EURANDOM PREPRINT SERIES

2017-002

February 6, 2017

The singel server queue with mixing dependencies

Y. Raaijmakers, H.-J. Albrecher, O. Boxma

ISSN 1389-2355

The single server queue with mixing dependencies

Youri Raaijmakers[1], Hansjoerg Albrecher[2] and Onno Boxma[1 3]
(y.raaijmakers@student.tue.nl; hansjoerg.albrecher@unil.ch; o.j.boxma@tue.nl)

February 6, 2017

[1] Department of Mathematics and Computer Science, Eindhoven University of Technology, P.O. Box 513, 5600 MB Eindhoven, The Netherlands

[2] Department of Actuarial Science, Faculty of Business and Economics, University of Lausanne, UNIL-Dorigny, CH-1015 Lausanne, Switzerland and Swiss Finance Institute, Lausanne

[3] The research of Onno Boxma was partly funded by the NWO Gravitation Project NETWORKS, Grant Number 024.002.003

Abstract

We study a single server queue, where dependence is introduced between the service times, or between the inter-arrival times, or both between the service times and the inter-arrival times. This dependence arises via mixing, i.e., a parameter pertaining to the distribution of the service times, or of the interarrival times, is itself considered to be a random variable. We give a duality result between such queueing models and corresponding insurance risk models, for which the respective dependence structures have been studied before. For a number of examples we provide exact expressions for the waiting time distribution, and compare these to the ones for the standard M/M/1 queue. We also investigate the effect of dependence and derive first order asymptotics for some of the obtained waiting time tails. Finally, we discuss this dependence concept for the waiting time tail of the G/M/1 queue.

1 Introduction

In the literature on the single server queue, it is almost always assumed that the arrival process is a renewal process, and usually even a Poisson process [16]. In recent years, there has been a growing inclination to move away from these assumptions. There are several reasons for this. Firstly, the arrival process may vary in the course of a time period. This has led to the study of queues with time-inhomogeneous Poisson arrivals [21], and of Markov-modulated arrivals, in which the arrival process is Poisson with rate λ_i when some underlying Markov process is in state $i \in I$ (cf. [8, Ch.XI]). Secondly, the (homogeneous or nonhomogeneous) Poisson assumption not always aligns well with actual data; see, e.g., [24] and references therein. More specifically, Poisson processes sometimes underestimate the variability of the input stream. This has, among other reasons, given rise to Cox input processes, which are Poisson processes of which the time-dependent intensity, say $\Lambda(t)$, itself is a stochastic process. The variance of the number of arrivals of a Cox process in a given interval is larger than the mean (while they are equal for the Poisson process); this phenomenon is usually called overdispersion. In a few recent papers, $\{\Lambda(t), t \geq 0\}$ was taken to be a shot-noise process; see [5] for an example in insurance mathematics and [25] for an example in queueing. For a general recent treatment of the topic and a discussion of estimation procedures from data, see [6, Ch.V].

As stated in [23], call center planners differentiate between variables such as day of week, holidays and marketing activities, to obtain an estimate for the arrival rate; but still one often observes overdispersion. For that reason, Jongbloed and Koole [23] suggest to use a Poisson *mixture model*, i.e., they let the arrival rate of the Poisson process itself be a random variable Λ with some distribution $F_\Lambda(\cdot)$, mixing a Poisson distribution over $F_\Lambda(\cdot)$:

$$\mathbb{P}_\Lambda(X = j) = \int_0^\infty \frac{\lambda^j}{j!} \mathrm{e}^{-\lambda} \mathrm{d}F_\Lambda(\lambda), \quad j = 0, 1, \dots .$$

The resulting mixed Poisson process is a special case of a Cox process with constant stochastic intensity $\Lambda(t) \equiv \Lambda$. Jongbloed and Koole then proceed to estimate the mixing distribution, followed by a case study. A recent example of the application of such a mixing procedure to the optimal staffing of large-scale service centers is given in Section 4 of [29].

Queueing models with mixing also form the topic of the present paper. To consider mixing distributions is not that common in the queueing literature, but more so in the financial and insurance literature; e.g., Bühlmann [13] already considered them in the context of credibility-based dynamic premium rules. Mixing also has the interesting and useful feature of introducing *dependence* between successive interarrival times – or between successive service times or successive claim sizes, if mixing is done over a parameter pertaining to those random variables. In a recent paper [7], explicit ruin formulas are obtained for a generalization of the Cramér-Lundberg insurance risk model to the case in which one of the key parameters (claim size rate or arrival rate) is itself a random variable, and in which mixing takes place over its distribution. In that paper, much emphasis is placed on the observation that successive claim sizes, or arrival intervals, now are no longer independent, and it is demonstrated that the resulting dependence can be described by an Archimedean survival copula.

In the present paper we are going to consider an analogous kind of mixing for single server queues: Suppose that successive interarrival intervals $A_1, A_2, \ldots$ are exponentially distributed, with rate X, and assume that X is itself a random variable with distribution $F_X(\cdot)$. Then, for all $n = 1, 2, \ldots,$

$$\mathbb{P}(A_1 > a_1, \ldots, A_n > a_n | X = \lambda) = \prod_{i=1}^{n} \mathrm{e}^{-\lambda a_i}, \quad a_1, \ldots, a_n \geq 0,$$

and

$$\mathbb{P}(A_1 > a_1, \ldots, A_n > a_n) = \int_{\lambda=0}^{\infty} \mathrm{e}^{-\lambda \sum_{i=1}^{n} a_i} \mathrm{d}F_X(\lambda), \quad a_1, \ldots, a_n \geq 0.$$

While $A_1, A_2, \ldots$ are conditionally independent given $X = \lambda$, unconditionally they are *dependent.* It should be noticed that X is chosen once and for all, so that during the entire period of interest we have the same realization of X.

Instead of mixing with respect to the arrival rate, we could also mix with respect to a parameter pertaining to the service times. If the successive service times are denoted by $B_1, B_2, \ldots$, then the joint distribution of $B_1, \ldots, B_n$ is for all $n = 1, 2, \ldots$ given by

$$\mathbb{P}(B_1 > b_1, \ldots, B_n > b_n) = \int_{\mu=0}^{\infty} \prod_{i=1}^{n} \mathbb{P}(B_i > b_i | X = \mu) \mathrm{d}F_X(\mu), \quad b_1, \ldots, b_n > 0. \quad (1)$$

In [7, Prop 2.1], see also its Remark 2.6, it is shown that this dependence model can equivalently be described by having marginal service times $B_1, B_2, ...$ that are completely monotone, with a dependence structure due to an Archimedean survival copula with generator $\phi = (\tilde{F}_X)^{-1}$, where $\tilde{F}_X$ denotes the Laplace transform of the service rate.

The main goal of the present paper is to explore this mixing principle and the corresponding dependence structure in relatively simple single server queueing models.

For such models, we shall be able to obtain *explicit* expressions for the waiting time distribution, which allows us to get insight into the question how mixing affects the waiting time behaviour.

Before finishing this section with an outline of the rest of the paper, we mention a few more queueing papers in which mixing is used. In an unpublished report, Cohen [17] studied an $M/G/1$ queue with Pareto distributed service times: $\mathbb{P}(B > t) = \delta(\frac{\theta}{\theta+t})^{\nu}$. His goal was to construct a service time distribution that is on the one hand heavy-tailed (he took $1 < \nu < 2$), but that on the other hand also has a tractable expression and relatively simple Laplace transform. To reach that goal, he mixed the parameter of the Pareto distribution with respect to a Gamma distribution. For $\nu = \frac{3}{2}$, that gave rise to a service time distribution that is expressed in a complementary error function. Remarkably, the waiting time distribution in this case also can be obtained explicitly, and it involves two Erfc$(\cdot)$ functions, with arguments $(1-\sqrt{\rho})\sqrt{t}$ and $(1+\sqrt{\rho})\sqrt{t}$, where $\rho < 1$ is the traffic load of the server. In [11, 12] the latter shape was one of the pointers for the authors to realize that one can obtain a heavy-traffic limit ($\rho \to 1$) by scaling the waiting time by $(1-\rho)^2$ (more generally: by $(1-\rho)^{1/(\nu-1)}$) instead of by $1-\rho$, the appropriate scaling factor in the case of a finite second moment of the service times.

Abate, Choudhury and Whitt [2] had performed a kind of mirror operation, compared to [17]; they had taken a Pareto mixture of a Gamma distribution. In [1], Abate and Whitt generalize the class studied by Cohen [17], considering two classes of so-called Beta mixtures of exponentials. Their focus is not on waiting times; they aim for explicit representations of service time distributions, and of their Laplace transforms, moments and asymptotics.

The rest of the paper is organized as follows. Section 2 provides a duality result between queueing and insurance risk models with mixing. Section 3 illustrates the idea in detail for the M/M/1 queueing model. In Section 4 we then allow for general inter-arrival times, deriving an explicit expression for the waiting time distribution for some particular cases. Section 5 contains suggestions for further research.

