Queue and Loss Distributions in Finite-Buffer Queues

Florin Ciucu
University of Warwick
England

Felix Poloczek
Germany

Amr Rizk
TU Darmstadt
Germany

ABSTRACT

We derive simple bounds on the *queue distribution* in finite-buffer queues with Markovian arrivals. Our technique relies on a subtle equivalence between tail events and stopping times orderings. The bounds capture a truncated exponential behavior, involving joint horizontal and vertical shifts of an exponential function; this is fundamentally different than existing results capturing horizontal shifts only. Using the same technique, we obtain similar bounds on the *loss distribution*, which is a key metric to understand the impact of finite-buffer queues on real-time applications. Simulations show that the bounds are accurate in heavy-traffic regimes, and improve existing ones by orders of magnitude. Remarkably, in the limit regime with utilization $\rho = 1$ and iid arrivals, the bounds on the queue size distribution are insensitive to the arrivals distribution.

ACM Reference Format:
Florin Ciucu, Felix Poloczek, and Amr Rizk. 2019. Queue and Loss Distributions in Finite-Buffer Queues. In *Proceedings of ACM conference.* ACM, New York, NY, USA, Article 4, 16 pages.

1 INTRODUCTION

In practice, queueing systems have *finite* buffers to store packets (jobs, customers, etc.) when the service capacity is insufficient; when a buffer fills up then packets are discarded.

The analysis of finite-buffer queues is however challenging. The seminal work of Keilson [30] on the M/G/1/K queue showed that the distribution of the stationary queue size, denoted by Q_K, can be expressed in terms of the corresponding distribution in the infinite-buffer system. For this reason, it has somewhat been natural that the literature dealing with non-Poisson arrivals employed approximations of the form

$$\mathbb{P}(Q_K \geq \sigma) \approx \mathbb{P}(Q_\infty \geq \sigma) \ ,$$

where the subscript denotes the buffer size. The approximations consist of various correction terms, *independent* of σ, which essentially involve horizontal shifts of an exponential function, e.g.,

$$\mathbb{P}(Q_K \geq \sigma) \approx \beta e^{-\theta\sigma} \ ,$$

where β and θ are some parameters (see, e.g., Belhaj and Pap [5]). However, $\mathbb{P}(Q_K \geq \sigma)$ has intrinsically a *truncated* form because Q_K is bounded by K. This behavior involves joint horizontal and vertical shifts, e.g.,

$$\mathbb{P}(Q_K \geq \sigma) \approx \alpha + \beta e^{-\theta\sigma} \ , \tag{1}$$

where α is independent of σ.

In this paper we derive stochastic upper and lower bounds on the *queue distribution* in finite-buffer queueing systems with independent and identically distributed (iid), as well as Markovian arrivals. We consider a discrete-time queueing model (see, e.g., Cruz and Liu [18]) with fluid (infinitely divisible) arrivals and constant-rate service. For a discussion on extending our results to G/G/1/K queues whereby the arrivals are 'jobs', each with its own service time, see § 8.

Our approach is based on a non-trivial extension of Kingman's martingale-based technique to derive bounds in GI/G/1 queues [34]. This involves the construction of two stopping times N_- and N_+, where N_- is finite a.s. (almost surely), such that

$$\{Q_K \geq \sigma\} = \{N_+ < N_-\} \text{ a.s.}$$

Manipulating the two stopping times using martingale properties yields bounds on $\mathbb{P}(Q_K \geq \sigma)$, which explicitly capture the truncated behavior from (1).

Using the same technique, we also derive bounds on the *loss distribution* in finite-buffer queues. We point out that this metric is not only more powerful, but also practically more relevant than the common *loss probability* metric, i.e., the long-run fraction of lost packets (see § 2.2 for additional details). Alike Q_K, the loss probability $\mathbb{P}_L$ is also typically subject to approximations of the form $P_L \approx \beta\mathbb{P}(Q_\infty \geq K)$ for some correction factor β (Mignault *et al.* [39]).

