_1)P(E_{1,\gamma} > t) + br_1(1-r_2)P(E_{2,\gamma} > t) + \cdots \\ &\qquad + b\prod_{i=1}^{k-1} r_i(1-r_k)P(E_{k,\gamma} > t) + \cdots + b\prod_{i=1}^{D-1} r_i P(E_{D,\gamma} > t) \\ &= be^{-\gamma t}\sum_{k=0}^{D-1}\frac{(\gamma t)^k}{k!}\prod_{i=1}^{k} r_i, \end{aligned}$$

for $t \geq 0$. Since $0 \leq \frac{(\gamma t)^k}{k!}\prod_{i=1}^{k} r_i \leq \frac{(\gamma t)^k}{k!}$ for $k \geq 0$ and $\sum_{k=0}^{D-1}\frac{(\gamma t)^k}{k!}$ converges (to $e^{\gamma t}$) as D tends to infinity, the series $\sum_{k=0}^{D-1}\frac{(\gamma t)^k}{k!}\prod_{i=1}^{k} r_i$ is convergent and approaches 0 as $t \to \infty$. This allows us to define the random variable X_γ as $\lim_{D\to\infty} X_{\gamma,D} = X_{\gamma,\infty} = X_\gamma$, with

$$P(X_\gamma > t) = be^{-\gamma t}\sum_{k=0}^{\infty}\frac{(\gamma t)^k}{k!}\prod_{i=1}^{k} r_i,$$

for $t \geq 0$.

Step 2: Consider the Cox random variable Z_ϵ, with phase parameters being all different. It is defined by the parameters α_j ($\alpha_j > 0$, $\alpha_j \neq \alpha_m$ for $j \neq m$, $1 \leq j, m \leq N$), and p_j ($p_j \in [0,1]$, $0 \leq j \leq N$) with $p_N = 0$. The quantity p_j is the probability to enter phase $j+1$ after leaving phase j and the parameter α_j is the rate of the exponential distribution describing the random duration spent at phase j. Let us now consider specific parameters (b and r_i for $i > 0$) for the random variable X_γ. We choose

$$r_i = \frac{\sum_{n=1}^{N}\left(\frac{\gamma}{\gamma+\alpha_n}\right)^i \sum_{j=n}^{N}(1-p_j)\left(\prod_{m=1,m\neq n}^{j}\frac{\alpha_m}{\alpha_m-\alpha_n}\right)\left(\prod_{m'=1}^{j-1} p_{m'}\right)}{\sum_{n=1}^{N}\left(\frac{\gamma}{\gamma+\alpha_n}\right)^{i-1} \sum_{j=n}^{N}(1-p_j)\left(\prod_{m=1,m\neq n}^{j}\frac{\alpha_m}{\alpha_m-\alpha_n}\right)\left(\prod_{m'=1}^{j-1} p_{m'}\right)},$$

for $i > 0$ and $b = p_0$. Remark that $P(Z_\epsilon > 0) = p_0\sum_{n=1}^{N}\sum_{j=n}^{N}(1-p_j)\left(\prod_{m=1,m\neq n}^{j}\frac{\alpha_m}{\alpha_m-\alpha_n}\right)\left(\prod_{m'=1}^{j-1} p_{m'}\right)$. We also know that $P(Z_\epsilon > 0) = p_0$. Hence, $\sum_{n=1}^{N}\sum_{j=n}^{N}(1-p_j)\left(\prod_{m=1,m\neq n}^{j}\frac{\alpha_m}{\alpha_m-\alpha_n}\right)\left(\prod_{m'=1}^{j-1} p_{m'}\right) = 1$, which implies

$$\prod_{i=1}^{k} r_i = \sum_{n=1}^{N}\left(\frac{\gamma}{\gamma+\alpha_n}\right)^k \sum_{j=n}^{N}(1-p_j)\left(\prod_{m=1,m\neq n}^{j}\frac{\alpha_m}{\alpha_m-\alpha_n}\right)\left(\prod_{m'=1}^{j-1} p_{m'}\right),$$

for $k \geq 1$. Using the previous relation in the expression of $P(X_\gamma > t)$ derived in Step 1, we obtain

$$\begin{aligned} P(X_\gamma > t) &= p_0 e^{-\gamma t} \sum_{k=0}^{\infty} \frac{(\gamma t)^k}{k!} \sum_{n=1}^{N} \left(\frac{\gamma}{\gamma+\alpha_n}\right)^k \sum_{j=n}^{N} (1-p_j) \left(\prod_{m=1, m\neq n}^{j} \frac{\alpha_m}{\alpha_m - \alpha_n}\right) \left(\prod_{m'=1}^{j-1} p_{m'}\right) \\ &= p_0 e^{-\gamma t} \sum_{n=1}^{N} \left(\sum_{k=0}^{\infty} \frac{(\gamma t)^k}{k!} \left(\frac{\gamma}{\gamma+\alpha_n}\right)^k\right) \sum_{j=n}^{N} (1-p_j) \left(\prod_{m=1, m\neq n}^{j} \frac{\alpha_m}{\alpha_m - \alpha_n}\right) \left(\prod_{m'=1}^{j-1} p_{m'}\right) \\ &= p_0 \sum_{n=1}^{N} e^{-\gamma t \frac{\alpha_n}{\gamma+\alpha_n}} \sum_{j=n}^{N} (1-p_j) \left(\prod_{m=1, m\neq n}^{j} \frac{\alpha_m}{\alpha_m - \alpha_n}\right) \left(\prod_{m'=1}^{j-1} p_{m'}\right). \end{aligned}$$

Therefore, as γ tends to infinity, $P(X_\gamma > t)$ converges to

$$p_0 \sum_{n=1}^{N} e^{-\alpha_n t} \sum_{j=n}^{N} (1-p_j) \left(\prod_{m=1, m\neq n}^{j} \frac{\alpha_m}{\alpha_m - \alpha_n}\right) \left(\prod_{m'=1}^{j-1} p_{m'}\right),$$

which is exactly $P(Z_\epsilon > t)$.