2 A duality result

There are close connections between the classical single server queue and the classical Sparre-Andersen insurance risk model. In particular, there are duality results that relate survival and ruin probabilities in the insurance risk model to waiting time distributions in the 'corresponding' queueing model. Our goal in this section is to extend one such duality result to the case of mixing outlined in Section 1.

Let us first briefly sketch the standard duality result, referring to [9], pages 45 and 161, for more detail. Let $R(t)$ be the surplus of an insurance company, with initial surplus $R(0) = x$, and with premium rate c which is normalized to be 1. Claims arrive with interarrival times $A_1, A_2, \dots$, which are i.i.d.; the claim sizes $C_1, C_2, \dots$ are also i.i.d., and independent of the interarrival times. In the 'corresponding' single server queue, customers arrive with interarrival times $A_1, A_2, \dots$, which are i.i.d., and require service times $C_1, C_2, \dots$ which are also i.i.d., and independent of the interarrival times; and the server works at unit speed.

The probability $P_n(x)$ that the surplus stays positive after the first n claims (the so-called survival probability) is

$$P_n(x) = \mathbb{P}(x+A_1-C_1 > 0, x+\sum_{j=1}^{2}(A_j-C_j) > 0, \ldots, x+\sum_{j=1}^{n}(A_j-C_j) > 0), \quad n = 1, 2, \ldots. \tag{2}$$

Denoting the waiting time of the nth arriving customer by W_n, $n = 1, 2, \ldots$, and using the well-known recursion

$$W_{n+1} = \max(0, W_n + C_n - A_n), \quad n = 1, 2, \ldots,$$

one can easily prove that

$$\mathbb{P}(W_{n+1} < x | W_1 = 0) = \mathbb{P}(C_n - A_n < x, \sum_{j=n-1}^{n} (C_j - A_j) < x, \ldots, \sum_{j=1}^{n}(C_j - A_j) < x). \tag{3}$$

It is immediately obvious from (2) and (3) that

$$P_n(x) = \mathbb{P}(W_{n+1} < x | W_1 = 0), \tag{4}$$

since $C_j - A_j$ has the same distribution as $C_{n-j} - A_{n-j}$, $j = 1, \ldots, n$. Furthermore, one also has

$$\lim_{n\to\infty} P_n(x) = \lim_{n\to\infty} \mathbb{P}(W_{n+1} < x), \quad x > 0. \tag{5}$$

In fact, these duality results remain valid if all C_j and A_j are dependent in the same way, since one only needs to consider the differences $C_j - A_j$. See [10] for some further extensions to multi-dimensional queues and insurance risk models.

If mixing with respect to a service time parameter X (which could for example be the mean service time) with distribution $F_X(\cdot)$ is applied to the service times in the queueing model, and similarly to the claim sizes in the insurance risk model, then these service times (or claim sizes) become dependent, having the joint distribution (cf. (1))

$$\mathbb{P}(C_1 > c_1, \ldots, C_n > c_n) = \int_{\mu=0}^{\infty} \prod_{i=1}^{n} \mathbb{P}(C_i > c_i | X = \mu) \mathrm{d}F_X(\mu), \quad c_1, \ldots, c_n > 0.$$

However, the random variables $C_1, \ldots, C_n$ now are exchangeable (cf. Section 7.2 of [28]), and hence one still has (4), while also (5) remains valid; see also Section 7.1 of [28]. A similar reasoning holds when there is mixing with respect to the interarrival times, or both to service times/claim sizes and interarrival times. Hence we have:

Theorem 2.1. *The duality results (4) and (5) remain valid for the Sparre Andersen insurance risk model and the 'corresponding' queueing model, when mixing is applied to the service times/claim sizes and/or the interarrival times.*

The implication of this theorem is, that (i) certain results for survival probabilities (or ruin probabilities) in [7] immediately translate into waiting time results, and that (ii) some waiting time results which we shall derive in subsequent sections, immediately translate into results for survival probabilities.

3 An M/M/1 queue with dependence

We first consider the following classical model, viz., the $M/M/1$ queue. Customers arrive according to a Poisson process with intensity λ and each customer requires an exponentially distributed service time with mean $1/\mu$. If the service times are independent it is well known that

$$\mathbb{P}(W > x) = \frac{\lambda}{\mu} e^{-(\mu-\lambda)x}, \qquad x \geq 0, \tag{6}$$

where W denotes the waiting time in stationarity of an arbitrary customer. This steady-state distribution exists if $\rho = \frac{\lambda}{\mu} < 1$ and for $\rho \geq 1$ we know $\mathbb{P}(W > x) = 1$ for all finite $x > 0$.

Now let us consider an extension to this model. We assume that the service rate is itself not a constant but a random variable, that takes the same value for all service times. Successive service times $B_1, B_2, ...$ now are no longer independent. However, they are conditionally independent, as for each n

$$\mathbb{P}(B_1 > b_1, ..., B_n > b_n | X = \mu) = \prod_{k=1}^{n} e^{-\mu b_k}, \tag{7}$$

where the service rate X is a *random variable*; given $X = \mu$, the service times are independent and exponentially distributed with mean $1/\mu$. Note that now, in contrast to the standard M/M/1 model, the marginal distribution of B_k is no longer exponential. Note that one can view the unconditional probability $\mathbb{P}(B_i > b)$ as $\mathbb{P}(\hat{B}_i/X > b)$, where $\hat{B}_i \sim \text{Exp}(1)$, which identifies $B_1, \ldots, B_n$ as an L_1 Dirichlet sequence, see [18] for a general discussion from this perspective.

If the *mixing* cumulative density function (cdf) is denoted by F_X, then for the dependence model (7) the tail of the waiting time is given by

$$\mathbb{P}(W_\mu > x) = \int_0^\infty \mathbb{P}(W > x) \mathrm{d}F_X(\mu), \qquad x \geq 0, \tag{8}$$

where W_μ denotes the waiting time of a customer, and where $\mathbb{P}(W > x)$ is given by (6). In Equation (8) the random variable X is chosen once and for all. This means that during the period of interest we take the same realization of X.

Next, we give examples regarding mixing over the service rate.

3.1 Mixing with respect to the service rate μ

Case I.1. Let X be a random variable with density function

$$f_X(\mu) = \begin{cases} 0 & \text{, if } 0 < \mu \leq \lambda, \\ \alpha e^{-\alpha(\mu-\lambda)} & \text{, if } \mu > \lambda, \end{cases} \tag{9}$$

where $\alpha > 0$.

Note that this choice for the density function satisfies the stability condition $\rho = \lambda\mathbb{E}[B] < 1$ for each possible realization $X = \mu$, because the density has only mass on $\mu > \lambda$. The tail of the service times is given by

$$\begin{aligned}\bar{F}_B(x) &= \int_0^\infty e^{-\mu x} f_X(\mu)\mathrm{d}\mu \\ &= \int_\lambda^\infty e^{-\mu x}\alpha e^{-\alpha(\mu-\lambda)}\mathrm{d}\mu \\ &= \frac{\alpha}{\alpha+x}e^{-\lambda x}.\end{aligned}$$

An immediate result is that (see below (1)),

$$\phi^{-1}(t) = \frac{\alpha}{\alpha+t}e^{-\lambda t},$$

and with this expression for the inverse generator we can get an explicit expression for the covariance under the dependence structure due to the Archimedean survival copula:

$$\mathrm{Cov}(B_i, B_j) = \alpha\left(\frac{1}{\lambda} + \alpha e^{-2\alpha\lambda}\left(e^{3\alpha\lambda} - \mathrm{Ei}\left[-\alpha\lambda\right]\right)\mathrm{Ei}\left[-\alpha\lambda\right]\right),$$

where $\mathrm{Ei}[x] = -\int_{-x}^\infty \frac{e^{-t}}{t}\mathrm{d}t$ is the exponential integral, cf. [3].