Our results are obtained in both *underload* (i.e., utilization $\rho < 1$) and *overload* ($\rho > 1$) regimes. Using a limit argument we immediately obtain bounds in the *border* regime $\rho = 1$. Remarkably, in the iid case, the bounds on the queue size distribution are insensitive to the arrivals distribution, i.e.,

$$\mathbb{P}(Q_K \geq \sigma) \leq 1 - \frac{\sigma}{C+K} \ ,$$

for all $0 \leq \sigma \leq K$, where C is the service capacity. Similar results are obtained for the loss distribution, yet they are not subject to an insensitivity property.

In comparison to related work, the key benefits of our bounds are negligible numerical complexity and expressivity. Moreover, unlike existing approximations, our results capture the fundamental truncated behavior in finite-buffer queueing systems. Using simulations, it is further shown that the bounds are accurate in heavy-traffic regimes and improve upon existing ones by orders of magnitude.

Next, we first summarize related work in § 2 and then describe the finite-buffer queueing model in § 3. We present the main results in § 4, i.e., bounds on the queue and loss distributions, in both underload and overload regimes. We instantiate these results to iid and Markovian arrivals in § 5 and § 6, respectively. Numerical comparisons with related work are presented in § 7, and possible improvements and extensions in § 8. Brief conclusions are drawn

ACM conference, 2018

ACM ISBN 123-4567-24-567/08/06.

in § 9; an Appendix includes some proofs and additional case studies.

2 RELATED WORK

We first review work concerning the closely related queue distribution and loss probability metrics, and then discuss the loss distribution metric along with other related ones.

2.1 Queue Distribution and Loss Probability

The stationary queue distribution in M/G/1/K queues was obtained by Keilson [30] in terms of the queue distribution in M/G/1 queues (with infinite buffer size). This idea was leveraged in several queueing models with non-Poisson arrivals, and with various degrees of accuracy (e.g., Biskidian *et al.* [6], Gouweleeuw and Tijms [23], Kim and Shroff [32]); for a numerical evaluation of some approximations see [5]. These results immediately lend themselves to the (packet) *loss probability*, i.e., the long-run fraction of lost packets. An exact link between the two distributions was provided by Ishizaki and Takine [27] in the case of state-dependent Markovian arrivals and deterministic service times, under the mild assumption that no arrivals occur in one state.

An exact analysis of the general N/G/1/K ('N' stands for the N-process, also known as Batch Markovian arrival process (BMAP), see, e.g., Lucantoni [38]) queue was carried out by Blondia [7] using matrix analytical techniques, which pose computational complexity issues. Other exact results (for the waiting-time distribution) were obtained by Miyazawa for the GI/GI/1/K queue in terms of transforms [40]. Related "transform-free" results in product-form were obtained by Kim and Chae [33], yet they rely on additional terms posing computational problems except in few cases (e.g., exponential service times). The exact (queue) distribution in the GI/M/1/K queue was recently obtained by Kempa [31] in recursive form involving the Laplace transform of the inter-arrivals; another recursive algorithm for queues with state-dependent Markovian arrivals was given by Gupta and Rao [24]. A diffusion approximation in G/GI/n/K queues was given in Whitt [55]. Computationally efficient algorithms were obtained by Chaudhry *et al.* [15]. Scalable solutions were also investigated by Nagarajan *et al.* [41] in the case of a superposition of Markov-Modulated On-Off (MMOO) sources (see Baiocchi *et al.* [4] as well). It was shown that using a 2-state Markov-Modulated Poisson Process (MMPP) approximation, whereby the matching depends on the buffer size, the loss rates are accurate over a broad range of buffer sizes. A renewal approximation (i.e., matching a GI/D/1/K queue) was shown to perform well only in heavy-traffic, whereas fluid approximations (also explored by Tucker [54], and Yang and Tsang [58]) perform well except in small-buffer regimes. In turn, a Poisson approximation was shown to be inaccurate. A recursive algorithm for estimating the loss probability with arbitrary accuracy was proposed by Sericola [50] for more general Markovian queues. Asymptotic loss rates in queues with heavy-tailed On-Off processes were obtained by Jelenković and Momčilović [28]; remarkably, the approximations are accurate for a broad range of finite values of K.