Step 3: Consider an arbitrarily Cox random variable denoted by Z. It is defined by the parameters μ_j ($\mu_j > 0$, $1 \leq j \leq N$), and p_j ($p_j \in [0,1]$, $0 \leq j \leq N$) with $p_N = 0$. The quantity p_j is the probability to enter phase $j+1$ after leaving phase j and the parameter μ_j is the rate of the exponential distribution describing the random duration spent at phase j. Let us now consider the particular Cox random variable Z_ϵ defined by the parameters $\mu_j(1+\epsilon)^j$ and p_j for $\epsilon > 0$ and $0 \leq j \leq N$. In what follows, we show that for sufficiently small values of ϵ, the rates of Z_ϵ are all different. If $\mu_j \leq \mu_m$ for $j \neq m$, then $\mu_j(1+\epsilon)^j \neq \mu_m(1+\epsilon)^m$. In the opposite case when $\mu_j > \mu_m$, the equation in ϵ: $\mu_j(1+\epsilon)^j = \mu_m(1+\epsilon)^m$ has a unique solution. This solution is $\epsilon = \left(\frac{\mu_j}{\mu_m}\right)^{\frac{1}{m-j}} - 1$. We therefore choose ϵ such that

$$\epsilon < \min_{0 \leq j < m \leq N, \mu_j > \mu_m} \left\{ \left(\frac{\mu_j}{\mu_m}\right)^{\frac{1}{m-j}} - 1 \right\}.$$

This choice ensures that the exponential rates of Z_ϵ are all different.

We next focus on the convergence in distribution of the sequence Z_ϵ as ϵ tends to zero. Recall first that the Levy continuity theorem for Laplace transforms states that a sequence of random variables converges in distribution if and only if the sequence of their respected Laplace transforms also converges. It suffices then to prove the convergence in distribution of the Laplace transform of

Z_ϵ to that of Z. The Laplace transforms of Z_ϵ and Z are denoted by $G_{Z_\epsilon}(.)$ and $G_Z(.)$, respectively. We have

$$G_{Z_\epsilon}(s) = p_0 \sum_{k=1}^{N-1} (1-p_k) \prod_{i=1}^{k-1} p_i \prod_{j=1}^{k} \left(\frac{\mu_j (1+\epsilon)^j}{s + \mu_j (1+\epsilon)^j} \right),$$

and

$$G_Z(s) = p_0 \sum_{k=1}^{N-1} (1-p_k) \prod_{i=1}^{k-1} p_i \prod_{j=1}^{k} \left(\frac{\mu_j}{s+\mu_j} \right),$$

for $s \geq 0$. Therefore,

$$|G_{Z_\epsilon}(s) - G_Z(s)| \leq p_0 \sum_{k=1}^{N-1} (1-p_k) \prod_{i=1}^{k-1} p_i \left| \prod_{j=1}^{k} \left(\frac{\mu_j (1+\epsilon)^j}{s + \mu_j (1+\epsilon)^j} \right) - \prod_{j=1}^{k} \left(\frac{\mu_j}{s+\mu_j} \right) \right|,$$

where $|x|$ is the absolute value of $x \in \mathbb{R}$. The summation and the products in the expression of $|G_{Z_\epsilon}(s) - G_Z(s)|$ are finite. Moreover,

$$\lim_{\epsilon \to 0} \frac{\mu_j (1+\epsilon)^j}{s + \mu_j (1+\epsilon)^j} = \frac{\mu_j}{s+\mu_j}.$$

We thus conclude that $|G_{Z_\epsilon}(s) - G_Z(s)|$ tends to 0 as ϵ tends to zero. The proof of the theorem is completed. $\square$

Table 1 gives the parameters of X_γ that ensure the convergence in distribution of X_γ to some classical distributions. The proofs of convergence to the various phase-type distributions (exponential, Erlang, hyperexponential and hypoexponential) in the table are similar to the proof of Theorem 1, except for the deterministic distribution which is slightly different. It is as follows. Assume the parameters of X_γ as in point 4 of Table 1. We may write

$$P(X_\gamma > t) = e^{-\frac{n}{\tau}t} \sum_{k=0}^{n} \frac{\left(\frac{n}{\tau}t\right)^k}{k!} = P(Y \leq n),$$

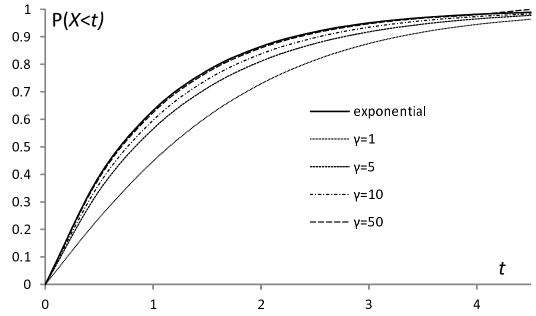
where Y, a random variable, follows a Poisson distribution with parameter $\frac{n}{\tau}t$, for $t \geq 0$. We have $P(Y \leq n) = P\left(\frac{Y - \frac{n}{\tau}t}{\sqrt{\frac{n}{\tau}t}} \leq \frac{n(1-t/\tau)}{\sqrt{\frac{n}{\tau}t}} \right)$. Using the Central Limit Theorem, the distribution of $\frac{Y - \frac{n}{\tau}t}{\sqrt{\frac{n}{\tau}t}}$ converges to a normal distribution with mean 0 and standard deviation 1 as n tends to infinity. If $\tau > t$, then $\frac{n(1-t/\tau)}{\sqrt{\frac{n}{\tau}t}}$ tends to $+\infty$ as n tends to infinity, therefore, $P(X_\gamma > t)$ tends to one.

Table 1 Parameters of X_γ to Fit Classical Distributions.