From (8) it follows that,

$$\begin{aligned}\mathbb{P}(W_\mu > x) &= \int_\lambda^\infty \alpha e^{-\alpha(y-\lambda)}\frac{\lambda}{y}e^{-(y-\lambda)x}\mathrm{d}y \\ &= \alpha\lambda\int_0^\infty \frac{e^{-(\alpha+x)z}}{z+\lambda}\mathrm{d}z \\ &= -\alpha\lambda e^{(x+\alpha)\lambda}\mathrm{Ei}[-\alpha\lambda], \quad x \geq 0. \qquad (10)\end{aligned}$$

Once an expression for the tail of the waiting time is found, one can also get an expression for the mean waiting time and the Laplace-Stieltjes Transform (LST). In particular,

$$\begin{aligned}\mathbb{E}[W_\mu] &= \int_0^\infty \mathbb{P}(W_\mu > x)\mathrm{d}x \\ &= \int_0^\infty \alpha\lambda\int_0^\infty \frac{e^{-(\alpha+x)z}}{z+\lambda}\mathrm{d}z\mathrm{d}x \\ &= \alpha\int_0^\infty e^{-\alpha z}\frac{\lambda}{z(z+\lambda)}\mathrm{d}z \\ &= \alpha\int_0^\infty e^{-\alpha z}\left(\frac{1}{z} - \frac{1}{z+\lambda}\right)\mathrm{d}z \\ &= \infty,\end{aligned}$$

and,

$$\begin{aligned}\mathbb{E}[e^{-sW_\mu}] &= \int_0^\infty e^{-sx}\alpha\lambda \int_0^\infty \frac{z}{(z+\lambda)} e^{-(\alpha+x)z}\mathrm{d}z\mathrm{d}x + 1 - \int_0^\infty \alpha\lambda\frac{e^{-\alpha z}}{z+\lambda}\mathrm{d}z \\ &= \alpha\lambda\int_0^\infty \frac{ze^{-\alpha z}}{(s+z)(z+\lambda)}\mathrm{d}z + 1 - \int_0^\infty \alpha\lambda\frac{e^{-\alpha z}}{z+\lambda}\mathrm{d}z.\end{aligned}$$

Case I.2. In the previous example we considered a density function that satisfies the stability condition. However, one may also wonder what happens if a density function that violates this stability condition is taken. If X is Gamma distributed with density

$$f_X(\mu) = \frac{\beta^\alpha}{\Gamma(\alpha)}\mu^{\alpha-1}e^{-\beta\mu}, \qquad \mu > 0,$$

it follows that the tail of the service times equals

$$\bar{F}_B(x) = \int_0^\infty e^{-x\mu}\frac{\beta^\alpha}{\Gamma(\alpha)}\mu^{\alpha-1}e^{-\beta\mu}\mathrm{d}\mu = \left(\frac{\beta}{\beta+x}\right)^\alpha,$$

which is the tail of a Pareto distribution. Further, the generator of the Archimedean survival copula is simply given by

$$\phi(t) = \beta(t^{-1/\alpha} - 1),$$

and one can observe that this generator corresponds to the Clayton copula (see Example 2.3 of [7]). For the covariance it follows that for $\alpha > 2$,

$$\mathrm{Cov}(B_i, B_j) = \frac{\beta^2}{(\alpha-2)(\alpha-1)^2}, \tag{11}$$

and the Spearman's rank correlation coefficient (or Spearman's rho), denoted by ρ_S (cf. [26]; ρ_S is a nonparametric measure of rank correlation) is given by

$$\rho_S(B_i, B_j) = 12\int_0^1\int_0^1 \left(\frac{1}{a^{-1/\alpha} + b^{-1/\alpha} - 1}\right)^\alpha \mathrm{d}a\mathrm{d}b - 3, \tag{12}$$

and only depends on α. Spearman's rho assesses monotonic relationships; a perfect Spearman rank correlation of $+1$ or -1 occurs when each of the variables is a perfect monotonic function of the other. Figure 1 depicts the value of ρ_S as a function of α graphically.

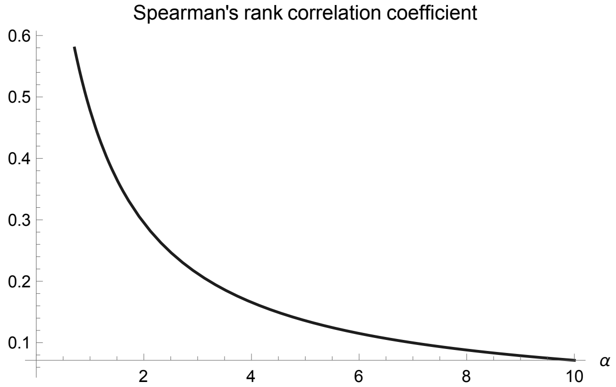

Figure 1: $\rho_S(B_i, B_j)$ for the dependent model with Pareto distributed service times as a function of α.

From (8), we now get

$$\begin{aligned}\mathbb{P}(W_\mu > x) &= \int_0^\infty \mathbb{P}(W > x)\mathrm{d}F_X(\mu)\\ &= \int_0^\lambda \mathbb{P}(W > x)\mathrm{d}F_X(\mu) + \int_\lambda^\infty \mathbb{P}(W > x)\mathrm{d}F_X(\mu)\\ &= \mathbb{P}(X \le \lambda) + \int_\lambda^\infty \frac{\lambda}{y} e^{-(y-\lambda)x} \frac{\beta^\alpha}{\Gamma(\alpha)} y^{\alpha-1} e^{-\beta y}\mathrm{d}y\\ &= 1 - \frac{\Gamma(\alpha,\beta\lambda)}{\Gamma(\alpha)} + \lambda e^{\lambda x}\left(\frac{\beta}{\beta+x}\right)^{\alpha-1} \beta \frac{\Gamma(-1+\alpha,(\beta+x)\lambda)}{\Gamma(\alpha)},\end{aligned} \tag{13}$$

where $\Gamma(\alpha, x) = \int_x^\infty w^{\alpha-1}e^{-w}\mathrm{d}w$ is the incomplete Gamma function (see also [7, Example 2.3], where this expression was already established for the ruin probability with initial surplus x in the respective risk model).

Like in [20], where the asymptotic behavior of the ruin probability of [7] was studied, we can establish the first order asymptotic behavior for the tail of the waiting time in the same way. Indeed, using

$$\Gamma(\alpha, x) \sim x^{\alpha-1}e^{-x}\left(1 + \frac{\alpha-1}{x} + \frac{(\alpha-1)(\alpha-2)}{x^2} + o\left(\frac{1}{x^2}\right)\right), \quad x \to \infty,$$

(see e.g. [3, Equ.6.5.32]), one gets

$$\begin{aligned}\mathbb{P}(W_\mu > x) &= 1 - \frac{\Gamma(\alpha,\beta\lambda)}{\Gamma(\alpha)} + \lambda e^{\lambda x}\left(\frac{\beta}{\beta+x}\right)^{\alpha-1} \beta \frac{\Gamma(-1+\alpha,(\beta+x)\lambda)}{\Gamma(\alpha)}\\ &\sim 1 - \frac{\Gamma(\alpha,\beta\lambda)}{\Gamma(\alpha)} + e^{-\beta\lambda}\frac{(\lambda\beta)^\alpha}{\lambda\Gamma(\alpha)}\left(\frac{1}{\beta+x} + \frac{\alpha-2}{\lambda(\beta+x)^2} + o\left(\frac{1}{x^2}\right)\right), \quad x \to \infty.\end{aligned} \tag{14}$$

From this expression one immediately sees that W_μ has an atom at infinity with probability mass $1 - \frac{\Gamma(\alpha,\beta\lambda)}{\Gamma(\alpha)}$ (which is the probability to violate the stability condition

and corresponds to the event of ruin in the analogous risk theory case).
In Figure 2 we compare the exact expression (13) with the first-order approximation (14) on a log-scale for the parameter values $\lambda = 3$, $\alpha = 4$ and $\beta = 1$.

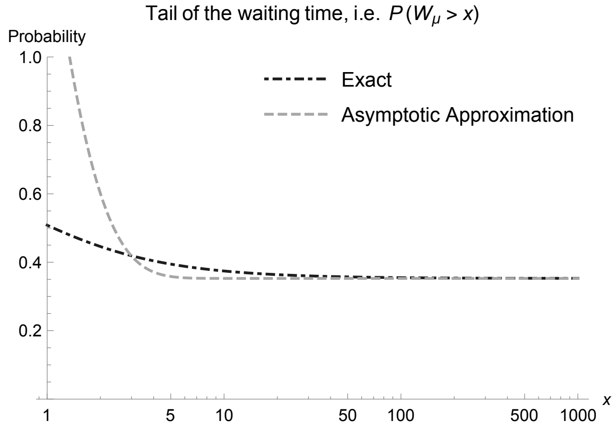

Figure 2: $\mathbb{P}(W_\mu > x)$ and its first-order asymptotic approximation with Pareto distributed service times ($\lambda = 3$, $\alpha = 4$ and $\beta = 1$).

Case I.3. Like in [7, Example 2.4], we next consider Lévy distributed service rates with

$$f_X(\mu) = \frac{\alpha}{2\sqrt{\pi\mu^3}} e^{-\frac{\alpha^2}{4\mu}}, \qquad \mu > 0.$$

Then the tail of the service times is

$$\bar{F}_B(x) = \int_0^\infty e^{-x\mu} \frac{\alpha}{2\sqrt{\pi\mu^3}} e^{-\frac{\alpha^2}{4\mu}} \mathrm{d}\mu = e^{-\alpha\sqrt{x}},$$

from which one sees that the marginal service times are Weibull distributed and the dependence structure is given by a Gumbel copula with generator

$$\phi(t) = (-(\ln t)/\alpha)^2.$$

With this dependence structure the covariance is given by

$$\mathrm{Cov}(B_i, B_j) = \frac{8}{\alpha^4}, \tag{15}$$

and Spearman's rho is for each α equal to

$$\rho_S(B_i, B_j) = 12 \int_0^1 \int_0^1 e^{-\sqrt{\log(a)^2 + \log(b)^2}} \mathrm{d}a\mathrm{d}b - 3 = 0.682. \tag{16}$$

Interestingly, the expression for Spearman's rho ρ_S in (12) for the dependency model with Gamma mixing over the service times equals ρ_S in (16) for the Lévy mixing when $\alpha = 0.5$.

