On Ultra-Sharp Queueing Bounds

Florin Ciucu
University of Warwick

Sima Mehri
University of Warwick

Amr Rizk
University of Duisburg-Essen

***Abstract*—We present a robust method to analyze a broad range of classical queueing models, e.g., the $GI/G/1$ queue with renewal arrivals, an $AR/G/1$ queue with alternating renewals (AR), as a special class of Semi-Markovian processes, and Markovian fluids queues. At the core of the method lies a standard change-of-measure argument to reverse the sign of the *negative* drift in the underlying random walks. Combined with a suitable representation of the overshoot, we obtain exact results in terms of series. Closed-form and computationally fast bounds follow by taking the series' first terms, which are the dominant ones because of the *positive* drift under the new probability measure. The obtained bounds generalize the state-of-the-art class of martingale bounds and can be much sharper by orders of magnitude.**

I. Introduction

The steady-state waiting time W in the $GI/G/1$ queue is the maximal value of a random walk, i.e.,

$$W = \max_{n\geq 0}\{X_1 + \cdots + X_n\}$$

whose expected increment $X_n = S_n - T_n$ satisfies $\mathbb{E}[X_n] < 0$ for stability; T_n are the (renewal) inter-arrival times and S_n are the corresponding service times. In general, there do not exist exact and closed-form solutions for the distribution of W. Existing (exact) approaches involve inversion techniques and numerical methods, which are themselves very challenging due to dealing with Wiener-Hopf type of integral equations (e.g., Cohen [11]).

The $SM/SM/1$ queue generalizes the $GI/G/1$ queue by relaxing the underlying renewal assumption, i.e., the arrival and service processes follow some Semi-Markovian structures. Studying such models is not only driven by measurement studies (e.g., Crovella and Bestavros [12] or Mi *et al.* [20]) but also by the fundamentally different queueing behavior when compared to the baseline renewal models (e.g., Tin [28] or Patuwo *et al.* [22]). The standard analysis of $SM/SM/1$ models is computational in nature involving, e.g., transform methods (e.g., Çinlar [7] or Adan and Kulakarni [1]) or matrix analytical methods (e.g., Neuts [21] or Akar and Sohraby [3]).

The Markovian fluid queue is another classical model in which the arrival process is driven by a Markovian structure whereas the service is fluid. The Markov-Modulated Fluid model appeared in the seminal paper of Anick, Mitra, and Sondhi [4]: the arrival processes consists of several multiplexed On-Off sources and the queueing model can be exactly analyzed using matrix analysis and ordinary differential equations. The more general case requires however more advanced techniques, e.g., spectral decomposition (Akar and Sohraby [2]) or Wiener-Hopf factorization (Rogers [24]). The numerical complexity of such techniques becomes however prohibitive when a large number of sources are multiplexed (Shroff and Schwartz [26]).

A robust alternative to computational techniques for such queueing models is Effective Bandwidth, which enables an asymptotically exact analysis, e.g.,

$$\mathbb{P}(W > \sigma) \approx \gamma_1 e^{-\theta\sigma} \ ,$$

for some asymptotic constant γ_1 and some asymptotic decay rate θ; see the standard review on the topic by Kelly [17]. The *asymptotically exact* representation $f(\sigma) \approx g(\sigma)$ means that $\lim_{\sigma\to\infty} \frac{f(\sigma)}{g(\sigma)} = 1$.

Another robust analytical alternative is to represent the tail probability $\mathbb{P}(W > \sigma)$ in terms of rigorous stochastic bounds, i.e.,

$$\mathbb{P}(W > \sigma) \leq \gamma_2 e^{-\theta\sigma} \ \forall \sigma \geq 0 \ ,$$

for some asymptotic constant γ_2. Such results can be obtained using martingale-based techniques starting with the work of Kingman [18] and stochastic network calculus (Jiang and Liu [16]), which is the probabilistic extension of the deterministic network calculus conceived by Cruz [13].

There are very few studies investigating the accuracy of effective bandwidth approximations or stochastic bounds obtained using either martingale techniques or stochastic network calculus. Using Markovian sources, Choudhury *et al.* [8] showed that effective bandwidth approximations can either overestimate or underestimate the corresponding exact results, in finite regimes, by orders of magnitude; the underlying reason is that while the asymptotic decay rate θ is exact, the corresponding asymptotic constant γ_1 can be extremely inaccurate, especially when many sources are multiplexed. For similar traffic sources, Ciucu *et al.* [10] showed that stochastic network calculus bounds overestimate exact results by orders of magnitude; moreover, related martingale bounds can be very sharp in heavy-traffic. The accuracy of martingale bounds does decay however at low utilizations or depending on the burstiness nature of the arrival processes (Poloczek and Ciucu [23], [9]).

In this paper we explore a new class of stochastic bounds for the three queueing models outlined above. The main idea rests on an exponential change-of-measure – a technique employed in several fields such as importance sampling or sequential hypothesis testing facing with similar technical problems as in queueing analysis. In our context, the crucial benefit of the exponential change-of-measure is reversing the sign of the drift in the underlying random walk: from negative in the original

(probability) space to positive in the transformed space. In this way, a fundamental technical difficulty in the original space can be addressed in a rather straightforwad manner in the new space. Furthermore, using an elementary representation of the overshoot of a point process, we obtain exact results on $\mathbb{P}(W > \sigma)$ in terms of an infinite sum with positive terms, each involving an integral; closed-form bounds could be obtained by solving for any number of terms/integrals.

In their simplest form, the proposed bounds recover Kingman's martingale-based bounds. They could be further arbitrarily sharpened, and yet retain a closed-form, but at the expense of solving for standard integrals. In other words, the proposed class of bounds is subject to an inherent tradeoff between simplicity/expressiveness and accuracy.

In the following, we first present the main ideas and results for the $GI/G/1$ renewal case, along with a numerical example illustrating the ability of the proposed bounds to gradually improve upon state-of-the-art martingale-based bounds. Then we present several extensions, first to the $AR/G/1$ queue, with an alternating renewal (AR) arrival process, and then to a Markovian fluid queue. We will particularly consider the cases of multiplexing both homogeneous and heterogeneous processes, and also the cases when the individual sources are either *more or less bursty than Poisson*; numerical evaluations against simulations confirm that the proposed bounds can be made arbitrarily sharp.

II. The $GI/G/1$ Queue

First we give some background from Sequential Analysis [29], [27] on 'change-of-measure'. Then, using an elementary observation on the overshoot of a renewal process, a new class of $GI/G/1$ bounds will be presented along with some numerical comparisons against the state-of-the-art.

Consider a stationary stochastic process $\mathbb{X} = (X_n)_n$, its natural filtration $\mathcal{F}_n = \sigma(X_1, \ldots, X_n)$, and $\mathcal{F} := \mathcal{F}_\infty$. For a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ denote by $\mathbb{P}_n$ the restriction of $\mathbb{P}$ on $\mathcal{F}_n$.

Assume for some $\theta > 0$ that the moment generating function (MGF) $\phi(\theta) = \mathbb{E}[e^{\theta X}]$ is finite (X has the same law as X_1). For each $n \in \mathbb{N}$ and $A \in \mathcal{F}_n$ define

$$\mathbb{P}_{n,\theta}(A) := \mathbb{E}\left[\frac{e^{\theta(X_1+\cdots+X_n)}}{\phi(\theta)^n} 1_A\right] = \int_A \frac{e^{\theta(X_1+\cdots+X_n)}}{\phi(\theta)^n} d\mathbb{P}_n \ , \tag{1}$$

where 1_A denotes the indicator function. $\mathbb{P}_{n,\theta}$ is a probability measure on $(\Omega, \mathcal{F}_n)$ and has the absolute continuity properties $\mathbb{P}_{n,\theta} << \mathbb{P}_n$ and $\mathbb{P}_n << \mathbb{P}_{n,\theta}$. The corresponding Radon-Nikodym derivatives are

$$\frac{d\mathbb{P}_{n,\theta}}{d\mathbb{P}_n} = \frac{e^{\theta(X_1+\cdots+X_n)}}{\phi(\theta)^n} \text{ and } \frac{d\mathbb{P}_n}{d\mathbb{P}_{n,\theta}} = \frac{\phi(\theta)^n}{e^{\theta(X_1+\cdots+X_n)}} \ ,$$

respectively. Moreover, based on Kolmogorov's Extension Theorem, there exists a probability measure $\mathbb{P}_\theta$ on $(\Omega, \mathcal{F})$ such that $\mathbb{P}_{n,\theta}$ is the restriction of $\mathbb{P}_\theta$ on $\mathcal{F}_n$; denote by $\mathbb{E}_\theta$ the corresponding expectation.