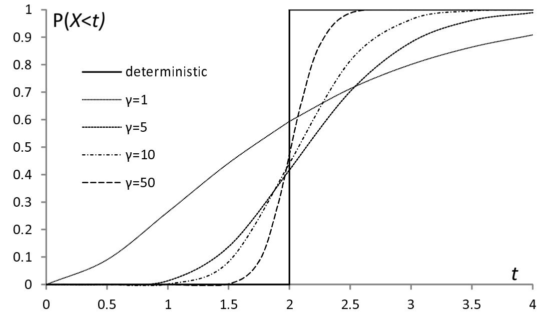
Distribution	**Parameters for X_γ**
1. Infinite patience	$b=1$, $r_i=1$ for $i>0$
2. Infinite impatience	$b=0$
3. Exponential (β)	$b=1$ and $r_i=\frac{\gamma}{\gamma+\beta}$ for $i>0$
4. Deterministic (τ)	$b=1$, $r_i=1$ for $0<i\leq n$ and $r_i=0$ for $i>n$, where $n\in\mathbb{N}$, and $\gamma=\frac{n}{\tau}$
5. Erlang (N,β)	$b=1$ and $r_i=\frac{\gamma}{\gamma+\beta}\cdot\frac{\sum_{n=0}^{N-1}\binom{i}{n}\left(\frac{\beta}{\gamma}\right)^n}{\sum_{n=0}^{N-1}\binom{i-1}{n}\left(\frac{\beta}{\gamma}\right)^{n-1}}$ for $i>0$
6. Hyperexponential (α_n,p_n) with $\alpha_n>0$ and $p_n\in[0,1]$, $p_1+p_2+\cdots+p_N=1$ for $1\leq n\leq N$	$b=1$ and $r_i=\frac{\sum_{n=1}^{N}p_n\left(\frac{\gamma}{\gamma+\alpha_n}\right)^i}{\sum_{n=1}^{N}p_n\left(\frac{\gamma}{\gamma+\alpha_n}\right)^{i-1}}$ for $i>0$
7. Hypoexponential (α_n) with $\alpha_n>0$, $\alpha_n\neq\alpha_m$ for $n\neq m$, $1\leq n,m\leq N$	$b=1$ and $r_i=\frac{\sum_{n=1}^{N}\left(\frac{\gamma}{\gamma+\alpha_n}\right)^i\prod_{m\neq n}\frac{\alpha_m}{\alpha_m-\alpha_n}}{\sum_{n=1}^{N}\left(\frac{\gamma}{\gamma+\alpha_n}\right)^{i-1}\prod_{m\neq n}\frac{\alpha_m}{\alpha_m-\alpha_n}}$ for $i>0$

Otherwise, $\frac{n(1-t/\tau)}{\sqrt{\frac{n}{\tau}t}}$ tends to $-\infty$, therefore, $P(X_\gamma>t)$ tends to 0. This corresponds to the cdf of a deterministic distribution with parameter τ.

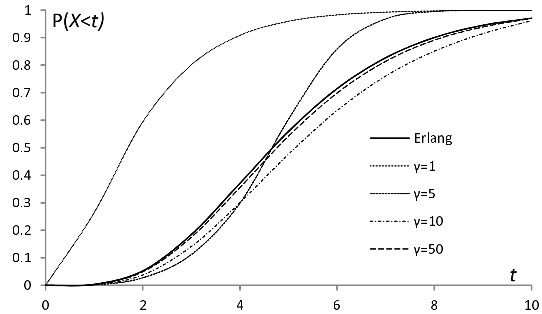
Figure 6 illustrates the convergence of $X_{\gamma,D}$ for the points 3 to 6 in Table 1. The value of γ has a significant impact on the approximation. Increasing it means that more states are required for the truncation to not have a too significant influence on the accuracy of the approximation. However, at the same time, having a large γ models better a continuous time. Therefore, D has to tend “faster” to infinity than γ. We choose $D=1+\gamma^2$ in the illustrations in Figure 6.



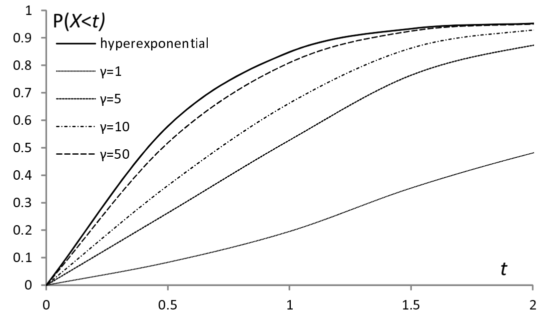
(a) Exponential Distribution with rate 1.



(b) Deterministic distribition with parameter 2.



(c) Erlang distribution with 5 phases and rate 1.



(d) Hyperexponential distribution with parameters $(1,10\%),(10,90\%)$.

Figure 2 Convergence of $X_{\gamma,D}$.

The convergence in distribution is the weakest type of convergence for random variables. We next show an example of non-convergence in probability for X_γ. This implies that X_γ does not converge in convergence senses that are stronger than convergence in distribution.

Proposition 1. *With $b=1$ and $r_i=\frac{\gamma}{\gamma+\beta}$ for $i>0$, X_γ does not converge in probability to an exponential random variable with parameter β.*

Proof of Proposition 1. We denote by Y the exponential random variable with parameter β. Recall that the definition of the convergence in probability is that $P(|X_\gamma - Y|>\epsilon)$ tends to zero as γ tends to infinity, for any $\epsilon>0$. The density function of X_γ in the variable t is $f_{X_\gamma}(t)=\gamma e^{-\gamma t}\sum_{k=0}^{\infty}\frac{(\gamma t)^k}{k!}\left(\frac{\gamma}{\gamma+\beta}\right)^{k+1}\mathbb{I}_{t\geq 0}$ and the density function of $-Y$ is $f_{-Y}(t)=\beta e^{\beta t}\mathbb{I}_{t\leq 0}$, where $\mathbb{I}_A$ is the indicator function of a given subset A.