For the tail of the waiting times we get,

$$\begin{aligned}\mathbb{P}(W_\mu > x) &= \int_0^\infty \mathbb{P}(W > x)\mathrm{d}F_X(\mu)\\ &= \mathbb{P}(X \leq \lambda) + \int_\lambda^\infty \frac{\lambda}{y} e^{-(y-\lambda)x} \frac{\alpha}{2\sqrt{\pi y^3}} e^{-\frac{\alpha^2}{4y}} \mathrm{d}y\\ &= \operatorname{Erfc}\left(\frac{\alpha}{2\sqrt{\lambda}}\right) - \frac{2\sqrt{\lambda}}{\alpha\sqrt{\pi}} e^{-\frac{\alpha^2}{4\lambda}}\\ &\quad + \lambda e^{\lambda x} \frac{1+\alpha\sqrt{x}}{\alpha^2} e^{-\sqrt{x}\alpha} \cdot \operatorname{Erfc}\left(\sqrt{\lambda x} - \frac{\alpha}{2\sqrt{\lambda}}\right)\\ &\quad + \lambda e^{\lambda x} \frac{-1+\alpha\sqrt{x}}{\alpha^2} e^{\sqrt{x}\alpha} \cdot \operatorname{Erfc}\left(\sqrt{\lambda x} + \frac{\alpha}{2\sqrt{\lambda}}\right), \end{aligned} \tag{17}$$

where $\operatorname{Erfc}(x) = 1 - \operatorname{Erf}(x) = \frac{2}{\sqrt{\pi}} \int_x^\infty e^{-w^2} \mathrm{d}w$ is the error function, cf. also [7, Ex.2.4].

Equivalently, using properties of the (generalized) gamma function , Equation (17) can be written as

$$\begin{aligned}\mathbb{P}(W_\mu > x) &= \operatorname{Erfc}\left(\frac{\alpha}{2\sqrt{\lambda}}\right) + \frac{\lambda\sqrt{x}}{\alpha} e^{\lambda x} \left(-\frac{2}{\sqrt{\pi\lambda x}} e^{-\lambda x - \frac{\alpha^2}{4\lambda}} \right.\\ &\quad + \left(1 + \frac{1}{\alpha\sqrt{x}}\right) e^{\alpha\sqrt{x}} \cdot \operatorname{Erfc}\left(\sqrt{\lambda x} - \frac{\alpha}{2\sqrt{\lambda}}\right)\\ &\quad \left. + \left(1 - \frac{1}{\alpha\sqrt{x}}\right) e^{-\alpha\sqrt{x}} \cdot \operatorname{Erfc}\left(\sqrt{\lambda x} + \frac{\alpha}{2\sqrt{\lambda}}\right)\right)\\ &= \operatorname{Erfc}\left(\frac{\alpha}{2\sqrt{\lambda}}\right) + \frac{\alpha\lambda\sqrt{x^3}}{2\sqrt{\pi}} e^{\lambda x} \cdot \Gamma\left(-\frac{3}{2}, \lambda x, \frac{\alpha^2 x}{4}\right), \end{aligned} \tag{18}$$

where $\Gamma(\zeta, y, d)$ denotes the generalized incomplete Gamma function

$$\Gamma(\zeta, y, d) = \int_y^\infty t^{\zeta-1} e^{-t-\frac{d}{t}} \mathrm{d}t, \qquad y \geq 0, \quad \zeta \in \mathbb{R}, \quad d > 0.$$

In [14] it was shown that

$$\begin{aligned}\Gamma(\zeta, y, d) \sim y^{\zeta-1} e^{-y-\frac{d}{y}} &\left(1 + \frac{d + (\zeta-1)y}{y^2}\right.\\ &\left. + \frac{d^2 + (2\zeta-4)dy + (\zeta^2 - 3\zeta + 2)y^2}{y^4} + o\left(\frac{1}{y^2}\right)\right), \quad y \to \infty,\end{aligned}$$

so that we get from Equation (18) that

$$\begin{aligned}\mathbb{P}(W_\mu > x) \sim \operatorname{Erfc}\left(\frac{\alpha}{2\sqrt{\lambda}}\right) &+ \frac{\alpha}{2\sqrt{\pi\lambda^3}} e^{-\frac{\alpha^2}{4\lambda}}\\ &\cdot \left(\frac{1}{x} + \frac{\alpha^2 - 10\lambda}{4\lambda^2 x^2} + o\left(\frac{1}{x^2}\right)\right), \quad x \to \infty.\end{aligned}$$

In Figure 3 we compare the asymptotic approximation with the exact expression for the tail of the waiting time given in (17) for the values $\alpha = 1$ and $\lambda = 3$.

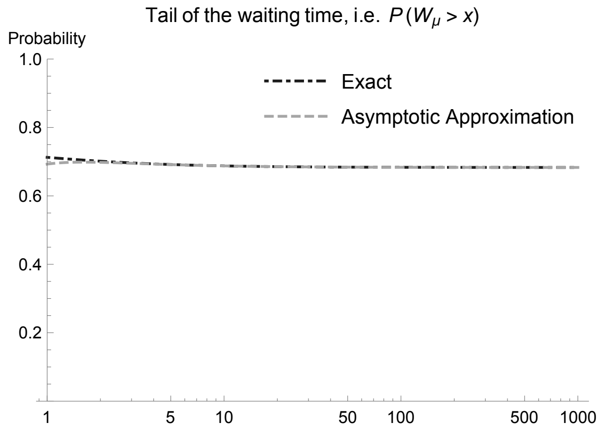

Figure 3: $\mathbb{P}(W_\mu > x)$ and its asymptotic approximation with Weibull distributed service times.

3.2 Mixing with respect to the arrival rate λ

The mixing idea of Section 3.1 can also be used for mixing over the Poisson arrivals with rate λ. If the random variable Λ with cdf F_Λ denotes the arrival intensity, then the resulting waiting time tail is

$$\mathbb{P}(W_\lambda > x) = \int_0^\infty \mathbb{P}(W > x)\mathrm{d}F_\Lambda(\lambda), \qquad x \geq 0. \tag{19}$$

Case II.1. Consider the density function

$$f_\Lambda(\lambda) = \begin{cases} \frac{\alpha e^{-\alpha\lambda}}{1-e^{-\alpha\mu}}, & \text{if } 0 < \lambda < \mu, \\ 0, & \text{if } \lambda \geq \mu, \end{cases} \tag{20}$$

for the arrival intensity, for which the stability condition holds again. Then the resulting marginal tail of the generic inter-arrival time A is

$$\bar{F}_A(t) = \int_0^\infty e^{-\lambda t} f_\Lambda(\lambda)\mathrm{d}\lambda = \frac{\alpha}{\alpha+t}\frac{1-e^{-(\alpha+t)\mu}}{1-e^{-\alpha\mu}}.$$

The resulting tail of the waiting time is

$$\begin{aligned}\mathbb{P}(W_\lambda > x) &= \int_0^\mu \frac{y}{\mu} e^{-(\mu-y)x}\frac{\alpha e^{-\alpha y}}{1-e^{-\alpha\mu}}\mathrm{d}y \\ &= \frac{\alpha e^{-\mu x}}{\mu(1-e^{-\alpha\mu})}\left[\frac{1-e^{(x-\alpha)\mu}\left((\alpha-x)\mu+1\right)}{(x-\alpha)^2}\right].\end{aligned} \tag{21}$$

Here the probability that in stationarity a customer experiences a waiting time is

$$\mathbb{P}(W_\lambda > 0) = \frac{1}{\alpha\mu} - \frac{e^{-\alpha\mu}}{1-e^{-\alpha\mu}}.$$

Case II.2. Like in Case I.2, where we took the service rate Gamma distributed, we now consider the example where the arrival intensity is Gamma distributed, i.e.

$$f_\Lambda(\lambda) = \frac{\beta^\alpha}{\Gamma(\alpha)}\lambda^{\alpha-1}e^{-\beta\lambda}, \qquad \lambda > 0,$$

see also [7, Example 3.1] for the respective insurance risk model. Then for the tail of the inter-arrival times we get

$$\bar{F}_A(x) = \int_0^\infty e^{-x\lambda}\frac{\beta^\alpha}{\Gamma(\alpha)}\lambda^{\alpha-1}e^{-\beta\lambda}\mathrm{d}\lambda = \left(\frac{\beta}{\beta+x}\right)^\alpha,$$

so that A is Pareto distributed (for the covariance and the Spearman's rho we correspondingly get the same expressions as in (11) and (12)).