2.2 Loss Distribution and Distance, and Other Related Metrics

In the study of *finite* queueing systems, Ramaswami [46] argued that the loss probability may be insufficient to understand the loss behavior because although the loss probability can be very small, sources can experience many *consecutive* losses. The explanation is that the *conditional loss probability*, i.e., the probability of losing packets in a slot conditioned on a loss in the previous slot, is high; this was shown using simulations [46] and network measurements by Bolot [10], Yajnik *et al.* [57], and Handley [26]. Consecutive losses (a.k.a., *packet gap*) can be detrimental to network performance, e.g., in scenarios involving Forward Error Correction (FEC) or audio/video transmissions (for a more elaborate discussion see Jiang and Schulzrinne [29]).

Blondia and Casals [9] derived the conditional loss probability in a finite queue with D-BMAP (discrete Batch Markovian Arrival Processes); the result was obtained in analytic form (involving an infinite sum) using matrix analytical techniques. The same authors addressed earlier in [8] the case of a superposition of (discrete) On-Off (Markovian) sources and obtained the *loss distribution* of a tagged source, by enforcing however an additional artificial assumption on the arrivals of the tagged source. An earlier study by Li [36] addressed the case of a superposition of (continuous) On-Off sources; using a stationary analysis the author obtained the average *blocking period* (the maximum interval whereby losses occur continuously) and the average loss rate within it. A key insight was that the length of the blocking period, as well as the behavior of packet loss within such a period, are invariant to the buffer size; the average non-blocking periods would obviously be affected.

Other derivations of conditional loss probabilities were carried out by Schulzrinne *et al.* [49] for Interrupted Poisson Processes (IPP) arrivals and by Takine *et al.* [52] for more general state-dependent Markovian arrivals. In the latter work, dealing with the exponential numerical complexity (in K) in computing the stationary probabilities of an underlying Markov chain was resolved through a recursive algorithm. The authors also derived the distribution of the blocking period (a.k.a. *loss distance*); interestingly, its average has the same expression as the average of the geometric distribution (the Bernoulli probability being the complement of the conditional loss probability), although the underlying Bernoulli trials are not independent. As a side remark, both the loss probability and loss distance were later specified in an IETF RFC document [35] as the key metrics to characterize the performance of real-time applications (e.g., audio and video) from the users' perspective.

The multiclass G/G/N/K queue was addressed by Ferrandiz and Lazar [20], who obtained closed-form results for the average loss distance and average packet gap. A valuable insight for network monitoring, in terms of reducing computational and storage costs, is that the packet gap only depends on the behavior of two consecutively lost packets; in particular, if the latter has a Markovian structure then the packet gap is geometrically distributed.

Other related "loss metrics" include the *loss period* (the difference between the arrival times of the last and first in a series of consecutively lost packets); its behavior was addressed by Fiems *et al.* [21], in an M/G/1/K queue, in terms of a joint transform with the number of losses within such a period. Another is the *block loss*

probability (the fraction of lost packets within a block of consecutive arrivals); a recursive formula was obtained by Cidon *et al.* [16] in an IPP/M/1/K queue, whereas an explicit expression was later derived by Gurewitz *et al.* [25] using ballot theorems in the M/M/1/K case; for a discussion of applications of such results to FEC schemes see [21]. Lastly, we mention the number of lost packets in a busy period (loss as well as blocking periods are sub-intervals of busy periods); asymptotic properties (in K) were obtained by Abramov [1]. Interestingly, in a M/GI/1/K queue at utilization $\rho = 1$, the mean number of lost packets in a busy period is 1 and is independent of K; for an elegant proof using stochastic couplings see Righter [47].