Let us now compute the density function of $X_\gamma - Y$, defined as $f_{X_\gamma - Y}(z)$, for $z \in \mathbb{R}$. We have

$$\begin{aligned}
f_{X_\gamma - Y}(z) &= \gamma\beta \int_z^\infty e^{-\gamma t} \sum_{k=0}^\infty \frac{(\gamma t)^k}{k!} \left(\frac{\gamma}{\gamma+\beta}\right)^{k+1} e^{\beta(z-t)}\, \mathrm{d}t \\
&= \beta \frac{\gamma}{\gamma+\beta} \frac{\beta}{\gamma+\beta} e^{-\gamma z} \sum_{k=0}^\infty \left(\frac{\gamma}{\gamma+\beta}\right)^{2k} \sum_{i=0}^k \frac{((\gamma+\beta)z)^i}{i!} \\
&= \frac{\gamma\beta^2 e^{-\gamma z \frac{\beta}{\gamma+\beta}}}{2\gamma\beta+\beta^2} \mathbb{I}_{z\geq 0},
\end{aligned}$$

for $z \geq 0$. After some algebra, we obtain

$$f_{X_\gamma - Y}(z) = \frac{\gamma\beta^2 e^{\beta z}}{2\gamma\beta+\beta^2} \mathbb{I}_{z\leq 0},$$

for $z \leq 0$. Using this density function, we may write

$$\begin{aligned}
P(|X_\gamma - Y| > \epsilon) = P(X_\gamma - Y > \epsilon) + P(X_\gamma - Y < -\epsilon) \\
= \frac{\beta(\gamma+\beta)}{2\gamma\beta+\beta^2} e^{-\gamma\epsilon\frac{\beta}{\gamma+\beta}} + \frac{\beta\gamma}{2\gamma\beta+\beta^2} e^{-\beta\epsilon}.
\end{aligned}$$

Therefore, as γ tends to infinity, $P(|X_\gamma - Y| > \epsilon)$ tends to $e^{-\epsilon\beta} \neq 0$. This finishes the proof of the proposition. $\square$

4. Transition Probabilities

Theorem 2 provides the expressions of the transition probabilities $p_{i,i-h}$ to move from phase i to phase $i-h$ in the Markov chain defined in Section 2, for $0 < i \leq D$ and $0 \leq h \leq i$.

Theorem 2. *We have*

$$p_{i,i-h} = \begin{cases} \prod_{k=1}^{i} q_k \ \textit{for } i = h, 0 < i \leq D, \\ (1 - q_{i-h}) \prod_{k=i-h+1}^{i} q_k, \ \textit{for } 0 \leq h < i \leq D, \end{cases} \tag{1}$$

where

$$q_k = \left[1 + \frac{b\lambda}{\gamma} \prod_{j=1}^{k} r_j\right]^{-1}, \ \textit{for } 0 < k \leq D. \tag{2}$$

Proof of Theorem 2. The number of customers that arrive at a given γ-phase is geometrically distributed with parameter $\frac{\gamma}{b\lambda+\gamma}$. Thus, the probability that exactly n customers arrive at the same γ-transition is $\left(\frac{b\lambda}{b\lambda+\gamma}\right)^n \frac{\gamma}{b\lambda+\gamma}$. The probability that a given customer does not abandon after k γ-transitions is $\prod_{j=1}^{k} r_j$. Therefore, the probability that a customer does abandon is $1-\prod_{j=1}^{k} r_j$. The probability $p_{i,i-h}$, for $0<i\leq D$ and $0\leq h<i$, is the probability that when the FIL leaves the head of the queue from state i (after an abandonment or to enter service), the new FIL is in state $i-h$. This probability is the probability that we do not have any customer at phases $i-h+1, i-h+2, \cdots, i$ and that at least one customer is present in the queue at phase $i-h$. The probability to not have any customer at a given phase k, $0<k\leq D$, is denoted by q_k. We may then write $p_{i,i-h}=(1-q_{i-h})\prod_{k=i-h+1}^{i} q_k$, for $0<i\leq D$ and $0\leq h<i$. The quantity $p_{i,0}$ is the probability that no other customers are present in the queue when the FIL leaves the queue, i.e., the queue becomes empty. So, $p_{i,0}=\prod_{k=1}^{i} q_k$, for $0<i\leq D$. We next compute q_k, for $0<k\leq D$. We have

$$q_k=\sum_{n=0}^{\infty}\left(\frac{b\lambda}{b\lambda+\gamma}\right)^n \frac{\gamma}{b\lambda+\gamma}\left(1-\prod_{j=1}^{k} r_j\right)^n,$$

from which we deduce after some algebra Equation (2). This finishes the proof of the theorem. □

For the extreme case with no balking and infinitely patient customer ($b=1$ and $r_j=1$ for $1\leq j\leq D$), we obtain $q_k=\frac{\gamma}{\lambda+\gamma}$. The expressions of $p_{i,i-h}$, for $0<i\leq D$ and $0\leq h\leq i$, then reduce to those found in Koole et al. (2012).

5. Numerical Illustration: Optimal Routing

We numerically illustrate the applicability of the FIL method for the optimal job routing in the canonical V-design queueing model with abandonments.

Optimization Problem. Two customer classes, A and B, have each their own queue and are both served by a common group of s servers. Within in each queue, customers are selected according to a first in, first out principle. Arrivals to each queue happen according to Poisson processes with rates λ_A and λ_B. Service times are assumed to be i.i.d. and exponentially distributed with rate μ for both classes. Abandonment times are approximated by two different independent homogeneous

γ-Cox random variables with parameters $r_{A,i}$ and $r_{B,j}$, for $i,j \geq 1$, where i and j represent the waiting phases of the FIL in queues A and B, respectively. We assume no balking for both customer classes, $b_A = b_B = 1$. The queueing model is shown in Figure 3. We denote by $p_{A,i,h}$ and $p_{B,j,h'}$ the transition probabilities from state i to state h in queue A, and from state j to state h' in queue B, respectively ($i,j > 0$, $0 \leq h \leq i$ and $0 \leq h' \leq j$). We restrict the analysis to non-preemptive and non-idling policies.

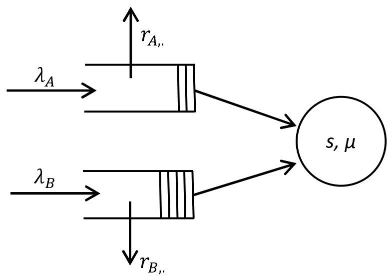


Figure 3 The V-Design Queue.