From the duality result of Section 2 in combination with [7, Example 3.1], or via a straightforward integration, it follows that

$$\begin{aligned}\mathbb{P}(W_\lambda > x) &= \int_0^\infty \mathbb{P}(W > x)\mathrm{d}F_\Lambda(\lambda)\\ &= \int_\mu^\infty \mathbb{P}(W > x)\mathrm{d}F_\Lambda(\lambda) + \int_0^\mu \mathbb{P}(W > x)\mathrm{d}F_\Lambda(\lambda)\\ &= \frac{\Gamma(\alpha,\beta\mu)}{\Gamma(\alpha)} + \frac{1}{\mu}e^{-x\mu}\left(\frac{\beta}{\beta-x}\right)^\alpha \frac{1}{\beta-x}\left(\alpha - \frac{\Gamma(1+\alpha,(\beta-x)\mu)}{\Gamma(\alpha)}\right). \qquad (22)\end{aligned}$$

Case II.3. If Λ is Lévy distributed with density function

$$f_\Lambda(\lambda) = \frac{\alpha}{2\sqrt{\pi\lambda^3}}e^{-\frac{\alpha^2}{4\lambda}}, \qquad \lambda > 0, \qquad (23)$$

then

$$\bar{F}_A(x) = \int_0^\infty e^{-x\lambda}\frac{\alpha}{2\sqrt{\pi\lambda^3}}e^{-\frac{\alpha^2}{4\lambda}}\mathrm{d}\lambda = e^{-\alpha\sqrt{x}},$$

so A is Weibull distributed (see also [7, Example 3.2]). The expressions for the covariance and Spearman's rho follow again from (15) and (16).

From duality, in combination with [7, Example 3.2], or via (19), we get

$$\begin{aligned}\mathbb{P}(W_\lambda > x) = \mathrm{Erfc}\left(\frac{\alpha}{2\sqrt{\mu}}\right) + \frac{i\alpha e^{-\mu x}}{4\mu\sqrt{x}}\left(e^{i\alpha\sqrt{x}}\cdot \mathrm{Erfc}\left(\frac{\alpha}{2\sqrt{\mu}} + i\sqrt{\mu x}\right)\right.\\ \left. -e^{-i\alpha\sqrt{x}}\cdot \mathrm{Erfc}\left(\frac{\alpha}{2\sqrt{\mu}} - i\sqrt{\mu x}\right)\right).\end{aligned}$$

With the identities of the Error function, given in [4], it can be shown that the imaginary part of this equation is equal to zero.

3.3 Mixing with respect to the traffic load ρ

Another possibility is to mix over the traffic load ρ. Here we describe two methods. The first one is to mix over μ but keep $\rho < 1$ fixed. This means that both μ and λ are determined by the mixing distribution. The second method, which is in line with Sections 3.1 and 3.2, is to mix over ρ and keep μ fixed. Let us first rewrite (6) as

$$\mathbb{P}(W > x) = \rho e^{-\mu(1-\rho)x}, \qquad x \geq 0.$$

Method 1: Let X be Gamma(α, β) distributed. Note that the stability condition is always satisfied because $\rho < 1$ is fixed. For the tail of the waiting time we then get the simple formula

$$\begin{aligned}\mathbb{P}(W_\mu > x) &= \int_0^\infty \rho e^{-y(1-\rho)x} \frac{\beta^\alpha}{\Gamma(\alpha)} y^{\alpha-1} e^{-\beta y} \mathrm{d}y \\ &= \rho \left(\frac{\beta}{\beta + (1-\rho)x} \right)^\alpha . \end{aligned} \tag{24}$$

Method 2: Let the random variable R denote the traffic load, and take the service rate μ fixed. Then consider the example where Λ is Unif$[0, \mu]$ distributed, from which it follows that the traffic load is uniformly distributed between 0 and 1, i.e., R is Unif$[0, 1]$ distributed. Then the tail of the waiting time is

$$\mathbb{P}(W_\lambda > x) = \int_0^1 \rho e^{-\mu(1-\rho)x} \mathrm{d}\rho = \frac{1}{(\mu x)^2} \left[e^{-\mu x} - 1 + \mu x \right] .$$

Note that in this example it is not possible to mix over μ for fixed λ and obtain the same result.

3.4 Sensitivity with respect to dependence

In the previous sections we provided explicit formulas for the tail of the waiting time in an M/M/1 queueing model with dependence. In this section we would like to assess the effects of the introduced dependence on the waiting time tail, where we use the independent M/M/1 model given by (6) with parameters $\lambda = 3$ and $\mu = 4$ as a benchmark. We restrict the analysis to a few examples.

Case I.2: Mixing over the service rate. Fix $\lambda = 3$ and for each choice of β (which is the parameter driving the dependence) we take α in such a way that $\mathbb{E}[X] = \alpha/\beta = 4$ remains constant, so that we can better compare the effect of dependence. The variance of X then is α/β^2 and decreases with increasing β, going to zero for $\beta \to \infty$ (in which case the service rate is fixed at 4, which corresponds to the case of independence). Figure 4 depicts the resulting waiting time tail for different values of β (cf. (13)).

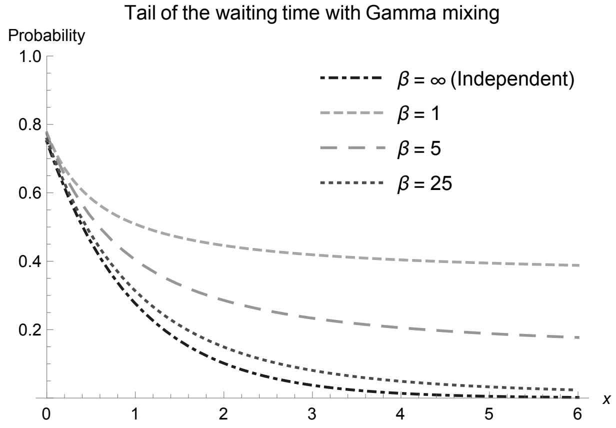

Figure 4: Tail of the waiting time for the independent model and the dependent model (13) with equal expected service rate.

One observes increasing dependence increases the waiting time tail for all values of x, which in fact suggests a stochastic ordering.

One can also assess the convergence of $\mathbb{P}(W_\mu > x)$ towards the limit value for increasing β. To that end, note that

$$\begin{aligned} \frac{\Gamma(a, a\zeta)}{\Gamma(a)} &= \frac{1}{2} \cdot \mathrm{Erfc}\left[\eta\sqrt{a/2}\right] + R_a(\eta), \\ R_a(\eta) &\sim \frac{e^{-\frac{1}{2}a\eta^2}}{\sqrt{2\pi a}} \sum_{n=0}^{\infty} \frac{c_n(\eta)}{a^n}, \quad a \to \infty, \end{aligned}$$

where $\eta = \sqrt{2(\zeta - 1 - \ln\zeta)}$ when $\zeta > 1$ and $\eta = -\sqrt{2(\zeta - 1 - \ln\zeta)}$ when $\zeta < 1$ (cf. [22]). Then

$$\begin{aligned} \mathbb{P}(W_\mu > x) &= 1 - \frac{\Gamma(c\beta, \beta\lambda)}{\Gamma(c\beta)} + \lambda e^{\lambda x}\left(\frac{\beta}{\beta + x}\right)^{c\beta} (\beta + x)\frac{\Gamma(-1 + c\beta, (\beta + x)\lambda)}{\Gamma(c\beta)} \\ &= 1 - \frac{\Gamma(c\beta, \beta\lambda)}{\Gamma(c\beta)} + \lambda e^{\lambda x}\left(\frac{\beta}{\beta + x}\right)^{c\beta} (\beta + x) \\ &\qquad \left(\frac{\Gamma(c\beta, (\beta + x)\lambda)}{(c\beta - 1)\Gamma(c\beta)} - \frac{((\beta + x)\lambda)^{c\beta - 1}e^{-(\beta + x)\lambda}}{(c\beta - 1)\Gamma(c\beta)}\right) \\ &\sim \frac{\lambda}{c}e^{-(c-\lambda)x}\left(1 + \frac{x(2 + cx)}{2\beta} + \frac{x^3(4 + 3cx)}{24\beta^2} + o\left(\frac{1}{\beta^2}\right)\right), \quad \beta \to \infty. \end{aligned}$$

Case II.2: Mixing over the arrival rate. In Figure 5 we compare in an analogous way the waiting time tail (22) for Gamma(α, β)-distributed arrival rate Λ, where for each choice of β, α is chosen such that $\mathbb{E}[\Lambda] = 3$, where $\mu = 4$ is fixed. One observes that now the respective curves of the independent model and dependent model do intersect. Figure 6 depicts this intersection point x as a function of β.

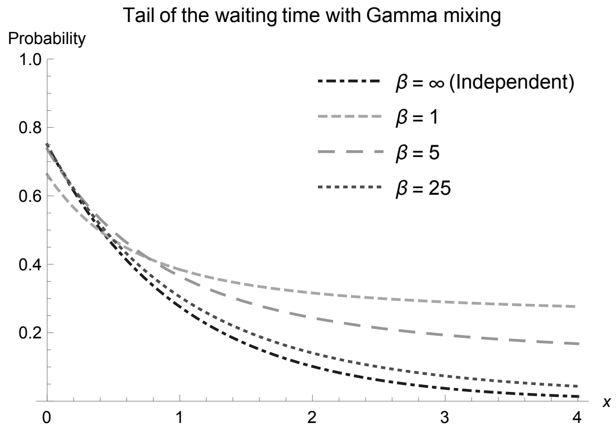

Figure 5: Tail of the waiting time for the independent model and the dependent model (22) with equal expected arrival rate.