3 MODEL AND METRICS

We consider the discrete-time queueing system from Figure 1, consisting of an arrival flow A served at a constant-rate $C > 0$, and a finite buffer size $K > 0$. The cumulative arrivals until time $n \geq 0$ are given by

$$A(n) := \sum_{k=1}^{n} a_k \ ,$$

where $(a_k)_{k\in\mathbb{N}}$ are the non-negative instantaneous arrivals. The bivariate extension of $A(n)$ is defined for $0 \leq k \leq n$ as $A(k,n) := A(n) - A(k)$; by convention $A(0) = 0$.

The arrival process is stationary and ergodic to guarantee the existence of stationary limits for the underlying queueing processes (see Jelenković and Momčilović [28]).

Let the *utilization* factor

$$\rho := \frac{\mathbb{E}[a_1]}{C} \ . \tag{2}$$

We shall mainly address the *underload* regime (i.e., $\rho < 1$), but also the *overload* (i.e., $\rho > 1$) and *border* (i.e., $\rho = 1$) regimes. Obviously, because K is finite, stability holds in all; for a broader discussion on stability issues see Chapter 2 in Baccelli and Brémaud [3].

3.1 Queue Process

One quantity of interest is the queue process $Q(n)$, which denotes the volume of (fluid) arrivals stored in the buffer at time n.[1] It is defined recursively for $n \geq 0$ as

$$Q(n+1) := \max\{0, \min\{Q(n) + a_{n+1} - C, K\}\} \ , \tag{3}$$

and $Q(0) := 0$. Its non-recursive representation is given in Cruz and Liu [18] (see Eq. (6) therein)

$$Q(n) = \max_{0\leq k\leq n} \left\{ \min\left\{ A(k,n) - (n-k)C, \min_{k\leq m\leq n} \{A(m,n) - (n-m)C + K\} \right\} \right\} .$$

We shall focus on the steady-state limit

$$Q := \max_{n\geq 0} \left\{ \min_{0\leq m<n} \left\{ A(n) - nC, A(m) - mC + K \right\} \right\} . \tag{4}$$

For brevity we wrote, and we shall write (unless otherwise specified), $A(n)$ instead of the corresponding time-reversed process $A^r(n) := \sum_{k=1}^{n} a_{-k}$ (obtained by extending $(a_k)_{k\in\mathbb{N}}$ to a stationary process $(a_k)_{k\in\mathbb{Z}}$ on the whole set of integers).

[1] Note that, unlike in the introduction, we dropped the subscript K from Q_K.

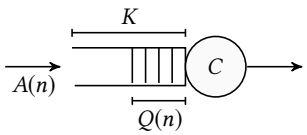


Figure 1: Finite-buffer queueing system: Flow $A(n)$ arriving at a server with capacity $C > 0$; buffer size $K > 0$; actual queue size (buffer content) $Q(n) \leq K$.

Obviously, the previous representation recovers Reich's equation in the case of an infinite-buffer ($K = \infty$), i.e.,

$$Q =_{\mathcal{D}} \max_{n\geq 0} \{A(n) - nC\} \ . \tag{5}$$

3.2 Loss Process

The other quantity of interest is the loss process $L(n)$, which denotes the volume of dropped/lost arrivals at time n as a consequence of having a finite-buffer. It is defined for $n \geq 1$ as

$$L(n) := \max\{Q(n-1) + a_n - C - K, 0\} \ .$$

Its non-recursive representation is also provided in [18] (see Eq. (12) therein) and we shall focus on the steady-state limit

$$L := \max_{n\geq 1} \left\{ \min_{1\leq m<n} \left\{ A(n) - nC - K, A(m) - mC \right\}, 0 \right\} . \tag{6}$$

Obviously, $L = 0$ in the limit $K \to \infty$.

4 MAIN RESULTS

We adopt a general arrival representation in terms of *martingale-envelopes* (see Poloczek and Ciucu [44]):

Definition 1. *The flow A admits a martingale-envelope if there exists a parameter θ and a function $h : \mathrm{Im}(a_1) \to \mathbb{R}^+$ such that the process*

$$M_n := h(a_n) e^{\theta(A(n)-nC)} \tag{7}$$

is a (discrete) martingale for $n \geq 0$.