The objective of the system manager is to minimize a linear combination of the stationary waiting time performance of the two customer classes. In the numerical experiments below, we consider two problem formulations: Formulation (1) consists of minimizing a linear combination of the expected waiting times in queues A and B, and Formulation (2) consists in minimizing a linear combination of percentiles of the waiting times in queues A and B.

The control action is to determine upon a service completion, when at least one customer is waiting in each queue, which customer should be prioritized. The system is modeled using a two dimensional continuous-time Markov chain. Since $A-$ and $B-$customers have the same service rate, we do not distinguish between them when they are in service. The system state is denoted by (i,j), where $i \geq -s$ and $j \geq 0$. States with $i \leq 0$ correspond to both queues empty and $s+i$ busy servers. Waiting times of customers are represented by states with positive indices $(i,j > 0)$.

We define for the two customer classes waiting cost functions denoted by $c_A(i)$ and $c_B(j)$, for $i,j \geq 0$. For Formulation (1), we choose increasing linear cost functions, and for Formulation (2), we choose step functions being 0 below a certain threshold and some non-zero value above this threshold.

Equations Setup. We use an MDP approach. Let V_n be the total expected value function n steps from the horizon and let us use backward recursion to determine the optimal policy. Since the system is uniformizable, we assume that $\lambda_A + \lambda_B + \gamma + s\mu = 1$. We denote by $W_n(i,j)$ the decision function to select customer A or B for service upon a service completion. Assume $V_0(i,j) = W_0(i,j) = 0$, for $i \geq -s$ and $j \geq 0$. We have

$$W_n(i,j) = \min\left(c_A(i) + \sum_{h=0}^{i} p_{A,i,h} V_{n-1}(h,j), c_B(j) + \sum_{h=0}^{j} p_{B,j,h} V_{n-1}(i,h)\right), \text{ for } i,j > 0, \tag{3}$$

$$W_n(i,0) = c_A(i) + \sum_{h=0}^{i} p_{A,i,h} V_n(h,0), \text{ for } i > 0,$$

$$W_n(0,j) = c_B(j) + \sum_{h=0}^{j} p_{B,j,h} V_n(0,h), \text{ for } j > 0,$$

for $n \geq 0$. We may write

$$V_{n+1}(i,0) = (\lambda_A + \lambda_B) V_n(i+1,0) + (s+i)\mu V_n(i-1,0) + (1 - \lambda_A - \lambda_B - (s+i)\mu) V_n(i,0), \text{ for } -s \leq i < 0, \tag{4}$$

$$V_{n+1}(0,0) = \lambda_A V_n(1,0) + \lambda_B V_n(0,1) + s\mu V_n(-1,0) + (1 - \lambda_A - \lambda_B - s\mu) V_n(0,0),$$

$$V_{n+1}(i,0) = \lambda_B V_n(i,1) + \gamma r_{A,i} V_n(i+1,0) + (s\mu + \gamma(1 - r_{A,i})) W_n(i,0) + (1 - \lambda_B - \gamma - s\mu) V_n(i,0), \text{ for } i > 0,$$

$$V_{n+1}(0,j) = \lambda_A V_n(1,j) + \gamma r_{B,j} V_n(0,j+1) + (s\mu + \gamma(1 - r_{B,j})) W_n(0,j) + (1 - \lambda_A - \gamma - s\mu) V_n(0,j), \text{ for } j > 0,$$

$$\begin{aligned} V_{n+1}(i,j) &= \gamma r_{A,i} r_{B,j} V_n(i+1,j+1) + s\mu W_n(i,j) + \gamma(1 - r_{A,i}) r_{B,j} \left(c_A(i) + \sum_{h=0}^{i} p_{A,i,h} V_n(h,j+1)\right) \\ &+ \gamma r_{A,i}(1 - r_{B,j}) \left(c_B(j) + \sum_{h=0}^{j} p_{B,j,h} V_n(i+1,h)\right) \\ &+ \gamma(1 - r_{A,i})(1 - r_{B,j}) \left(c_A(i) + c_B(j) + \sum_{h=0}^{i}\sum_{h'=0}^{j} p_{A,i,h} p_{B,j,h} V_n(h,h')\right) + (1 - \gamma - s\mu) V_n(i,j), \text{ for } i,j > 0, \end{aligned}$$

for $n \geq 0$. One way of obtaining the long-run average optimal actions is to apply the value iteration technique introduced by Bellman (1957) and Howard (1960), by recursively evaluating V_n using Equations (3) and (4), for $n \geq 0$. As n tends to infinity, the optimal policy converges to the unique average optimal policy. Moreover, the optimal long-run policy is independent of the choice of V_0. The convergence is due to the aperiodic irreducible finite-state Markov chains considered here. The aperiodicity is due to the fictitious transitions from a state to itself. Then, Theorem 8.5.3 part c of Puterman (1994) guarantees the existence of an optimal deterministic stationary policy.

Experiments. For $A-$customers ($B-$customers) abandonment times, we assume a 2-phase hyperexponential distribution with probability u_A (u_B) associated with the exponential rate $\alpha_{A,1}$ ($\alpha_{B,1}$) and probability $1-u_A$ ($1-u_B$) associated with the exponential rate $\alpha_{A,2}$ ($\alpha_{B,2}$). This choice is motivated in practice by Jouini et al. (2013) and Mandelbaum and Zeltyn (2013) where it has been shown that hyperexponential distributions fit well with real call center data. Using the fitting parameters from Table 1, one may write

$$r_{A,i} = \frac{u_A\left(\frac{\gamma}{\gamma+\alpha_{A,1}}\right)^i + (1-u_A)\left(\frac{\gamma}{\gamma+\alpha_{A,2}}\right)^i}{u_A\left(\frac{\gamma}{\gamma+\alpha_{A,1}}\right)^{i-1} + (1-u_A)\left(\frac{\gamma}{\gamma+\alpha_{A,2}}\right)^{i-1}},$$

for $i>0$, and

$$r_{B,j} = \frac{u_B\left(\frac{\gamma}{\gamma+\alpha_{B,1}}\right)^j + (1-u_B)\left(\frac{\gamma}{\gamma+\alpha_{B,2}}\right)^j}{u_B\left(\frac{\gamma}{\gamma+\alpha_{B,1}}\right)^{j-1} + (1-u_B)\left(\frac{\gamma}{\gamma+\alpha_{B,2}}\right)^{j-1}},$$

for $j>0$. Next, using Equation (2) in Theorem 2, we obtain

$$q_{A,k} = \left[1+\frac{\lambda_A}{\gamma}\left(u_A\left(\frac{\gamma}{\gamma+\alpha_{A,1}}\right)^k + (1-u_A)\left(\frac{\gamma}{\gamma+\alpha_{A,2}}\right)^k\right)\right]^{-1},$$