A key result which we will use throughout is Wald's Fundamental Identity (WFI): for a stopping time T and a non-negative random variable $Y \geq 0$ which is prior to T (i.e., $Y 1_{T=n}$ is measurable $\mathcal{F}_n$) the following holds

$$\mathbb{E}_\theta\left[Y 1_{T<\infty}\right] = \mathbb{E}\left[Y e^{\theta(X_1+\cdots+X_T)} \phi(\theta)^{-T} 1_{T<\infty}\right] \ . \tag{2}$$

Note that, in particular, if $T = n$ and Y is measurable $\mathcal{F}_n$, then

$$\mathbb{E}_\theta\left[Y\right] = \mathbb{E}\left[Y e^{\theta(X_1+\cdots+X_n)} \phi(\theta)^{-n}\right]$$

and

$$\mathbb{E}\left[Y\right] = \mathbb{E}_\theta\left[Y e^{-\theta(X_1+\cdots+X_n)} \phi(\theta)^{n}\right] \ . \tag{3}$$

A. $GI/G/1$: State-of-the-Art Bounds

Consider now the $GI/G/1$ queue. The inter-arrival and service times are denoted by the iid (independent and identically distributed) sequences $(T_n)_n$ and $(S_n)_n$, respectively. We assume that the service times are light-tailed in the sense that there exists $\theta > 0$ such that $\phi(\theta) = 1$; note that such a condition depends on the inter-arrival times as well. Denote $X_n := S_n - T_n$ and assume for stability that $\mathbb{E}[X_1] < 0$. In steady-state, the waiting time W of an arbitrary job is

$$W = \max_{n \geq 0} \{X_1 + \cdots + X_n\} \ .$$

We shall focus on the tail $\mathbb{P}(W > \sigma)$, for some $\sigma \geq 0$, which is subject to the duality

$$\mathbb{P}(W > \sigma) = \mathbb{P}(T < \infty) \ ,$$

where T is the stopping time

$$T := \min\{n : X_1 + \cdots + X_n > \sigma\} \ .$$

By convention, the minimum of the empty set is denoted by ∞ and sums with no terms are set to zero; an important remark is that $\mathbb{P}(T = \infty) > 0$ because $\mathbb{E}[X_1] < 0$.

Kingman [18] proposed a method to bound $\mathbb{P}(W > \sigma)$ by first constructing the martingale[1]

$$M_n := e^{\theta(X_1+\cdots+X_n)} \ ,$$

where $\theta > 0$ satisfies $\phi(\theta) = 1$. It then follows from the Optional Sampling Theorem (see Mehri and Ciucu [19] for more details) that

$$\begin{aligned} 1 = \mathbb{E}\left[M_0\right] = \mathbb{E}\left[M_T 1_{T<\infty}\right] = E\left[e^{\theta(X_1+\cdots+X_T)} 1_{T<\infty}\right] \\ \geq e^{\theta\sigma}\mathbb{P}(T < \infty) = e^{\theta\sigma}\mathbb{P}(W > \sigma) \ , \end{aligned} \tag{4}$$

and consequently the Kingman bound

$$\mathbb{P}(W > \sigma) \leq e^{-\theta\sigma} \ .$$

The inequality in (4) can be refined as

$$E\left[e^{\theta(X_1+\cdots+X_T)} 1_{T<\infty}\right] \geq \inf_{x \geq 0} K(x) e^{\theta\sigma}\mathbb{P}(T < \infty) \ ,$$

[1] Briefly, an integrable stochastic process M_n is a martingale if its conditional expected increment is zero, i.e., $\mathbb{E}\left[M_{n+1} - M_n \mid \mathcal{F}_n\right] = 0 \ \forall n$, where $\mathcal{F}_n = \sigma(M_1, \ldots, M_n)$.

where $K(x) := \mathbb{E}\left[e^{\theta(X_1-x)} \mid X_1 \geq x\right]$, which leads to the Ross [25] bound

$$\mathbb{P}(W > \sigma) \leq \frac{1}{\inf_{x\geq 0} K(x)} e^{-\theta\sigma} \ ,$$

which is sharper than Kingman's bound because $K(x) \geq 1 \ \forall x \geq 0$.

B. A New Class of Bounds

The weakness of the Kingman and Ross bounds lies in the coarse treatment of the dependency inside $E\left[e^{\theta(X_1+\cdots+X_T)}1_{T<\infty}\right]$; in particular, Kingman's bound simply resorts to bounding the first term by the non-random quantity $e^{\theta\sigma}$.

A more effective method is to apply WFI from (2) with $Y = 1$, i.e.,

$$\mathbb{E}\left[1_{T<\infty}\right] = \mathbb{E}_\theta\left[e^{-\theta(X_1+\cdots+X_T)}1_{T<\infty}\right] \ . \tag{5}$$

The crucial observation is that $T < \infty$ a.s. on $(\Omega, \mathcal{F}, \mathbb{P}_\theta)$. Indeed, since $\phi(\theta)$ is convex and $\phi(0) = \phi(\theta) = 1$, it follows that $\phi'(\theta) = \mathbb{E}[Xe^{\theta X}] > 0$ and hence

$$\mathbb{E}_\theta[X] = \mathbb{E}\left[Xe^{\theta X}\right] > 0 \ .$$

Thus $\mathbb{P}_\theta(T < \infty) = 1$ and (5) becomes

$$\mathbb{P}(W > \sigma) = \mathbb{E}_\theta\left[e^{-\theta(X_1+\cdots+X_T)}\right] = e^{-\theta\sigma}\mathbb{E}_\theta\left[e^{-\theta R_\sigma}\right] \ ,$$

where

$$\mathbb{R}_\sigma := X_1 + \cdots + X_T - \sigma \tag{6}$$

is the overshoot. Note that Kingman's bound (aka Lundberg inequality in Financial Mathematics) simply follows from $\mathbb{R}_\sigma \geq 0$. A standard approach to continue is to express the overshoot R_σ as the remaining lifetime relative to a new renewal process constructed in terms of the (positive) ladder heights of the original (and not necessarily positive) process X_n; finding the distribution of the ladder heights is however an open problem in the general case. The classical related result is the so-called Cramér-Lundberg approximation when $\sigma \to \infty$ [5].

Because our main target are non-asymptotic regimes (i.e., finite values of σ), we will rely on the following elementary expansion of the overshoot's tail

$$\{R_\sigma > x\} = \cup_{n\geq 1}\{\sum_{i=1}^n X_i > \sigma + x, \max_{1\leq k\leq n-1}\sum_{i=1}^k X_i \leq \sigma\} \ , \tag{7}$$

for any $x > 0$, in terms of a union of disjoint events; the disjointness follows from the fact that the events $\{\sum_{i=1}^k X_i \leq \sigma\}$ and $\{\sum_{i=1}^k X_i > \sigma + x\}$ are disjoint themselves for all $k \geq 1$ and $x > 0$. Using this expansion we obtain the following exact result on the tail $\mathbb{P}(W > \sigma)$.

Theorem 1. *The waiting time distribution satisfies*

$$\mathbb{P}(W > \sigma) = e^{-\theta\sigma}\left(1 - \sum_{n=1}^\infty g_n(\sigma)\right) \tag{8}$$

for all $\sigma > 0$, where

$$g_n(\sigma) := \mathbb{E}\left[\left(e^{\theta\sum_{i=1}^n X_i} - e^{\theta\sigma}\right)1_{T=n}\right] \ \forall n \geq 1 \ .$$

Note that, according to (3), $g_n(\sigma)$ can be rewritten as

$$g_n(\sigma) := \mathbb{E}_\theta\left[\left(1 - e^{\theta(\sigma-\sum_{i=1}^n X_i)}\right)1_{T=n}\right] \ \forall n \geq 1 \ . \tag{9}$$

Since $g_n(\sigma) \geq 0$, upper bounds on $\mathbb{P}(W > \sigma)$ follow by taking any number of terms $g_n(\sigma)$ in the sum from (12). These can be written recursively for $n \geq 2$ as

$$\begin{aligned} g_n(\sigma) &= \mathbb{E}\left[\mathbb{E}\left[\left(e^{\theta\sum_{i=1}^n X_i} - e^{\theta\sigma}\right)1_{\{T=n\}}1_{X_1\leq\sigma} \mid X_1\right]\right] \\ &= \mathbb{E}\left[e^{\theta X_1}1_{X_1\leq\sigma}\mathbb{E}\left[\left(e^{\theta(X_2+\cdots+X_n)} - e^{\theta(\sigma-X_1)}\right)1_A \mid X_1\right]\right] \\ &= \mathbb{E}\left[e^{\theta X_1}1_{X_1\leq\sigma}g_{n-1}(\sigma - X_1)\right] \end{aligned} \tag{10}$$

and $g_1(\sigma) = \mathbb{E}\left[\left(e^{\theta X_1} - e^{\theta\sigma}\right)1_{X_1>\sigma}\right]$, where A is the event $\{\sum_{i=2}^n X_i > \sigma - X_1, \max_{2\leq k\leq n-1}\sum_{i=2}^k X_i \leq \sigma - X_1\}$.