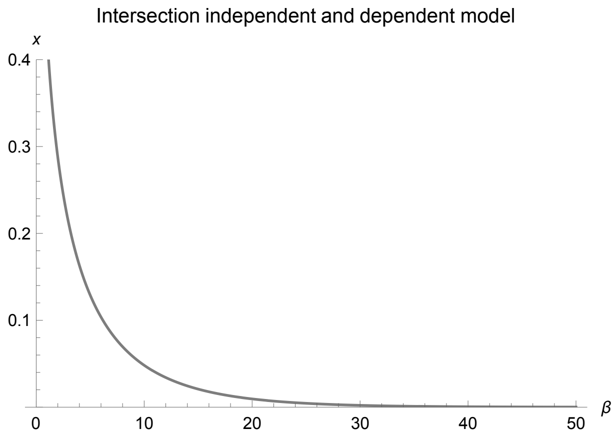

Figure 6: Intersection point of independent and dependent model (22).

For completeness we also give the asymptotic behavior of $\mathbb{P}(W_\lambda > x)$ for $\beta \to \infty$ for this case:

$$\begin{aligned}\mathbb{P}(W_\lambda > x) &= \frac{\Gamma(c\beta, \beta\mu)}{\Gamma(c\beta)} + \frac{1}{\mu}e^{-x\mu}\left(\frac{\beta}{\beta - x}\right)^{c\beta}\frac{1}{\beta - x}\left(c\beta - \frac{\Gamma(1 + c\beta, (\beta - x)\mu)}{\Gamma(c\beta)}\right)\\ &\sim \frac{c}{\mu}e^{-(\mu - c)x}\left(1 + \frac{x(2 + cx)}{2\beta} + \frac{x^2(3(cx)^2 + 20cx + 24)}{24\beta^2} + o\left(\frac{1}{\beta^2}\right)\right), \quad \beta \to \infty,\end{aligned}$$

so the rate of convergence of the tail of this model to the one of the independent M/M/1 model is of order $1/\beta$.

Finally, we consider the case of mixing over ρ according to Method 1, cf. (24). Figure 7 shows that in this case the waiting time tail is less sensitive to dependence introduced via Gamma mixing. The respective asymptotic behavior is given by

$$\mathbb{P}(W_\lambda > x) \sim \rho e^{-c(1-\rho)x}\left(1+\frac{c(1-\rho)^2x^2}{2\beta}\right.$$
$$\left. + \frac{c(1-\rho)^3x^3(3cx(1-\rho)-8)}{24\beta^2} + o\left(\frac{1}{\beta^2}\right)\right), \quad \beta\to\infty,$$

again leading to a convergence rate of order $1/\beta$.

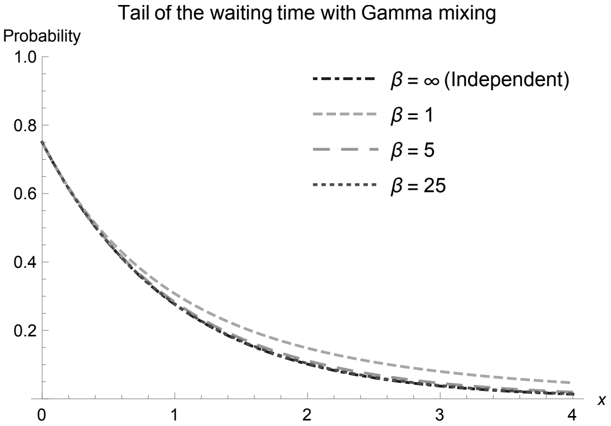

Figure 7: Tail of the waiting time for the independent model and the dependent model with mixing according to Method 1.

3.5 Independent parallel mixing

Until now we studied mixing over either the service rate or the arrival intensity, but one can also mix over *both* the service rate and the arrival intensity. In [7, Example 4.1] this idea was already introduced for the analogous insurance risk model, without providing further details. If we (independently) mix over both the service rate and arrival intensity, the tail of the waiting time is given by

$$\mathbb{P}(W_{\lambda\mu} > x) = \int_0^\infty\int_0^\infty \mathbb{P}(W > x)\mathrm{d}F_X(\mu)\mathrm{d}F_\Lambda(\lambda). \tag{25}$$

In the next example we assume the density of X to satisfy (9) and Λ to be Gamma(δ,η) distributed. One can check that in this example the stability condition is always satisfied. From (25) it now follows that

$$\begin{aligned}\mathbb{P}(W_{\lambda\mu} > x) &= \int_0^\infty\int_\lambda^\infty \frac{\lambda}{\mu}e^{-(\mu-\lambda)x}\alpha e^{-\alpha(\mu-\lambda)}\frac{\eta^\delta}{\Gamma(\delta)}\lambda^{\delta-1}e^{-\eta\lambda}\mathrm{d}\mu\mathrm{d}\lambda\\ &= \int_0^\infty \alpha\lambda\int_0^\infty \frac{e^{-(\alpha+x)z}}{z+\lambda}\mathrm{d}z\frac{\eta^\delta}{\Gamma(\delta)}\lambda^{\delta-1}e^{-\eta\lambda}\mathrm{d}\lambda\\ &= \frac{\alpha\delta}{(1+\delta)}\frac{\eta^\delta}{(x+\alpha)^{(\delta+1)}}\cdot {}_2F_1\left(1+\delta,1+\delta,2+\delta,-\frac{x+\alpha-\eta}{x+\alpha}\right),\end{aligned} \tag{26}$$

where

$$ {}_2F_1\left(a,b,c,z\right)=\sum_{n=0}^{\infty}\frac{(a)_n\cdot(b)_n}{(c)_n}\frac{z^n}{n!} $$

is a hypergeometric function with the Pochhammer symbol

$$ (u)_n=\begin{cases}1, & \text{if } n=0,\\ u\cdot(u+1)\cdot\ldots\cdot(u+n-1), & \text{if } n>0.\end{cases} $$

Remark: If we take δ and η such that $\mathbb{E}[\Lambda]\to c$ and $\mathrm{Var}(\Lambda)\to 0$, we get back to (10).

In Figure 8 we compare the waiting time tails for the independent model, the dependent model given by (10) with parameters $\alpha=1$ and $\lambda=3$, and the dependent model given by (26) with parameters $\alpha=1$, $\delta=3$ and $\eta=1$. One observes that, for this case, introducing additional dependence in the inter-arrival times when the service times are already dependent does not have a significant influence on the tail of the waiting time.

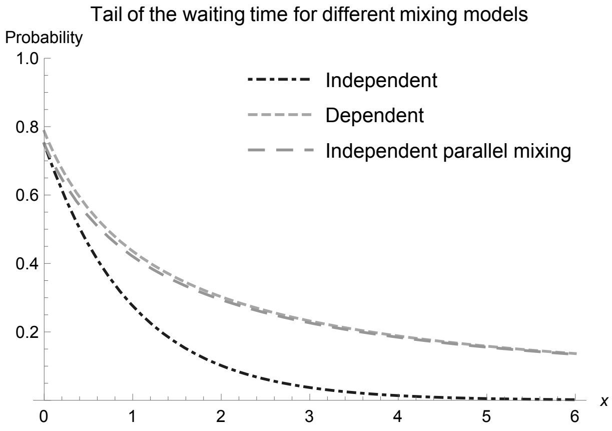

Figure 8: Tail of the waiting time for different mixing models.

Now let us consider the model where X is Gamma(δ,η) distributed and Λ has the density given by (20). Then

$$ \begin{aligned} \mathbb{P}(W_{\lambda\mu}>x) &= \int_0^\infty\int_0^\mu \frac{\lambda}{\mu}e^{-(\mu-\lambda)x}\frac{\alpha e^{-\alpha\lambda}}{1-e^{-\alpha\mu}}\frac{\eta^\delta}{\Gamma(\delta)}\mu^{\delta-1}e^{-\eta\mu}\mathrm{d}\lambda\mathrm{d}\mu \\ &= \int_0^\infty \frac{\alpha e^{-\mu x}}{\mu(1-e^{-\alpha\mu})}\left[\frac{1-e^{(x-\alpha)\mu}\left((\alpha-x)\mu+1\right)}{(x-\alpha)^2}\right]\frac{\eta^\delta}{\Gamma(\delta)}\mu^{\delta-1}e^{-\eta\mu}\mathrm{d}\mu \\ &= \frac{\alpha\left(\frac{\eta}{\delta}\right)^\delta}{(x-\alpha)^2(-1+\delta)}\left(\alpha\cdot\mathrm{Zeta}\left(-1+\delta,\frac{x+\eta}{\alpha}\right)-\alpha\cdot\mathrm{Zeta}\left(-1+\delta,\frac{\alpha+\eta}{\alpha}\right)\right. \\ &\qquad \left.+(x-\alpha)(-1+\delta)\cdot\mathrm{Zeta}\left(\delta,\frac{\alpha+\eta}{\alpha}\right)\right), \end{aligned} \tag{27} $$

where $\operatorname{Zeta}(s,a)=\sum_{k=0}^{\infty}\frac{1}{(k+a)^s}$.