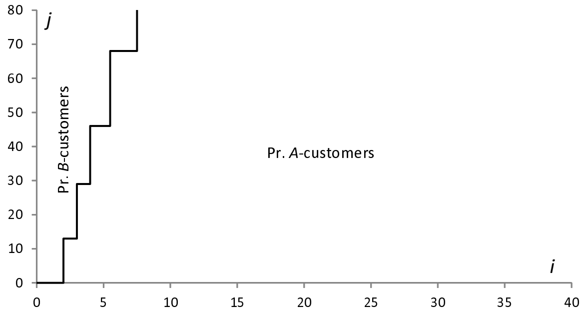
and

$$q_{B,k} = \left[1+\frac{\lambda_B}{\gamma}\left(u_B\left(\frac{\gamma}{\gamma+\alpha_{B,1}}\right)^k + (1-u_B)\left(\frac{\gamma}{\gamma+\alpha_{B,2}}\right)^k\right)\right]^{-1},$$

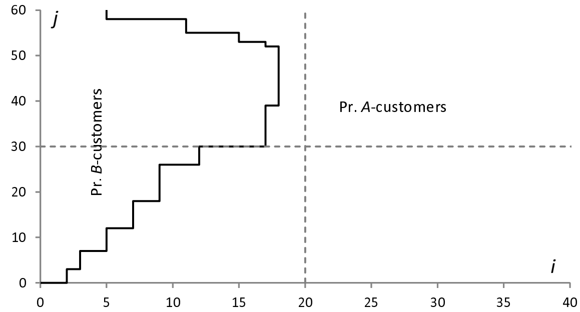
for $0<k\le D$. Finally, Equation (1) in Theorem 2 leads to the transition probabilities $p_{A,i,h}$ and $p_{B,j,h'}$ ($i,j>0$, $0\le h\le i$ and $0\le h'\le j$). Applying now the value iteration technique, we obtain the optimal routing policies for Formulations (1) and (2). They are shown in Figures 4(a) and 4(b), respectively.

The chosen numerical setting in the figures gives a higher importance for $A-$customers. From Figure 4(a), we observe that the optimal policy is of switch type, and that the switching curve is increasing. The interpretation is intuitive. The higher is the waiting time of the FIL in one queue, the more likely this customer will be prioritized upon the next service completion.

In Figure 4(b), the optimal policy is also of switch type. However, the switching curve is no longer monotone. We observe as expected that $A-$customers are most of the time prioritized. They



(a) $c_A(i) = 5i/\gamma$, $c_B(j) = j/\gamma$, $i, j > 0$.



(b) $c_A(i) = 5$ for $i > 20$ and 0 otherwise, $c_B(j) = 1$, for $j > 30$ and 0 otherwise.

Figure 4 Optimal Policies ($\lambda_A = \lambda_B = 5$, $\mu = 1$, $s = 11$, $u_A = 0.1$, $\alpha_{A,1} = 1$, $\alpha_{A,2} = 5$, $u_B = 0.3$, $\alpha_{B,1} = 2$, $\alpha_{B,2} = 3$, $\gamma = 30$, $D = 120$).

lose priority when the FIL waiting time in queue A is small or when that in queue B is around (below or above) the B waiting threshold (30 time units). The interest from serving a queue B FIL, with an age higher than the threshold, is that the B waiting customers after her are likely to have an age below the threshold. This does not happen however when the B FIL elapsed waiting time is much higher than the threshold. There is no longer a reason to select this customer for service since the following B waiting customers are likely to have ages higher than the threshold. It can be better to let those customers abandon so as we thereafter encounter a new B FIL with a better elapsed waiting time, i.e., close to the threshold.

6. Numerical Illustration: Performance Analysis

We show the applicability of our results for the numerical computation of the M/M/s+GI performance measures. The performance analysis of this queueing model is known from Baccelli and Hebuterne (1981). We illustrate how the FIL process converges to that of the M/M/s+GI queue.

State Definition. The M/M/s+GI system is analyzed using a one dimensional continuous-time Markov chain. We denote by x a state of the system for $-s \leq x \leq D$, where x represents the servers state or the waiting time in the queue. More precisely, states with $-s \leq x \leq 0$ correspond to an empty queue and $s + x$ busy agents. States with $0 < x \leq D$ correspond to the phase at which the FIL in the queue is waiting and all agents are busy. Lumping together the states representing free servers and the waiting time of the FIL in the queue in one dimension can be done as servers cannot be free while customers are waiting.

Transitions. We next describe the 6 possible transition types in the Markov chain.

1. An arrival with rate λ while the queue is empty ($-s \leq x \leq 0$), which changes the state to $x+1$. If $-s \leq x < 0$, then the number of busy servers is increased by 1. If $x = 0$, then the FIL entity is created.

2. A service completion with rate $(s+x)\mu$ while the queue is empty ($-s < x \leq 0$), which changes the state to $x-1$. The number of busy servers is decremented by 1.

3. A service completion with rate $s\mu p_{x,x-h}$ while the queue is not empty ($0 < x \leq D$), which changes the state to $x-h$, that is, the new FIL is in waiting phase $x-h$.

4. A phase increase which does not lead to an abandonment with rate γr_x while the queue is not empty and the FIL is not in waiting phase D ($0 < x < D$), which changes the state to $x+1$. The waiting phase of the FIL is incremented by 1.