Proof. Fixing $\sigma \geq 0$ we can write

$$\begin{aligned} \mathbb{E}_\theta\left[e^{-\theta R_\sigma}\right] &= \int_0^1 \mathbb{P}_\theta\left(e^{-\theta R_\sigma} > y\right)dy \\ &= 1 - \int_0^1 \mathbb{P}_\theta\left(R_\sigma > -\frac{\ln y}{\theta}\right)dy \\ &= 1 - \int_0^\infty \mathbb{P}_\theta(R_\sigma > z)\theta e^{-\theta z}dz \\ &= 1 - \mathbb{P}_\theta(R_\sigma > Z) = 1 - \sum_{n=1}^\infty g_n(\sigma) \ , \end{aligned}$$

where Z is an exponential random variable with parameter θ, and independent of $\mathcal{F} := \mathcal{F}_\infty$, whereas the terms $g_n(\sigma)$ follow from the expansion (7) by first denoting for convenience $U := \sum_{i=1}^n X_i$ and $V := \max_{1\leq k\leq n-1}\sum_{i=1}^k X_i$, i.e.,

$$\begin{aligned} g_n(\sigma) &= \mathbb{P}_\theta\left(\sum_{i=1}^n X_i > \sigma + Z, \max_{1\leq k\leq n-1}\sum_{i=1}^k X_i \leq \sigma\right) \\ &= \mathbb{E}\left[e^{\theta U}1_{\{U>\sigma+Z, V\leq\sigma\}}\right] = \mathbb{E}\left[\mathbb{E}\left[e^{\theta U}1_{\{U>\sigma+Z,V\leq\sigma\}} \mid \mathcal{F}_n\right]\right] \\ &= \mathbb{E}\left[1_{V\leq\sigma}\mathbb{E}\left[e^{\theta U}1_{U>\sigma+Z} \mid \mathcal{F}_n\right]\right] \ . \end{aligned}$$

In the second line we rewrote $\mathbb{E}_\theta$ in terms of $\mathbb{E}$ according to (1), and in the last line we used the measurability of $1_{V\leq\sigma}$ with respect to $\mathcal{F}_n$. Denoting $U := \sum_{i=1}^n X_i$ we can expand the inner conditional expectation as

$$\begin{aligned} \mathbb{E}\left[e^{\theta U}1_{U>\sigma+Z} \mid \mathcal{F}_n\right] &= \mathbb{E}\left[e^{\theta U}1_{U>\sigma}1_{U>\sigma+Z} \mid U\right] \\ &= e^{\theta U}1_{U>\sigma}\mathbb{E}\left[1_{U>\sigma+Z} \mid U\right] \\ &= e^{\theta U}1_{U>\sigma}\mathbb{P}\left(Z < U - \sigma \mid U\right) \\ &= e^{\theta U}1_{U>\sigma}\left(1 - e^{-\theta(U-\sigma)}\right) \\ &= 1_{U>\sigma}\left(e^{\theta U} - e^{\theta\sigma}\right) \ . \end{aligned}$$

Therefore

$$\begin{aligned} g_n(\sigma) &= \mathbb{E}\left[\left(e^{\theta\sum_{i=1}^n X_i} - e^{\theta\sigma}\right)1_{\{U>\sigma,V\leq\sigma\}}\right] \\ &= \mathbb{E}\left[\left(e^{\theta\sum_{i=1}^n X_i} - e^{\theta\sigma}\right)1_{T=n}\right] \ . \end{aligned}$$

□

C. Example: $M/D/1$

For numerical illustration we consider the $M/D/1$ queue with $T_n \sim Exp(\lambda)$ and deterministic service time S. For $\sigma < S$, we obtain by elementary integration the following bound when using only the term $g_1(\sigma)$ in the sum from (12)

$$\mathbb{P}(W > \sigma) \leq 1 - \frac{\theta}{\lambda + \theta} e^{-\lambda(S-\sigma)} \ ,$$

where $\rho := \lambda S < 1$ denotes the utilization factor. In turn, by only using the first two terms, $g_1(\sigma)$ and $g_2(\sigma)$ we obtain

$$\mathbb{P}(W > \sigma) \leq 1 - \left(1 + \theta S e^{-\theta S} - e^{-2\theta S}\right) e^{-\lambda(2S-\sigma)} \ .$$

Similar bounds can be obtained for $S \leq \sigma < 2S$; these are not shown here for brevity. Fig. 1 shows the two sets of bounds against the standard Ross bound and also simulations based on 10^7 independent runs; we note that $M/D/1$ has an exact result, yet it is subject to significant numerical complications due to an underlying sum with nearly cancelling very-large terms (see Iversen and Staalhagen [15]).

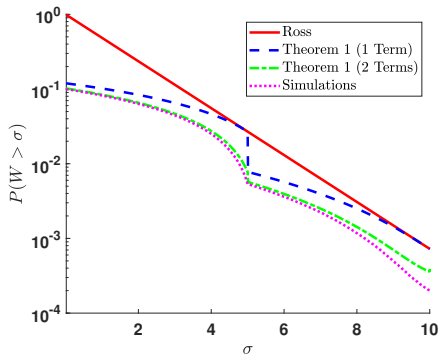


Fig. 1: Waiting-time CCDF for $M/D/1$; $\lambda = 0.02$, $S = 5$, $\rho = 0.1$

Besides the obvious improvement of Ross' bound, the crucial observation is that the gradual improvements appear to decay exponentially. This is supported by the recursive representation of $g_n(\sigma)$ from (10). Moreover, from the alternative representation from (9), the dominant terms in the sum $\sum_n g_n(\sigma)$ are seemingly the first ones, given the positive drift $\mathbb{E}_\theta[X_1] > 0$. A fundamental open question is whether there exists $k \geq 0$ such that $\sum_{n \geq k} g_n(\sigma)$ is an analytic series; if so, then the first k terms from the sum would be largely sufficient to shed most of numerical inaccuracies in Ross' bound.

III. The $AR/G/1$ Queue

Here we extend the $GI/G/1$ results to a queue with non-renewal arrivals. We consider the case of an alternating renewal (AR) process driven by a deterministic Markov chain a_n with two states 1 and 2, i.e.,

$$\mathbb{P}(a_n = 3 - j \mid a_{n-1} = j) = 1 \ \forall n \geq 1, \ j \in \{1, 2\} \ .$$

In state j, the inter-arrivals form an iid sequence $\left(T_n^{(j)}\right)_n$ with the same law as $T^{(j)}$; denote by $T_n := T_n^{(a(n))}$ the inter-arrival time at time n. The service times are denoted by the iid sequence $(S_n)_n$ with the same law as S. Assume that $\mathbb{P}(a_0 = 1) = .5$ and the stability condition $\mathbb{E}[X_n] < 0$ where $X_n := S_n - T_n$.

Denoting the MGFs $\phi_j(\theta) = \mathbb{E}\left[e^{\theta(S - T^{(j)})}\right]$ for $j \in \{1, 2\}$, the light-tailed condition of the service times is slightly more involved than in the $GI/G/1$ case. Indeed, assume that there exist θ_j such that $\phi_j(\theta_j) = 1$ for $j \in \{1, 2\}$. Assuming without loss of generality that $\theta_1 \leq \theta_2$, we additionally assume that $\phi_1(\theta_2) < \infty$. Using the continuity and convexity of $\phi_j(\theta)$, along with $\phi_j(0) = 0$, it then follows that there exists $\theta \in [\theta_1, \theta_2]$ such that $\phi_1(\theta)\phi_2(\theta) = 1$.