In Figure 9 we compare the waiting time tails of the independent model, the dependent model given by (21) and the dependent model given by (27). It can be seen that, likewise, introducing additional dependence between the service times in the model that already features dependence between interarrival times does not change the tail of the waiting time significantly.

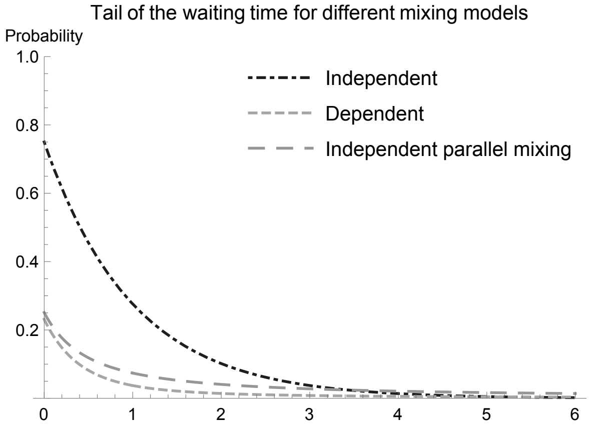

Figure 9: Tail of the waiting time for different mixing models.

In the last example we consider the model where X has density (9) and Λ is Lévy(δ) distributed, cf. (23). From (25) it follows that,

$$\begin{aligned}\mathbb{P}(W_{\lambda\mu}>x)&=\int_0^\infty\int_\lambda^\infty\frac{\lambda}{\mu}e^{-(\mu-\lambda)x}\alpha e^{-\alpha(\mu-\lambda)}\frac{\delta}{2\sqrt{\pi\lambda^3}}e^{-\frac{\delta^2}{4\lambda}}\mathrm{d}\mu\mathrm{d}\lambda\\&=\int_0^\infty\alpha\lambda\int_0^\infty\frac{e^{-(\alpha+x)z}}{z+\lambda}\mathrm{d}z\frac{\delta}{2\sqrt{\pi\lambda^3}}e^{-\frac{\delta^2}{4\lambda}}\mathrm{d}\lambda\\&=\frac{\alpha\delta}{2\sqrt{x+\alpha}}\big(\pi\cos(\sqrt{x+\alpha}\delta)+2\cdot\operatorname{CosIntegral}(\sqrt{x+\alpha}\delta)\sin(\sqrt{x+\alpha}\delta)\\&\qquad-2\cdot\operatorname{SinIntegral}(\sqrt{x+\alpha}\delta)\cos(\sqrt{x+\alpha}\delta)\big)\,,\end{aligned}\tag{28}$$

where $\operatorname{CosIntegral}(x)=-\int_x^\infty\frac{\cos(t)}{t}\mathrm{d}t$ and $\operatorname{SinIntegral}(x)=\int_0^x\frac{\sin(t)}{t}\mathrm{d}t$.

In Figure 10 the waiting time tails of the independent model and the dependent model given by (28) are compared for the values $\alpha=1$ and $\delta=1$.

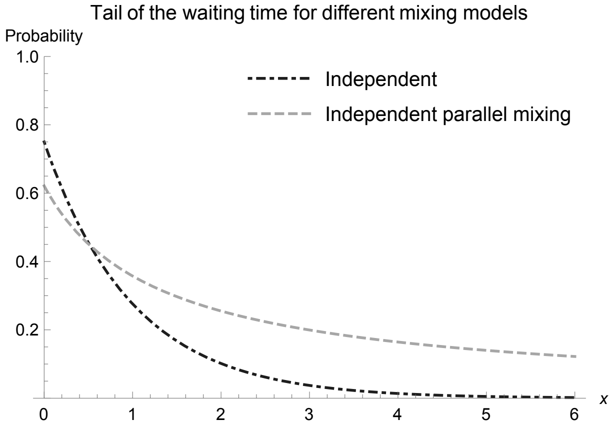

Figure 10: Tail of the waiting time for the independent model and the model with independent parallel mixing.

4 A G/M/1 queue with dependence

We now proceed to the G/M/1 model. Let $A(\cdot)$ be the inter-arrival time distribution with LST $\tilde{A}(\cdot)$. Furthermore, let μ denote the service rate. For this queueing model it is known (cf. Chapter II.3 of [16]) that,

$$\mathbb{P}(W > x) = \sigma e^{-\mu(1-\sigma)x}, \qquad x \geq 0,$$

where σ is the unique zero in $[0,1]$ of

$$\sigma = \tilde{A}(\mu(1-\sigma)). \tag{29}$$

We first study mixing over the service rate μ and subsequently mixing over the traffic load ρ.

4.1 Mixing with respect to the service rate μ

If the density function of X is given by (9), it follows that the general expression for the tail of the waiting time is,

$$\mathbb{P}(W_\mu > x) = \int_\lambda^\infty \alpha e^{-\alpha(y-\lambda)}\sigma(y)e^{-y(1-\sigma(y))x}\mathrm{d}y, \tag{30}$$

where we wrote $\sigma(y)$ instead of σ to emphasize that the zero in $[0,1]$ of (29) depends on the actual service rate.

To get a better understanding of this model, consider the special case of an E_2/M/1 queueing model, i.e. the interarrival times are Erlang$(2, 2\lambda)$ distributed. From (29) one immediately obtains

$$\sigma = 2\rho + \frac{1}{2} - \sqrt{2\rho + \frac{1}{4}}.$$

Substituting this expression for σ into (30) gives

$$\mathbb{P}(W_\mu > x) = \int_\lambda^\infty \alpha e^{-\alpha(y-\lambda)} \left(\frac{2\lambda}{y} + \frac{1}{2} - \sqrt{\frac{2\lambda}{y} + \frac{1}{4}} \right) e^{-y\left(\frac{1}{2} - \frac{2\lambda}{y} + \sqrt{\frac{2\lambda}{y} + \frac{1}{4}}\right)x} \mathrm{d}y. \quad (31)$$

Figure 11 graphically compares this expression for $\lambda = 3$ and $\alpha = 1$ (i.e. $\mathbb{E}[X] = 4$) with the independent model ($\lambda = 3$, $\mu = 4$).

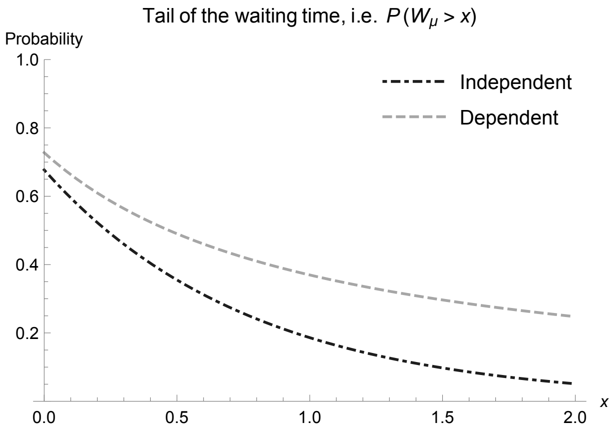

Figure 11: $\mathbb{P}(W_\mu > x)$ for the independent and dependent model in an E_2/M/1 queue.

4.2 Mixing with respect to the traffic load ρ

We now turn to the mixing over the traffic load for the G/M/1 model. For Method 1 from Section 3.3 and X ~Exp(β), one gets

$$\mathbb{P}(W_\mu > x) = \int_0^\infty \beta e^{-\beta y} \sigma e^{-y(1-\sigma)x} \mathrm{d}y. \quad (32)$$

For the E_2/M/1 queue with mean interarrival time $1/\lambda$, σ depends on λ and μ only via their quotient ρ. Since ρ is kept constant in Method 1, formula (32) simpliefies further to

$$\mathbb{P}(W_\mu > x) = \frac{\sigma\beta}{\beta + (1-\sigma)x}, \quad (33)$$

where σ satisfies (29). Figure 12 depicts this function together with the independent model for $\lambda = 3$, $\beta = \frac{1}{4}$ (such that $\mathbb{E}[X] = 4$) and $\mu = 4$, respectively.

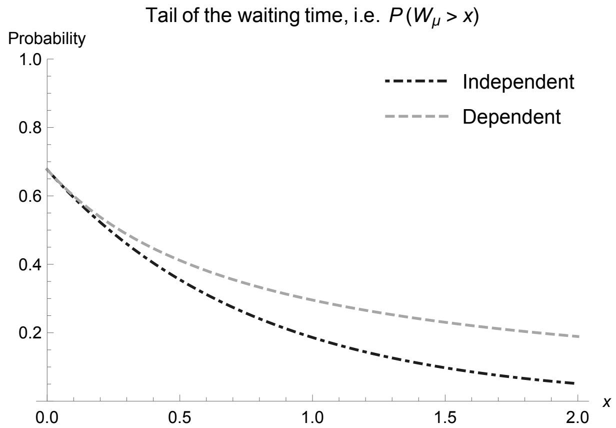

Figure 12: Tail of the waiting time for the independent and dependent model in an $E_2/M/1$ queue when mixing according to Method 1.