5. A phase increase which leads to an abandonment with rate $\gamma(1-r_x)p_{x,x-h}$ while the queue is not empty and the FIL is not in waiting phase D ($0 < x < D$), which changes the state to $x-h$, that is, the new FIL is in waiting phase $x-h$.

6. A phase increase with rate $\gamma q_{D,D-h}$ while the FIL is in waiting phase D, which changes the state to $D-h$, that is, the new FIL is in waiting phase $D-h$.

When the FIL changes because of a service completion (transition Type 3), because of an abandonment (transition Type 5) or because of a rejection (transition Type 6), the waiting time phase changes from $x > 0$ to $x-h$ with probability $p_{x,x-h}$ (given in Theorem 2).

Equilibrium Equations. We denote by π_x the stationary probability to be in state x for $-s \leq x \leq D$. Let S be the state space. Consider the cut between $A = \{-s, \cdots, x\}$ and $S \backslash A$, where $-s \leq x < D$. By equating flows across the cut, one may write

$$\lambda\pi_x = (s+x+1)\mu\pi_{x+1}, \text{ for } -s \leq x < 0, \tag{5}$$

$$\lambda\pi_0 = \sum_{i=1}^{D-1}(s\mu+\gamma(1-r_i))\pi_i \cdot p_{i,0} + (s\mu+\gamma)\pi_D \cdot p_{D,0}, \tag{6}$$

$$\gamma\pi_x = \sum_{i=x+1}^{D-1}(s\mu+\gamma(1-r_i))\pi_i \left(1-\sum_{k=x+1}^{i} p_{i,k}\right) + (s\mu+\gamma)\pi_D \left(1-\sum_{k=x+1}^{D} p_{D,k}\right), \tag{7}$$

for $0<x<D$. Using Equation (7), one may obtain an expression of π_x as a function of π_D. Equation (6) then leads to the expression of π_0 as a function of π_D. Finally, the remaining probabilities are obtained from Equation (5). Using next the fact that all probabilities sum up to one, one may deduce π_D.

The Embedded Markov Chain. Arriving customers either enter service upon arrival, enter service from the queue after some wait, abandon after experiencing some wait or are rejected after D phases of wait. Call the instants when one of these four events occurs Q-instants. To compute the performance measures, we use an embedded Markov chain approach in which we use the system state probabilities seen at Q-instants.

The Q-instants are determined by λ-transitions from states with a vacant server (transition Type 1), $s\mu+\gamma(1-r_x)$-transitions from the other states except state 0 (transition Types 3 and 5) and γ-transitions from state D (transition Type 6). The system state probability at Q-instants is denoted by $\widetilde{\pi}_x$ and is given by

$$\widetilde{\pi}_x = \frac{\Lambda_x\pi_x}{\sum_{i=-s}^{D}\Lambda_i\pi_i},$$

where

$$\Lambda_x = \begin{cases} \lambda \text{ for } -s\le x<0, \\ s\mu+\gamma(1-r_x) \text{ for } 0<x<D, \\ s\mu+\gamma \text{ for } x=D. \end{cases}$$

From the stationary probabilities at Q-instants, we next show how the performance measures can be derived.

Performance Measures. Let W, a random variable, be the unconditional customer waiting time in the queue. A customer entering service when $x<0$ goes directly to a free server and experiences no waiting. When a customer enters service, abandons or is rejected from a state $x>0$, she has waited a sum of x exponentially distributed time periods, each with mean $1/\gamma$. Therefore, the expected waiting time, $E(W)$, in the queue can be written as

$$E(W)=\sum_{x=1}^{D}\frac{x}{\gamma}\widetilde{\pi}_x.$$

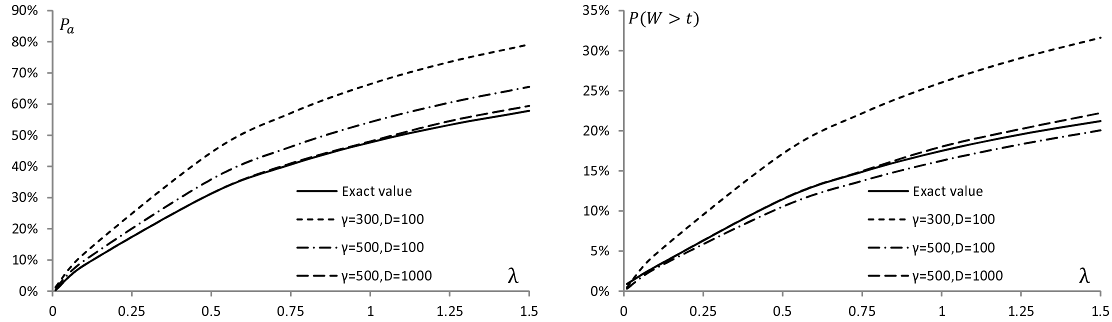
Let $F_{\gamma,x}(t)=1-\sum_{i=0}^{x-1}\frac{(\gamma t)^i}{i!}e^{-\gamma t}$ be the cdf of an Erlang random variable with shape parameter $x\geq 1$ and scale parameter $\gamma\in\mathbb{R}^+$. The waiting time distribution of a customer in the system can be deduced from

$$P(W>t)=\sum_{x=1}^{D}(1-F_{\gamma,x}(t))\widetilde{\pi}_x,$$

for $t\geq 0$. Customers abandon due to a $\gamma(1-r_x)$ transition from state x for $0<x<D$ or due to a γ-transition from state $x=D$. Since the overall mean flow of arrivals is λ, one may obtain the probability of abandonment, P_a, through