In the following, similarly as in the $GI/G/1$ case, we seek a suitable change-of-measure to reverse the sign of the expected increment $\mathbb{E}[X_n]$. Observe first that the process

$$M_n := h_{a_n} e^{\theta(X_1 + \cdots + X_n)}$$

is a martingale, where $\theta > 0$ satisfies

$$\sqrt{\mathbb{E}[e^{-\theta(T^{(1)} + T^{(2)})}]}\mathbb{E}[e^{\theta S}] = 1$$

according to the assumption of light-tailed service times, and $h_1 := 1$ and $h_2 := \sqrt{\frac{\mathbb{E}[e^{-\theta T^{(1)}}]}{\mathbb{E}[e^{-\theta T^{(2)}}]}}$; note that $h_2 \geq 1$ according to the earlier assumption that $\theta_1 \leq \theta_2$. The proof follows immediately from

$$\mathbb{E}\left[h_{a_{n+1}} e^{\theta(S_{n+1} - T_{n+1})} \mid a_n\right] = h_{a_n} \ \forall n \geq 1 \ .$$

The change-of-measure

$$\mathbb{P}_{n,\theta}(A) := \mathbb{E}\left[\frac{M_n}{\mathbb{E}[M_0]} 1_A\right]$$

for $A \in \mathcal{F}_n = \sigma(a_1, X_1, \ldots, a_n, X_n)$ entails the same properties as in the renewal case from § II, such as the existence of $\mathbb{P}_\theta$ on $(\Omega, \mathcal{F})$ and WFI

$$\mathbb{E}_\theta\left[Y 1_{T < \infty}\right] = \mathbb{E}\left[Y \frac{M_T}{\mathbb{E}[M_0]} 1_{T < \infty}\right] \tag{11}$$

for any stopping time T and any non-negative r.v. $Y \geq 0$ prior to T; a key reason is using a normalized martingale with $E\left[\frac{M_n}{\mathbb{E}[M_0]}\right] = 1$ (see Asmussen [5], pp. 358–359).

Theorem 2. *Under the above setting, the stationary waiting-time distribution satisfies for all $\sigma \geq 0$*

$$\mathbb{P}(W > \sigma) = e^{-\theta\sigma}\left(\frac{h_1 + h_2}{2} - \sum_{n=1}^{\infty} g_n(\sigma)\right) \ , \tag{12}$$

where $g_n(\sigma) := \mathbb{E}g_n(\sigma, a_0)$ and

$$g_n(\sigma, j) := \mathbb{E}\left[\left(h_{a_n} e^{\theta(X_1 + \cdots + X_n)} - e^{\theta\sigma}\right) 1_{\{T = n\}} \mid a_0 = j\right]$$

for $j \in \{1, 2\}$ and the standard stopping time

$$T := \inf\{n \geq 1 : X_1 + \cdots + X_n > \sigma\} \ .$$

Proof. Fixing $\sigma \geq 0$, WFI from (11) yields

$$\mathbb{P}(W > \sigma) = \mathbb{E}[1_{T < \infty}] = e^{-\theta\sigma}\mathbb{E}_\theta\left[\frac{e^{-\theta R_\sigma}}{h_{a_T}}\right]\mathbb{E}[h_{a_0}] \ .$$

Since $\frac{e^{-\theta R_\sigma}}{h_{a_T}} \in [0,1]$, from the definition of θ, we can write as in the proof of Theorem 1

$$\mathbb{E}_\theta\left[\frac{e^{-\theta R_\sigma}}{h_{a_T}}\right] = \int_0^1 \mathbb{P}_\theta\left(\frac{e^{-\theta R_\sigma}}{h_{a_T}} > y\right) dy$$
$$= 1 - \sum_{n=1}^{\infty} \frac{g_n(\sigma)}{E[h_{a_0}]} \ ,$$

where Z is an exponential random variable with parameter θ independent of $\mathcal{F} := \mathcal{F}_\infty$. Further, using the expansion from (7), and denoting for convenience $U := \sum_{i=1}^n X_i$ and $V := \max_{1\le k\le n-1} \sum_{i=1}^k X_i$

$$g_n(\sigma) = \mathbb{P}_\theta\left(U > \left(Z - \frac{\ln h_{a_n}}{\theta}\right)_+ + \sigma, V \le \sigma\right)\mathbb{E}[h(a_0)]$$
$$= \mathbb{E}\left[\mathbb{E}\left[h_{a_n} e^{\theta U} 1_{\{U > Z - \frac{\ln h_{a_n}}{\theta} + \sigma, U > \sigma, V \le \sigma\}} \mid \mathcal{F}_n\right]\right]$$
$$= \mathbb{E}\left[h_{a_n} e^{\theta U}\left(1 - \frac{1}{h_{a_n}} e^{\theta(\sigma - U)}\right) 1_{\{U>\sigma, V\le\sigma\}}\right]$$
$$= \mathbb{E}\left[\left(h_{a_n} e^{\theta U} - e^{\theta\sigma}\right) 1_{\{T=n\}}\right] \ .$$

□

Since $g_n(\sigma) \ge 0$, upper bounds on $\mathbb{P}(W > \sigma)$ follow by taking any number of terms $g_n(\sigma)$ in the sum from (12). In particular, dismissing all terms yields the Kingman bound

$$\mathbb{P}(W > \sigma) \le \frac{h_1 + h_2}{2} e^{-\theta\sigma} \ .$$

The terms $g_n(\sigma)$ can be written recursively for $n \ge 2$

$$g_n(\sigma, j) := \mathbb{E}\left[\left(h_{a_n} e^{\theta(X_1 + \cdots + X_n)} - e^{\theta\sigma}\right) 1_{\{T=n\}} \mid a_0 = j\right]$$
$$= \mathbb{E}\Bigg[e^{\theta\left(S_1 - T_1^{3-j}\right)} g_{n-1}\left(\sigma - S_1 + T_1^{(3-j)}, 3 - j\right)$$
$$1_{S_1 - T_1^{(3-j)} \le \sigma}\Bigg]$$

and $g_1(\sigma, j) = \mathbb{E}\left[\left(h_{a_1} e^{\theta\left(S_1 - T_1^{3-j}\right)} - e^{\theta\sigma}\right) 1_{S_1 - T_1^{(3-j)} > \sigma}\right]$ for $j \in \{1, 2\}$.

For numerical evaluations we consider the extension of the $M/D/1$ queue, denoted here by $MM/D/1$, in which the alternating renewals are $T^{(j)} \sim Exp(\lambda_j)$ for $j \in \{1, 2\}$ with $\lambda_1 \ge \lambda_2$, and deterministic service times S. The condition on θ becomes

$$\sqrt{\frac{(\lambda_1 + \theta)(\lambda_2 + \theta)}{\lambda_1 \lambda_2}} = e^{\theta S}$$

which can be solved numerically under the stability condition $\rho := \frac{\lambda_1 + \lambda_2}{2} S < 1$.

For $\sigma < S$, Theorem 2 yields the following bound when using only $g_1(\sigma)$:

$$\mathbb{P}(W > \sigma) \le 1 - \frac{1}{2}\left(1 - e^{-\theta S}\right)\left(e^{-\lambda_1(S-\sigma)} - e^{-\lambda_2(S-\sigma)}\right)$$
$$+ \frac{1}{2} e^{-\theta S}\left(h_2 e^{-\lambda_1(S-\sigma)} - e^{-\lambda_2(S-\sigma)}\right) \ ,$$

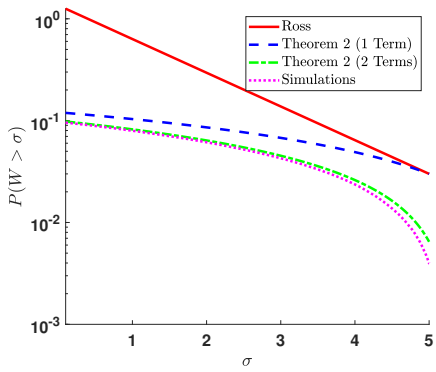


Fig. 2: Waiting-time CCDF for $MM/D/1$; $\lambda_1 = 0.03$, $\lambda_2 = 0.01$, $S = 5$, $\rho = 0.1$

where $h_2 = \sqrt{\frac{\lambda_1(\lambda_2+\theta)}{\lambda_2(\lambda_1+\theta)}}$.

Fig. 2 illustrates the accuracy of the bounds (see the caption for the parameters' values); the expression of the bounds when accounting for $g_2(\sigma)$ as well is not shown for brevity. Similar observations as in Fig. 1 can be drawn, i.e., the Kingman/Ross bounds overestimate by one order of magnitude, whereas the improvements when using $g_1(\sigma)$ and $g_2(\sigma)$ are drastic and almost match the simulations.