Recall that $A(n)$ denotes the reversed process, and note that $M_0 = h(a_0)$; Im() denotes the image of a (random) function.

Besides an integrability condition, the crucial property of the martingale is that

$$\mathbb{E}[M_{n+1} \mid \mathcal{F}_n] = M_n \ ,$$

for all $n \geq 1$, where $\mathcal{F}_n$ is the σ-algebra ('*information*') generated by the increments of A until time n. A related concept is that of a stopping-time, which is essentially a random variable N such that the event $\{N = n\}$ is $\mathcal{F}_n$-measurable. More informally, $\{N = n\}$ only depends on the past+present but not the future (i.e., information after time n).

The expression of a martingale-envelope is driven by the expression of the steady-state queue size Q from (5), and in particular the cumulative drift $A(n) - nC$. Martingales are convenient to bound probability events of the form $\mathbb{P}(Q \geq \sigma)$ using Doob's Optional-Stopping Theorem; for a follow-up discussion see § 4.2. The martingale-envelope from (7) essentially transforms the cumulative drift, which is either a supermartingale or submartingale, depending whether $\rho < 1$ or $\rho > 1$, respectively, into a martingale.

The parameter θ is positive when $\rho < 1$ or negative when $\rho > 1$. In general, θ can be regarded as the decay rate of the stationary queue process in an infinite-buffer system, whereby $A(n)$ is served at rate C. In turn, $h()$ is defined on the set of values of a_1, and it does not have a concrete meaning besides its '*role*' to encode the correlation structure of the arrivals; if the increments are iid then $h() = 1$. Explicit constructions for these parameters will be provided in § 5 and § 6.

For our main results we need to define four additional parameters related to the construction of the key stopping times N_+ and N_- from the main results.

Definition 2. *Assume the flow A admits a martingale-envelope with θ and h. Define*

$$H_+ := \inf\{h(x) \mid x > C\} \ , \quad H_- := \inf\{h(x) \mid x < C\} \ ,$$

and also

$$H'_+ := \sup\{h(x) \mid x > C\} \ , \quad H'_- := \sup\{h(x) \mid x < C\} \ .$$

For instance, H_+ is constructed by only accounting for the increments of $A(n)$ satisfying a strictly positive drift $a_n - C$. H_+ and H_- will appear in the upper bounds, whereas H'_+ and H'_- will appear in the lower bounds.

Next, we give the main results of the paper, i.e., bounds on the queue and loss distributions.

4.1 Queue Distribution. Underload ($\rho < 1$)

Assume the '*stability*' condition $\rho < 1$ which implies that $\theta > 0$ in the construction of arrival-envelopes; the system is nevertheless stable for any value of ρ.

Theorem 3 (Queue distribution (underload)). *In the queueing scenario above, assume the flow A admits a martingale-envelope M_n with parameters $\theta > 0$ and h. Then, the following upper bound holds for the queue size distribution Q, for $0 < \sigma \leq K$*

$$\mathbb{P}(Q \geq \sigma) \leq \frac{\mathbb{E}[h(a_0)] - H_- e^{\theta(\sigma-K-C)}}{H_+ e^{\theta\sigma} - H_- e^{\theta(\sigma-K-C)}} \ .$$

Further, if $a_n \leq a_{max}$ for some constant $a_{max} > 0$ and all $n \geq 0$, then additionally the following lower bound on Q holds

$$\mathbb{P}(Q \geq \sigma) \geq \frac{\mathbb{E}[h(a_0)] - H'_- e^{\theta(\sigma-K)}}{H'_+ e^{\theta(\sigma+a_{max}-C)} - H'_- e^{\theta(\sigma-K)}} \ .$$

We tacitly assume that the denominators in the bounds are positive (all our examples satisfy this property); otherwise the inequality signs have to be reversed and the bounds themselves change, e.g., the upper becomes lower.