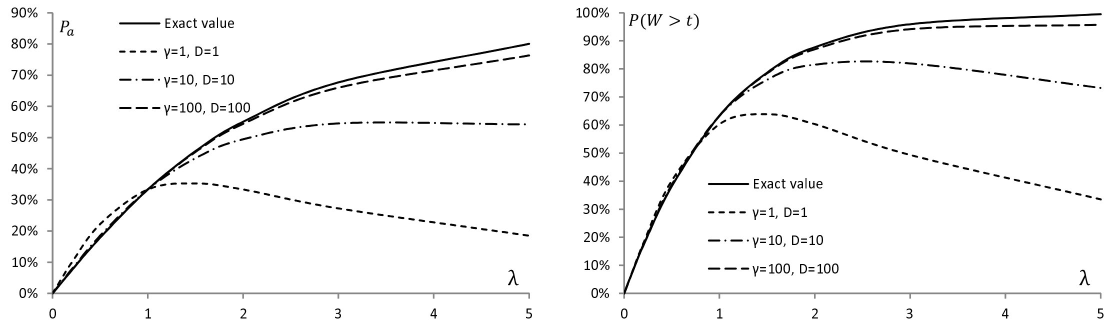
$$P_a=\sum_{x=1}^{D-1}\frac{\gamma(1-r_x)}{\lambda}\pi_x+\frac{\gamma}{\lambda}\pi_D.$$

Illustration. Based on the analysis above, we are ready to illustrate the convergence of the FIL process to some classical ones. This is given in Figures 5 and 6 for the processes associated to the M/M/s+M (exponential abandonment with rate β) and M/M/s+D (deterministic abandonment time τ) queues, respectively. We use the fitting parameters as given in lines 3 and 4 of Table 1. In the numerical computation, the two parameters γ and D should be carefully chosen (Koole et al. 2012). The truncation parameter D introduces the risk of having a large probability mass in the truncated state, particularly for large values of γ. The value of γ has an important influence on the approximation. Increasing γ means that more states are required for the truncation. At the same time, γ should be sufficiently large to represent the continuous elapsing of time.



(a) Proportion of Abandonment. (b) $P(W > t)$.

Figure 5 Convergence of the Performance Measures in the M/M/s+M queue ($s = 1$, $\mu = 1$, $t = 0.1$, $\beta = 10$).



(a) Proportion of Abandonment. (b) $P(W > t)$.

Figure 6 Convergence of the Performance Measures in the M/M/s+D Queue ($s = 1$, $\mu = 1$, $t = 0.1$, $\tau = 1$).

7. Concluding Remarks

We considered multi-server queueing systems with general abandonment. Abandonment times are approximated by a homogeneous Cox distribution. We proved that this distribution arbitrarily closely approximate any non-negative distribution. We proposed a Markov process that explicitly models the waiting time of the first customer in line which has led to a bounded jump Markov process allowing for uniformization. This method is applicable for the performance evaluation and the optimization of queueing systems where routing decisions are based on actual waiting times, and not only their expected values. Illustrations of the applicability of the results were given for the dynamic control of the V-design queue and for the performance analysis of queueing systems with customer abandonment.

An interesting future research direction is to include the modeling of general service times so as to enlarge the class of possible practical applications. It would be also interesting to extend the set of scheduling policies by relaxing the FCFS order.

Acknowledgments

This work was supported by Agence Nationale de la Recherche under the project ANR-JCJC-SIMI3-2012-OPERA. The authors want to express their gratitude to the review team for their useful comments, that significantly improved this paper. The authors would like also to thank Sébastien Thorel from INTERACTIV GROUP and René Bekker from VU University Amsterdam for their helpful discussions.

References

Baccelli F, Hebuterne G (1981) On queues with impatient customers. *Performance'81* North-Holland Publishing Company, 159-179.

Bailey ED, Sweeney T (2003) Considerations in establishing emergency medical services response time goals. *Prehospital Emergency Care* 7(3):397–399.

Bellman R (1957) *Dynamic programming* (Princeton University Press, Princeton).

Bhulai S, Brooms AC, Spieksma FM (2014) On structural properties of the value function for an unbounded jump markov process with an application to a processor sharing retrial queue. *Queueing Systems* 76(4):425–446.

Down DG, Koole G, Lewis ME (2011) Dynamic control of a single-server system with abandonments. *Queueing Systems* 67(1):63–90.

Howard RA (1960) *Dynamic Programming and Markov Processes.*

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

Koole G, Nielsen BF, Nielsen TB (2012) First in line waiting times as a tool for analysing queueing systems. *Operations research* 60(5):1258–1266.

Legros B (2016) 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.

Legros B, Jouini O, Koole G (2016) Optimal scheduling in call centers with a callback option. *Performance Evaluation* 95:1–40.

Mandelbaum A, Zeltyn S (2013) Data-stories about (im)patient customers in tele-queues. *Queueing Systems* 75(2):115–146.

Puterman M (1994) *Markov Decision Processes* (John Wiley and Sons).

Schaßberger R (1973) *Warteschlangen* (Springer-Verlag Vienna).

Benjamin Legros is Professor in operations management at EM Normandie. He received a B.Sc. degree in industrial engineering from CentraleSupélec in 2006. He carried out his Ph.D. research on the optimization of multi-channel call centers at CentraleSupélec and received a Ph.D degree in 2013. His current research interests are in stochastic modeling, queueing theory and operations management of call centers.

Oualid Jouini is Professor in operations management at CentraleSupélec. He received a B.Sc. degree in Industrial Engineering from Ecole Nationale d'Ingénieurs de Tunis in 2001 and a M.Sc. degree in Industrial Engineering from CentraleSupélec in 2003. He received a Ph.D. degree on the optimization of call centers at CentraleSupélec in 2006 and held a postdoc position at University of Minnesota in 2007. He holds the chair *Call Centers* at CentraleSupélec. His current research interests are in stochastic modeling and service operations management. His main application areas are call centers and healthcare systems.

Ger Koole Dr. Ger M. Koole is Professor at VU University Amsterdam. He graduated at Leiden University in 1992 on a thesis on the control of queueing systems and held post-doc positions at CWI Amsterdam and INRIA Sophia Antipolis. He holds the chair "optimization of business processes" which makes him responsible for research in applied operations research and for the bachelor and master programs in business analytics.

He has supervised 14 PhD students and published over 90 papers in the international literature. Next to his academic work Dr. Koole co-founded three companies: the call center planning company CCmath, the internet advertisement company Adscience, and the hotel revenue management company IrevenU. He is also a founder of PICA, the VU university/medical center joint knowledge center on health care operations management, and of ACBA, the multi-disciplinary "Amsterdam Center for Business Analytics" which has been very successful in obtaining company funding for academic research.