IV. Markovian Fluid Queues

Here we analyze a queueing model with Markovian arrivals and fluid service. First we present the main results and then we consider two examples with multiplexing homogeneous and heterogeneous sources.

Let a_n be a stationary and ergodic Markov chain with state space $\mathcal{S} = \{s_i \mid 1 \le i \le m\}$. Assume for simplicity that a_n is reversible and denote the stationary distribution $\pi(i) = \mathbb{P}(a_n = s_i)$ and the transition matrix $P(i,j) = \mathbb{P}(a_{n+1} = s_j \mid a_n = s_i)$.

Consider a queue with instantaneous arrivals a_n and constant service rate C, and assume the stability condition $\mathbb{E}[a_1] < C$. Because of the fluid service, we next focus on the stationary queue size which is subject to the convenient expression

$$Q = \max_{n\ge 0}\{X_1 + \cdots + X_n\}$$

where $X_n := a_n - C$, and which is subject to the duality

$$\mathbb{P}(Q > \sigma) = \mathbb{P}(T < \infty) \ ,$$

where T is the stopping time

$$T = \min\{n \ge 0 : X_1 + \cdots + X_n > \sigma\} \ . \qquad (13)$$

Similarly as in the $GI/G/1$ and $AR/G/1$ cases, we seek a suitable change-of-measure to reverse the sign of the expected increment $\mathbb{E}[a_1 - C]$. For some $\theta \ge 0$, consider the transformed transition matrix $P_\theta(i,j) := P_{i,j} e^{\theta s_j}$ and denote by $\lambda(\theta)$

its spectral radius and by $v = (v_1, \ldots, v_m)$ the corresponding non-negative right eigenvector, according to the Perron-Frobenius Theorem. From the stability condition, the equation $\lambda(\theta) = e^{\theta C}$ has a unique solution and the process

$$M_n = h_{a_n} e^{\theta(X_1 + \cdots + X_n)}$$

is an (arrival) martingale (see, e.g., Duffield [14]), where $h_{s_i} := v_i$.

We use the same change-of-measure as in the $AR/G/1$ case, i.e.,

$$\mathbb{P}_{n,\theta}(A) := \mathbb{E}\left[\frac{M_n}{\mathbb{E}[M_0]} 1_A\right] \tag{14}$$

for $A \in \mathcal{F}_n = \sigma(X_1, \ldots, X_n)$, entailing the same WFI from (11).

The next theorem provides an exact result on the distribution of Q.

Theorem 3. *The stationary queue size has the distribution*

$$\mathbb{P}(Q > \sigma) = e^{-\theta\sigma}\left(\frac{\mathbb{E}[h_{a_0}]}{H} - \sum_{n=1}^{\infty} g_n(\sigma)\right) , \tag{15}$$

for all $\sigma \geq 0$ where

$$g_n(\sigma) := \mathbb{E}\left[\left(\frac{h_{a_n}}{H} e^{\theta \sum_{i=1}^n X_i} - e^{\theta\sigma}\right) 1_{\{T=n\}}\right] ,$$

T is the stopping time from (13) *and*

$$H := \min h_{a_T} = \min\{h_{a_n} : a_n > C\} .$$

Proof. We first need to prove the positivity of the increments $\mathbb{E}_\theta[X_1]$ under the constructed measure $\mathbb{P}_\theta$; unlike in the $GI/G/1$ renewal case, the argument is slightly more compounded. Define first the auxiliary (new) measure

$$\tilde{\mathbb{E}}[Y] := \frac{\mathbb{E}[h_{a_1} Y]}{\mathbb{E}[h_{a_1}]} ,$$

for some non-negative random variable $Y \geq 0$; note that $\tilde{\mathbb{P}}(A) = \tilde{\mathbb{E}}(1_A)$ for some event A. Because M_n is a martingale and a_n is stationary

$$\tilde{\mathbb{E}}[e^{\theta X_1}] = \frac{\mathbb{E}[h_{a_1} e^{\theta X_1}]}{\mathbb{E}[h_{a_1}]} = 1 .$$

Then, using Jensen's inequality for $\tilde{\theta} < \theta$

$$\frac{\mathbb{E}[h_{a_1} e^{\tilde{\theta} X_1}]}{\mathbb{E}[h_{a_1}]} = \tilde{\mathbb{E}}[e^{\tilde{\theta} X_1}] \leq \left(\tilde{\mathbb{E}}[e^{\theta X_1}]\right)^{\frac{\tilde{\theta}}{\theta}} = 1 = \frac{\mathbb{E}[h_{a_1} e^{\theta X_1}]}{\mathbb{E}[h_{a_1}]} .$$

Since $\mathbb{E}[h_{a_1} X_1 e^{\theta X_1}] < \infty$ and

$$h_{a_1} \frac{e^{\theta X_1} - e^{\tilde{\theta} X_1}}{\theta - \tilde{\theta}} \leq h_{a_1} X_1 e^{\theta X_1}$$

from the Mean Value theorem, Fatou's lemma implies that

$$\mathbb{E}_\theta[X_1] = \mathbb{E}[h_{a_1} X_1 e^{\theta X_1}] \geq \mathbb{E}\left[h_{a_1} \lim_{\tilde{\theta} \nearrow \theta} \frac{e^{\theta X_1} - e^{\tilde{\theta} X_1}}{\theta - \tilde{\theta}}\right]$$
$$\geq \limsup_{\tilde{\theta} \nearrow \theta} \mathbb{E}\left[h_{a_1} \frac{e^{\theta X_1} - e^{\tilde{\theta} X_1}}{\theta - \tilde{\theta}}\right] \geq 0 .$$

Therefore, the overshoot R_σ defined as in (6) exists and is non-negative.

For the remaining proof fix $\sigma \geq 0$. Because H is non-random, applying WFI from (11) with $Y = \frac{\mathbb{E}[M_0]}{M_T}$ yields

$$\mathbb{P}(Q > \sigma) = \mathbb{E}[1_{T<\infty}] = e^{-\theta\sigma} \mathbb{E}_\theta\left[\frac{e^{-\theta R_\sigma} H}{h_{a_T}}\right] \frac{\mathbb{E}[h_{a_0}]}{H} .$$

Since $\frac{e^{-\theta R_\sigma} H}{h_{a_T}} \in [0, 1]$ we can write

$$\begin{aligned}
&\mathbb{E}_\theta\left[\frac{e^{-\theta R_\sigma} H}{h_{a_T}}\right] = \int_0^1 \mathbb{P}_\theta\left(\frac{e^{-\theta R_\sigma} H}{h_{a_T}} > y\right) dy \\
&= 1 - \int_0^1 \mathbb{P}_\theta\left(R_\sigma > \frac{\ln(H) - \ln(h_{a_T})}{\theta} - \frac{\ln y}{\theta}\right) dy \\
&= 1 - \int_0^\infty \mathbb{P}_\theta(R_\sigma > \frac{\ln(H) - \ln(h_{a_T})}{\theta} + z)\theta e^{-\theta z} dz \\
&= 1 - \mathbb{P}_\theta\left(R_\sigma > \frac{\ln(H) - \ln(h_{a_T})}{\theta} + Z\right) \\
&= 1 - \sum_{n=1}^{\infty} \frac{g_n(\sigma) H}{E[h_{a_0}]} ,
\end{aligned}$$

where Z is an exponential random variable with parameter θ independent of $\mathcal{F} := \mathcal{F}_\infty$ and, using the expansion from (7),

$$\begin{aligned}
g_n(\sigma) H &= \mathbb{P}_\theta\Bigg(U > \left(Z + \frac{\ln(H) - \ln(h_{a_n})}{\theta}\right)_+ + \sigma, \\
&\qquad V \leq \sigma\Bigg) \mathbb{E}[h_{a_0}] \\
&= \mathbb{E}\left[\mathbb{E}\left[h_{a_n} e^{\theta U} 1_{\{U > Z + \frac{\ln(H) - \ln(h_{a_n})}{\theta} + \sigma, V \leq \sigma, U > \sigma\}} \mid \mathcal{F}_n\right]\right] \\
&= \mathbb{E}\left[h_{a_n} e^{\theta U}\left(1 - \frac{H}{h_{a_n}} e^{\theta(\sigma - U)}\right) 1_{\{V \leq \sigma, U > \sigma\}}\right] \\
&= \mathbb{E}\left[\left(h_{a_n} e^{\theta \sum_{i=1}^n X_i} - H e^{\theta\sigma}\right) 1_{\{T=n\}}\right] ,
\end{aligned}$$

where $U := \sum_{i=1}^n X_i$ and $V := \max_{1 \leq k \leq n-1} \sum_{i=1}^k X_i$. In the first equality we used $1_{U > (z)_+ + \sigma} = 1_{U > z + \sigma} 1_{U > \sigma}$ for any real number z. □