Remark: Observe that for any $E_k/M/1$ queueing model ($k \in \mathbb{N}$) σ only depends on ρ, from which we conclude that (33) holds for all these models. Further, for a D/M/1 queue with $D = 1/\lambda$, it follows that (29) is equivalent to $\sigma = e^{-\frac{1-\sigma}{\rho}}$, and (31) again holds for this value of σ.

For Method 2 from Section 3.3, one gets

$$\mathbb{P}(W_\lambda > x) = \int_0^1 \sigma e^{-\mu(1-\sigma)x} \mathrm{d}\rho, \tag{34}$$

where σ can depend on ρ.

For the $E_2/M/1$ queueing model, a substitution of the expression for σ into (34) gives

$$\begin{aligned}\mathbb{P}(W_\lambda > x) &= \int_0^1 \left(2\rho + \frac{1}{2} - \sqrt{2\rho + \frac{1}{4}}\right) e^{-\mu(1-\left(2\rho+\frac{1}{2}-\sqrt{2\rho+\frac{1}{4}}\right))x} \mathrm{d}\rho \\ &= \frac{2e^{-x\mu} - 2 + 3x\mu - \sqrt{x\mu} \cdot e^{x\mu} \int_0^{\sqrt{x\mu}} e^{-w^2} \mathrm{d}w}{4x^2\mu^2}.\end{aligned}$$

Recall that in Method 2 we considered $R \sim \mathrm{Unif}[0, 1]$ and consequently $\mathbb{E}[R] = 1/2$, so for a comparison with the independent model we choose $\lambda = 2$ and $\mu = 4$, cf. Figure 13.

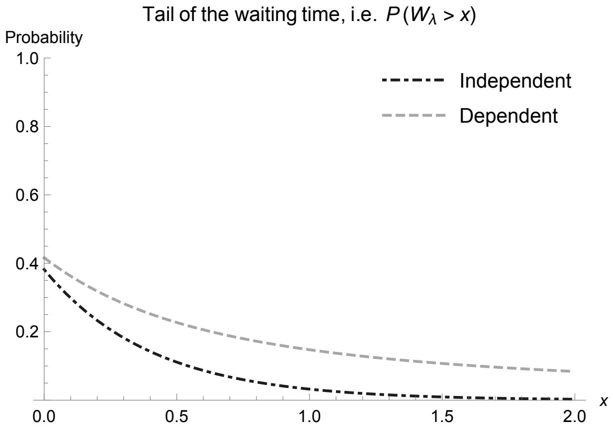

Figure 13: Tail of the waiting time for the independent and dependent model in an $E_2/M/1$ queue when mixing according to Method 2.

5 Conclusion

We have derived explicit expressions for the tail of the waiting time for a single server queue where dependence among service times and/or inter-arrival times is introduced via a mixing procedure. The resulting formulas allow to study the effects of dependence on the waiting distribution in a quantitative way. Such a mixing procedure is a versatile tool and in principle feasible for any kind of model for which explicit formulas for the waiting distribution exist. For instance, a similar analysis is possible for M/Ph/1 queues.

In future work, we would like to consider other performance measures, and more advanced queueing models. As observed by Michel Mandjes (private correspondence), it would also be interesting to assume that the mixing random variable takes a particular value only for a certain period of time; subsequently, another value is drawn for that random variable, etc. Another possible model variant, which was already proposed in [7, Example 4.2] for insurance risk, is a queueing model with comonotonic mixing (i.e., instead of mixing over both service times and inter-arrival times independently, the mixing random variable of λ is a deterministic function of the one for μ). In general, the advantage of the approach proposed in this paper is to obtain explicit expressions for the waiting time tail for queueing models with a certain type of dependence, which can be useful as benchmark expressions for other dependence structures that may be motivated by causal or statistical considerations in concrete applications.

References

[1] Abate, J. and Whitt, W. (1999). Modeling service-time distributions with non-exponential tails: Beta mixtures of exponentials. *Stochastic Models 15*, 517-546.

[2] Abate, J., Choudhury, G.L. and Whitt, W. (1994). Waiting-time tail probabilities in queues with long-tail service-time distributions. *Queueing Systems 16*, 311-338.

[3] Abramowitz, M. and Stegun, I. (1965). *Handbook of Mathematical Functions.* National Bureau of Standards Applied Mathematics Series 55.

[4] Abrarov, S. and Quine, B. (2015). A rational approximation for efficient computation of the Voigt function in quantitative spectroscopy. *Journal of Mathematics Research 7*, 163-174.

[5] Albrecher, H. and Asmussen, S. (2006). Ruin probabilities and aggregate claims distributions for shot noise Cox processes. *Scandinavian Actuarial Journal*, 86-110.

[6] Albrecher, H., Beirlant, J. and Teugels, J. (2017). *Reinsurance: Actuarial and Statistical Aspects.* Wiley, Chichester.

[7] Albrecher, H. Constantinescu, C. and Loisel, S. (2011). Explicit ruin formulas for models with dependence among risks. *Insurance: Mathematics and Economics 48*, 265-270.

[8] Asmussen, S. (2003). *Applied Probability and Queues.* Springer-Verlag, New York.

[9] Asmussen, S. and Albrecher, H. (2010) *Ruin Probabilities.* World Scientific Publ. Cy., Singapore.

[10] Badila, E.S., Boxma, O.J. and Resing, J.A.C. (2015). Two parallel insurance lines with simultaneous arrivals and risks correlated with inter-arrival times. *Insurance: Mathematics and Economics 61*, 48-61.

[11] Boxma, O.J. and Cohen, J.W. (1998). The $M/G/1$ queue with heavy-tailed service time distribution. *IEEE J. Sel. Areas in Communications 16*, 749-763.

[12] Boxma, O.J. and Cohen, J.W. (1999). Heavy-traffic analysis for the $GI/G/1$ queue with heavy-tailed distributions. *Queueing Systems 33*, 177-204.

[13] Bühlmann, H. (1972). Ruinwahrscheinlichkeit bei erfahrungstarifiertem Portefeuille. *Bulletin de l'Association des Actuaires Suisses 2*, 131-140.

[14] Chaudry, M., Temme, N. and Veling, E. (1996). Asymptotic and closed form of a generalized incomplete Gamma function. *Journal of Computational and Applied Mathematics 67*, 371-379.

[15] Chitnis, N. Hyman, J. and Cushing, J. (2008). Determining important parameters in the spread of malaria through the sensitivity analysis of a mathematical model. *Bulletin of Mathematical Biology 70*, 1272-1296.

[16] Cohen, J.W. (1982). *The Single Server Queue.* North-Holland Publ. Cy., Amsterdam.

[17] Cohen, J.W. (1997). On the $M/G/1$ queue with heavy-tailed service time distributions. CWI Report PNA-R9702.

[18] Constantinescu, C., Hashorva, E. and Ji, L. (2011). Archimedean copulas in finite and infinite dimensions – with application to ruin problems. *Insurance: Mathematics and Economics 49*, 487-495.

[19] Dunster, T., Paris, R. and Cang, S. (1998). On the high-order coefficients in the uniform asymptotic expansion for the incomplete Gamma function. *Methods and Applications of Analysis 5*, 223-247.

[20] Dutang, C., Lefèvre, C. and Loisel, S. (2013). On an asymptotic rule $A+B/u$ for ultimate ruin probabilities under dependence by mixing. *Insurance: Mathematics and Economics 53*, 774-785.

[21] Eick, S., Massey, W. and Whitt, W. (1993). The physics of the $M_t/G/\infty$ queue. *Management Science 39*, 241-252.

[22] Gautschi, W. (1998). The incomplete Gamma functions since Tricomi. In Tricomi's Ideas and Contemporary Applied Mathematics, Atti dei Convegni Lincei, n. 147, Accademia Nazionale dei Lincei, pages 203-237.

[23] Jongbloed, G. and Koole, G. (2001). Managing uncertainty in call centers using Poisson mixtures. *Applied Stochastic Models in Business and Industry 17*, 307-318.

[24] Kim, S.-H. and Whitt, W. (2014). Are call center and hospital arrivals well modeled by nonhomogeneous Poisson processes? *Manufacturing and Service Operations Management 16*, 464-480.

[25] Koops, D., Boxma, O.J. and Mandjes, M.R.H. (2016). Networks of $\cdot/G/\infty$ queues with shot-noise driven arrival intensities. *Eurandom Report 2016-010.*

[26] Nelsen, R. (1999). *An Introduction to copulas.* Springer, New York.

[27] Pedro, S., Tonnang, H. and Abelman, S. (2016). Uncertainty and sensitivity analysis of a rift valley fever model. *Applied Mathematics and Computation 279*, 170-186.

[28] Ross, S.M. (1996). *Stochastic Processes.* Wiley, New York.

[29] Zan, J., Hasenbein, J. and Morton, D.P. (2014). Asymptotically optimal staffing of service systems with joint QoS constraints. *Queueing Systems 78*, 359-386.