Because $g_n(\sigma) \geq 0$, upper bounds on $\mathbb{P}(W > \sigma)$ follow from Theorem 3 by taking any number of terms $g_n(\sigma)$ in the sum from (15). These could be described recursively for $n \geq 2$ in a slightly more complicated manner than in the $GI/G/1$ case due to necessary conditioning, i.e.,

$$\begin{aligned}
g_n(\sigma) &= \mathbb{E}\left[\mathbb{E}\left[\left(\frac{h_{a_n}}{H} e^{\theta \sum_{i=1}^n X_i} - e^{\theta\sigma}\right) 1_{\{T=n\}}\right] \mid X_1\right] \\
&= \mathbb{E}\Bigg[1_{X_1 \leq \sigma} e^{\theta X_1} \mathbb{E}\Big[\left(\frac{h_{a_n}}{H} e^{\theta U} - e^{\theta(\sigma - X_1)}\right) \\
&\qquad 1_{V \leq \sigma - X_1, U > \sigma - X_1}\Big] \mid X_1\Bigg] \\
&= \mathbb{E}\left[1_{X_1 \leq \sigma} e^{\theta X_1} g_{n-1}(\sigma - X_1, a_1)\right] ,
\end{aligned}$$

where $U := \sum_{i=2}^{n} X_i$, $V := \max_{2\le k\le n-1} \sum_{i=1}^{k} X_i$, and

$$g_n(\sigma, s_i) = \mathbb{E}\left[\left(\frac{h_{a_n}}{H} e^{\theta \sum_{i=1}^{n} X_i} - e^{\theta\sigma}\right) 1_{T=n} \mid a_0 = s_i\right]$$
$$= \sum_j P_{i,j} \mathbb{E}\left[\left(\frac{h_{a_n}}{H} e^{\theta \sum_{i=1}^{n} X_i} - e^{\theta\sigma}\right) 1_{T=n} \mid a_1 = s_j\right]$$
$$= \sum_j P_{i,j} \mathbb{E}\left[1_{X_1\le\sigma} e^{\theta X_1} \left(\frac{h_{a_n}}{H} e^{\theta U} - e^{\theta(\sigma - X_1)}\right) 1_{V\le\sigma - X_1, U>\sigma - X_1} \mid a_1 = s_j\right]$$
$$= \sum_j P_{i,j} \mathbb{E}\left[1_{X_1\le\sigma} e^{\theta X_1} g_{n-1}(\sigma - X_1, s_j) \mid a_1 = s_j\right]$$

for all states $s_i \in \mathcal{S}$. The initial conditions are

$$g_1(\sigma, s_i) = \mathbb{E}\left[\left(\frac{h_{a_1}}{H} e^{\theta X_1} - e^{\theta\sigma}\right) 1_{X_1>\sigma} \mid a_0 = s_i\right] .$$

A. Example 1: Homogeneous On-Off Processes

As a particular case, consider an aggregate of (independent and homogeneous) N On-Off sources/processes $a_n^{(s)}$ for $s = 1, \ldots, N$, each with state space $\{0, 1\}$. A convenient interpretation is that $a_n^{(s)} = 0$ if the source is 'Off' (or idle) at time n, and $a_n^{(s)} = 1$ if the source is 'On' and transmits R data units in a time unit. The probability transition matrix is

$$P = \begin{pmatrix} 1-\alpha & \alpha \\ \beta & 1-\beta \end{pmatrix} .$$

The stationary distribution is $\pi_0 = \frac{\beta}{\alpha+\beta}$ and $\pi_1 = \frac{\alpha}{\alpha+\beta}$ (the index 0 corresponds to the 'Off' state). Denoting by c the per-source capacity the utilization is $\rho = \frac{\mathbb{E}[a_0^{(1)}]}{c} = \frac{\pi_1 R}{c}$ and the overall capacity is $C = Nc$.

Rather than working with the transition matrix for the aggregate process having the state space $\{0, 1, \ldots, N\}$, each corresponding to the number of sources in the 'On' state, we follow a computationally much simpler approach from [23]. Define the spectral radius $\lambda(\theta)$ and the corresponding (non-negative) right eigenvector $h = (h_0, h_1)^T$ for a single source only, relative to the transformed transition matrix $P_\theta(i, j) := P_{i,j} e^{\theta j R}$ for $i, j \in \{0, 1\}$. Denoting by $\theta > 0$ the unique solution of $\lambda(\theta) = e^{\theta c}$, subject to the stability condition $\mathbb{E}[a_0^{(1)}] < c$, the arrival martingale is

$$M_n = h_{a_n} e^{\theta(X_1 + \cdots + X_n)} \tag{16}$$

where $a_n = \sum_{s=1}^{N} a_n^{(s)}$, $h_{a_n} = \prod_{s=1}^{N} h_{a_n^{(s)}}$, and $X_n = a_n R - C$.

Bounds on the tail probability $\mathbb{P}(Q > \sigma)$ follow using the change-of-measure from (14) and applying Theorem 3. Interestingly, the behavior of $\mathbb{P}(Q > \sigma)$ fundamentally differs on the value of $\alpha + \beta$ relative to 1. Observe first that H from Theorem 3 is

$$H = \inf h_{a_T} = \begin{cases} h_1^N & \text{if } h_1 \le h_0 \\ h_0^{N - \lceil C/R \rceil} h_1^{\lceil C/R \rceil} & \text{if } h_1 > h_0 \end{cases}$$

and $\mathbb{E}[h_{a_0^{(1)}}] = \pi_0 h_0 + \pi_1 h_1$, where T is the stopping time defined in (13).

The Kingman bound (see, e.g., Duffield [14]) is

$$\mathbb{P}(Q > \sigma) \le \frac{\mathbb{E}[h_{a_0^{(1)}}]^N}{H} e^{-\theta\sigma} . \tag{17}$$

In turn, a Theorem 3 bound when using only $g_1(\sigma)$ becomes

$$\mathbb{P}(Q > \sigma) \le e^{-\theta\sigma} \left(\frac{\mathbb{E}[h_{a_0^{(1)}}]^N}{H} - \sum_{i > \frac{C+\sigma}{R}} \left(\frac{h_1^i h_0^{N-i}}{H} e^{\theta(iR - C)} - e^{\theta\sigma}\right) \binom{N}{i} \pi_0^{N-i} \pi_1^i\right) .$$

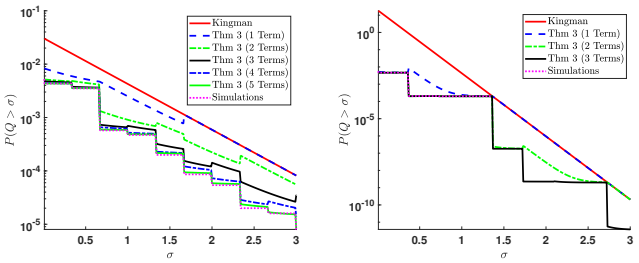
The crucial observation is that, if $h_1 > h_0$ which is equivalent to $\alpha + \beta < 1$ (see Buffet and Duffield [6]), then

$$\mathbb{P}(Q > \sigma) \approx \gamma^N e^{-\theta\sigma} ,$$

for some $\gamma < 1$. In turn, if $h_0 \ge h_1$ which is equivalent to $\alpha + \beta \ge 1$ ([6]) then

$$\mathbb{P}(Q > \sigma) \approx \zeta^N e^{-\theta\sigma} ,$$

but for some $\zeta \ge 1$! In the former regime, called *more bursty than Poisson*, the tail of the queue *decays* exponentially fast in the number of sources N, whereas in the latter regime, called *less bursty than Poisson*, the tail of the queue *grows* exponentially fast in the number of sources N. This behavior was conjectured by Choudhury *et al.* [8] and it is precisely captured by the Kingman bound from (17).



(a) *more bursty*: $\alpha = 0.1,\ \beta = 0.5$ (b) *less bursty*: $\alpha = 0.2,\ \beta = 0.9$

Fig. 3: Kingman and Theorem 3 bounds on the queue size CCDF $\mathbb{P}(Q > \sigma)$ for N On-Off Markov processes, for both *more* and *less bursty* than Poisson regimes ($N = 5$, $R = 1$, $\rho = 0.25$)

Fig. 3 illustrates the Kingman and Theorem 3 bounds, for several number of terms $g_n(\sigma)$, against simulations[2] for both the *more* and *less bursty* than Poisson regimes; the parameters' values are given in the caption. (a) indicates that the Kingman bound overestimates by roughly one order of magnitude, whereas the bounds from Theorem 2 quickly become arbitrarily sharp, e.g., by using $\sum_{n=1}^{5} g_n(\sigma)$ the match

[2]Simulation results are obtained from 10^7 independent runs, each accounting for 10^5 time slots; the running time using a C++ implementation on a standard CPU is about 2 days.

with simulations is almost perfect. We mention that alternative effective bandwidth approximations or stochastic network calculus bounds overestimate themselves the Kingman bounds by several orders of magnitude (see, e.g., [8], [23]).

In turn, in the *less-bursty* regime from (b), the Kingman bound overestimates simulations by roughly three orders of magnitude; counterintuitively, the effective bandwidth approximation was shown to *underestimate* by several orders of magnitude [8]! The bounds from Theorem 3 are very sharp: using only $g_1(\sigma)+g_2(\sigma)$ the match with simulations over the first two '*steps*' is almost perfect (we observe that simulations '*stop*' around $\sigma = 1.4$ due to using only 10^7 runs). The further behavior becomes apparent from using $g_1(\sigma)+g_2(\sigma)+g_3(\sigma)$; capturing the third '*step*' of $\mathbb{P}(Q > \sigma)$ using simulations would require about one year of running time.

B. Example 2: Heterogenous On-Off Processes

Here we generalize the results from § IV-A by considering N_l On-Off sources of type $l = 1, 2$, all being independent. We will use the same notation except for an additional super(sub)script, e.g., the probability transition matrices are $P^{(l)}$, and the transition probabilities are α_l and β_l. Without loss of generality we assume the same rate R for both source types; different rates could be considered by suitably scaling the transition probabilities. The overall capacity is C and assume the stability condition

$$N_1\pi_{1,1}R + N_2\pi_{2,1}R < C \ ,$$

where $\pi_{l,1}$ is the 'On' stationary probability for a type l source.

The key difficulty in the heterogeneous case is the construction of a martingale M_n containing a *single* exponential, as the one from (16) in the homogeneous case. To do so, we split the capacity C into $C = C_1 + C_2$, where C_l is the overall capacity allocated to the sources of type l. We parameterize $C_1 = wC$ where

$$\frac{N_1\pi_{1,1}R}{C} < w < 1 - \frac{N_2\pi_{2,1}R}{C} \ ,$$

such that a stability condition is satisfied for both source types.

As in the homogeneous case, for each source type l, denote by $\lambda_l(\theta)$ the spectral radius of the transformed transition matrix $P_\theta^{(l)}(i,j) = P_{i,j}^{(l)} e^{\theta R j}$ for $i, j \in \{0, 1\}$. For each w in the range above, denote by $\theta_{l,w}$ the unique solutions of

$$\lambda_l(\theta) = e^{\frac{\theta C_l}{N_l}} \tag{18}$$

for $l = 1, 2$; the unicity is guaranteed, as before, from stability. The crucial observation is that $\theta_{1,w} = 0$ when $w \to \frac{N_1\pi_{1,1}R}{C}$ as the utilization for the type 1 sources would approach 1; similarly, $\theta_{2,w} = 0$ at the other extreme $w \to 1 - \frac{N_2\pi_{2,1}R}{C}$ at which the utilization for the type 2 sources would approach 1. Therefore, from the continuity of the solutions θ in (18), depending on the (continuous) parameter w, it follows that there exists a value $w \in \left(\frac{N_1\pi_{1,1}R}{C}, 1 - \frac{N_2\pi_{2,1}R}{C}\right)$ such that

$$\theta := \theta_{1,w} = \theta_{2,w} \ .$$

For this value of θ we denote by $h_l = (h_{l,0}, h_{l,1})^T$ the (non-negative) right eigenvectors corresponding to the spectral radii $\lambda_l(\theta)$ for $l = 1, 2$. We thus obtain the martingales

$$M_{l,n} = h_{l,a_{l,n}} e^{\theta(X_{l,1}+\cdots+X_{l,n})}$$

for $l = 1, 2$, where $a_{l,n} := \sum_{s=1}^{N_l} a_{l,n}^{(s)}$, $h_{l,a_{l,n}} := \prod_{s=1}^{N_l} h_{l,a_{l,n}^{(s)}}$, and $X_{l,n} := a_{l,n}R - C_l$. Furthermore, using the sources' independence, we obtain the single martingale

$$M_n = h_{(a_{1,n},a_{2,n})} e^{\theta(X_1+\cdots+X_n)} \ ,$$

where $h_{(a_{1,n},a_{2,n})} := h_{1,a_{1,n}} h_{2,a_{2,n}}$ and $X_n := X_{1,n} + X_{2,n}$.

The rest follows as in the homogeneous case and Theorem 3. First, without loss of generality we assume that $\frac{h_{1,1}}{h_{1,0}} \le \frac{h_{2,1}}{h_{2,0}}$. Then $H = \inf h_{(a_{1,T},a_{2,T})}$, where T is the stopping time from (13), can be expressed as

$$\begin{cases} h_{1,1}^{N_1} h_{2,1}^{N_2} & \text{if } \frac{h_{1,1}}{h_{1,0}} \le \frac{h_{2,1}}{h_{2,0}} < 1 \\ h_{1,1}^{N_1} h_{2,1}^{L} h_{2,0}^{N_2-L} & \text{if } \frac{h_{1,1}}{h_{1,0}} \le 1 \le \frac{h_{2,1}}{h_{2,0}} \\ h_{1,1}^{N_1 \wedge \lceil C/R \rceil} h_{1,0}^{N_1 - N_1 \wedge \lceil C/R \rceil} h_{2,1}^{L} h_{2,0}^{N_2-L} & \text{if } 1 < \frac{h_{1,1}}{h_{1,0}} \le \frac{h_{2,1}}{h_{2,0}} \end{cases}$$

where $L := (\lceil C/R \rceil - N_1)_+$, $x_+ := \max(x, 0)$, and $x \wedge y := \min(x, y)$ for real numbers x and y.

Observing that

$$\mathbb{E}[h_{a_{1,0},a_{2,0}}] = \left(\frac{\alpha_1 h_{1,1} + \beta_1 h_{1,0}}{\alpha_1 + \beta_1}\right)^{N_1} \left(\frac{\alpha_2 h_{2,1} + \beta_2 h_{2,0}}{\alpha_2 + \beta_2}\right)^{N_2}$$

it follows that the Kingman bound is

$$\mathbb{P}(Q > \sigma) \le \frac{\mathbb{E}[h_{a_{1,0},a_{2,0}}]}{H} e^{-\theta\sigma}$$

whereas the exact result is

$$\mathbb{P}(Q > \sigma) = e^{-\theta\sigma} \left\{ \frac{\mathbb{E}[h_{a_{1,0},a_{2,0}}]}{H} - \sum_{k=1}^{\infty} g_k(\sigma) \right\} \ .$$

In particular,

$$g_1(\sigma) = \sum_{i+j > \frac{C+\sigma}{R}} \left(\frac{h_{1,1}^i h_{1,0}^{N_1-i} h_{2,1}^j h_{2,0}^{N_2-j}}{H} e^{\theta(i+j)R - \theta C} - e^{\theta\sigma} \right) \times q_i^{(1)} q_j^{(2)} \ ,$$

where

$$q_i^{(l)} = \binom{N_l}{i} \left(\frac{\beta_l}{\alpha_l + \beta_l}\right)^{N_l - i} \left(\frac{\alpha_l}{\alpha_l + \beta_l}\right)^{i} \ \forall i = 0, \ldots, N_l$$

and $l = 1, 2$. The general form of $g_n(\sigma)$ is

$$g_n(\sigma) = \sum q_{i_1}^{(1)} Q_{i_1,i_2}^{(1)} \cdots Q_{1_{n-1},i_n}^{(1)} q_{j_1}^{(2)} Q_{j_1,j_2}^{(2)} \cdots Q_{j_{n-1},j_n}^{(2)} \times \left(\frac{h_{1,1}^{i_n} h_{1,0}^{N_1-i_n} h_{2,1}^{j_n} h_{2,0}^{N_2-j_n}}{H} e^{\theta(i_1+j_1+\cdots+i_n+j_n)R - n\theta C} - e^{\theta\sigma} \right)$$

where $Q^{(l)}$ is the transition matrix of the Markov process $a_{l,n}$, i.e., $Q_{k,i}^{(l)} = \mathbb{P}(a_{l,n+1} = i \mid a_{l,n} = k)$, and the first sum is

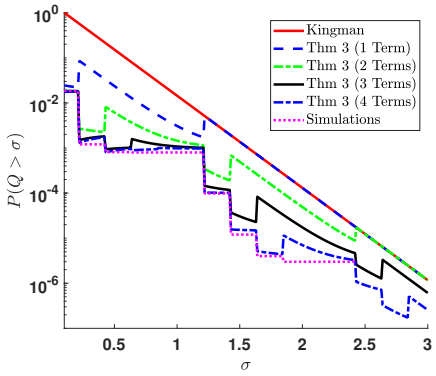


Fig. 4: Bounds on $\mathbb{P}(Q > \sigma)$ for N_1 *less bursty* than Poisson On-Off Markov processes and N_2 *more bursty* than Poisson On-Off Markov processes ($N_1 = 2$, $\alpha_1 = 0.2$, $\beta_1 = 0.9$, $N_2 = 2$, $\alpha_2 = 0.1$, $\beta_2 = 0.5$, $R = 1$, $\rho = 0.25$)

jointly taken after $\max_{1 \leq k \leq n-1} \left\{ \sum_{t=1}^{k} (i_t + j_t) - \frac{kC}{R} \right\} \leq \frac{\sigma}{R}$ and $\sum_{t=1}^{n} (i_t + j_t) - \frac{nC}{R} > \frac{\sigma}{R}$.

Fig. 4 illustrates the accuracy of the bounds by multiplexing the *less bursty* and *more bursty* than Poisson sources used in Figs. 3.(a-b). The inaccuracy of the Kingman bounds is roughly the average of the corresponding inaccuracies from Figs. 3.(a) and (b), due to multiplexing the same types of sources. The proposed bounds become very sharp with only the first four terms $g_n(\sigma)$, whereas simulations stop around $\sigma = 2.4$ due to restricting to 10^7 runs only. Lastly, as an overarching conclusion, simple exponential approximations/bounds are not fit for purpose, given the true behavior of the exact results apparent from simulations; this is progressively captured by the $g_n(\sigma)$ terms.

V. Conclusions

We have proposed a new class of exact results for some of the most popular queueing models, which in their most general forms are only subject to computationally complex numerical methods or proverbially inaccurate approximations/bounds. The new results are expressed as an infinite series with closed-form terms. Remarkably, for numerical purposes, the first few terms are the dominant ones, in the sense of being sufficient to render ultra-sharp queueing bounds; the key reason lies in the reversal of the drift's sign from the original queueing model. Several numerical examples confirm the ultra-sharpness of the new bounds, including the very challenging scenario with multiplexing heterogeneous Markovian sources, some of which being *more bursty than Poisson* and the others being *less bursty than Poisson*.

Acknowledgement

This work has been partially funded by the Engineering and Physical Sciences Research Council (EPSRC) through the project EP/T031115/1 and by the German Research Foundation (DFG), as part of project B4 within the Collaborative Research Center 1053 (MAKI).

References

[1] I. J. B. F. Adan and V. G. Kulkarni. Single-server queue with markov-dependent inter-arrival and service times. *Queueing Syst. Theory Appl.*, 54(1):79, 2006.
[2] N. Akar and K. Sohraby. Infinite- and finite-buffer markov fluid queues: A unified analysis. *Journal of Applied Probability*, 41(2):557–569, 2004.
[3] N. Akar and K. Sohraby. System-theoretical algorithmic solution to waiting times in semi-markov queues. *Perform. Evaluation*, 66(11):587–606, 2009.
[4] D. Anick, D. Mitra, and M. M. Sondhi. Stochastic theory of a data-handling system with multiple sources. *Bell Systems Technical Journal*, 61(8):1871–1894, Oct. 1982.
[5] S. Asmussen. *Applied Probability and Queues*. Springer, 2003.
[6] E. Buffet and N. G. Duffield. Exponential upper bounds via martingales for multiplexers with Markovian arrivals. *Journal of Applied Probability*, 31(4):1049–1060, Dec. 1994.
[7] E. Çinlar. Queues with semi-markovian arrivals. *Journal of Applied Probability*, 4(2):365–379, 1967.
[8] G. L. Choudhury, D. M. Lucantoni, and W. Whitt. Squeezing the most out of ATM. *IEEE Transactions on Communications*, 44(2):203–217, Feb. 1996.
[9] F. Ciucu and F. Poloczek. Two extensions of Kingman's GI/G/1 bound. *Proc. of the ACM on Measurement and Analysis of Computing Systems - ACM Sigmetrics / IFIP Performance*, 2(3):43:1–43:33, Dec. 2018.
[10] F. Ciucu, F. Poloczek, and J. Schmitt. Sharp per-flow delay bounds for bursty arrivals: The case of FIFO, SP, and EDF scheduling. In *Proc. of IEEE Infocom*, pages 1896–1904, Apr. 2014.
[11] J. W. Cohen. *The Single Server Queue (2nd Edition)*. Elsevier Science, 1982.
[12] M. E. Crovella and A. Bestavros. Self-similarity in world wide web traffic: Evidence and possible causes. In *ACM SIGMETRICS Conf. Measurement & Modeling of Comput. Syst.*, pages 160–169, May 1996.
[13] R. Cruz. A calculus for network delay, parts I and II. *IEEE Transactions on Information Theory*, 37(1):114–141, Jan. 1991.
[14] N. G. Duffield. Exponential bounds for queues with Markovian arrivals. *Queueing Systems*, 17(3-4):413–430, Sept. 1994.
[15] V. Iversen and L. Staalhagen. Waiting time distribution in M/D/1 queueing systems. *Electronics Letters*, 35(25):2184–2185, Dec. 1999.
[16] Y. Jiang and Y. Liu. *Stochastic Network Calculus*. Springer, 2008.
[17] F. P. Kelly. Notes on effective bandwidths. In *Stochastic Networks: Theory and Applications. (Editors: F.P. Kelly, S. Zachary and I.B. Ziedins) Royal Statistical Society Lecture Notes Series, 4*, pages 141–168. Oxford University Press, 1996.
[18] J. F. C. Kingman. A martingale inequality in the theory of queues. *Cambridge Philosophical Society*, 60(2):359–361, Apr. 1964.
[19] S. Mehri and F. Ciucu. On a continuous-time martingale and two applications. In *ACM Mobihoc*, pages 111–120, 2023.
[20] N. Mi, Q. Zhang, A. Riska, E. Smirni, and E. Riedel. Performance impacts of autocorrelated flows in multi-tiered systems. *Performance Evaluation*, 64(9):1082–1101, 2007. Performance 2007.
[21] M. F. Neuts. *Matrix-Geometric Solutions in Stochastic Models: An Algorithmic Approach*. Dover, 1981.
[22] B. E. Patuwo, R. L. Disney, and D. C. McNickle. The effect of correlated arrivals on queues. *IIE Transactions*, 25(3):105–110, 1993.
[23] F. Poloczek and F. Ciucu. Scheduling analysis with martingales. *Performance Evaluation (Special Issue: IFIP Performance 2014)*, 79:56 – 72, Sept. 2014.
[24] L. C. G. Rogers. Fluid models in queueing theory and wiener-hopf factorization of markov chains. *The Annals of Applied Probability*, 4(2):390–413, May 1994.
[25] S. M. Ross. Bounds on the delay distribution in GI/G/1 queues. *Journal of Applied Probability*, 11(2):417–421, June 1974.
[26] N. B. Shroff and M. Schwartz. Improved loss calculations at an ATM multiplexer. *IEEE/ACM Transactions on Networking*, 6(4):411–421, Aug. 1998.
[27] D. Siegmund. *Sequential Analysis: Tests and Confidence Intervals*. Springer Series in Statistics. Springer New York, 2013.
[28] P. Tin. A queueing system with markov-dependent arrivals. *Journal of Applied Probability*, 22(3):668–677, 1985.
[29] M. Woodroofe. *Nonlinear Renewal Theory in Sequential Analysis*. CBMS-NSF Regional Conference Series in Applied Mathematics. Society for Industrial and Applied Mathematics, 1982.