By letting $K \to \infty$ the upper bound in the infinite queue is

$$\mathbb{P}(Q \geq \sigma) \leq \frac{\mathbb{E}[h(a_0)]}{H_+} e^{-\theta\sigma} \ ,$$

which was obtained in [44], whereas the corresponding lower bound is

$$\mathbb{P}(Q \geq \sigma) \geq \frac{\mathbb{E}[h(a_0)]}{H'_+} e^{-\theta(\sigma+a_{\max}-C)} \ .$$

Proof. Consider Q's representation from (4). For $0 < \sigma \leq K$, define the stopping time N as the first point in time where the process within the max-operator first exceeds σ, i.e.,

$$N := \min\Big\{n \geq 0 \ \Big| \ \min\Big\{A(n) - nC, \min_{0 \leq m < n} \{A(m) - mC + K\}\Big\} \geq \sigma\Big\} \ . \tag{8}$$

By definition, it holds that[2]

$$\{Q \geq \sigma\} = \{N < \infty\} \ .$$

We now define the key stopping times N_+ and N_- as

$$N_+ := \min\{n \geq 0 \mid A(n) - nC \geq \sigma\} \tag{9}$$

$$N_- := \min\{m \geq 0 \mid A(m) - mC < \sigma - K\} \ . \tag{10}$$

Their construction is directly related to the expression of N from (8), and in particular the two terms in the outer 'min'.

Clearly, $P(N_+ = N_-) = 0$. Note also that $N_+, N_- \geq 1$ a.s. because

$$A(0) - 0C = 0 \in [\sigma - K, \sigma) \ .$$

Further, from the 'stability' condition $\rho < 1$, note that the process $A(m) - mC$ has a negative drift and hence N_- is finite a.s.

$$\mathbb{P}(N_- < \infty) = 1 \ . \tag{11}$$

We next show that the three stopping times N, N_+, and N_- are related by the fundamental relationship

$$\{N < \infty\} = \{N_+ < N_-\} \ . \tag{12}$$

Assume first that $N < \infty$. For all $k > N_-$ it holds:

$$\begin{aligned} &\min\Big\{A(k) - kC, \min_{0 \leq m < k} \{A(m) - mC + K\}\Big\} \\ &\leq \min_{0 \leq m < k} \{A(m) - mC + K\} \\ &\leq A(N_-) - N_-C + K \\ &< \sigma \ , \end{aligned}$$

such that necessarily $N \leq N_-$. As obviously $A(N) - NC \geq \sigma$, it follows that $N_+ \leq N$ and thus $N_+ < N_-$.

For the other direction, assume that $N_+ < N_-$. For $m < N_-$ it holds by definition

$$A(m) - mC + K \geq \sigma \ , \tag{13}$$

and hence also

$$\min\Big\{A(N_+) - N_+C, \min_{0 \leq m < N_+} \{A(m) - mC + K\}\Big\} \geq \sigma \ ,$$

and thus $N \leq N_+$. This completes the proof of (12) since $N_- < \infty$.

As a side remark, the finite and infinite-buffer cases are subject to very different behaviors concerning buffer overflows. In the finite case, at $N_+ \wedge N_-$ (which is a.s. finite) we know *for certain* whether $Q \geq \sigma$ or $Q < \sigma$, depending where the minimum is attained. In the infinite case, however, if $Q < \sigma$ then at no point in time we would know this fact for certain (we would know that $Q \geq \sigma$ at time N_+, should it be reached; note that $\mathbb{P}(N_+ = \infty) > 0$).

We can now derive the bounds on $\mathbb{P}(Q \geq \sigma)$. For the upper bound, apply the Optional-Stopping Theorem (see, e.g., Williams [56],

[2]We note that this duality has also been established in the context of ruin probabilities (e.g., Asmussen [2], pp. 1-2).

p. 100) to the martingale-envelope M from Definition 1 and the *bounded* stopping time $N_+ \wedge N_- \wedge n$ (for some $n \geq 0$):

$$\begin{aligned}\mathbb{E}[h(a_0)] &= \mathbb{E}[M_0] = \mathbb{E}[M_{N_+\wedge N_-\wedge n}] \\ &= \mathbb{E}[M_{N_+\wedge N_-} 1_{\{N_+\wedge N_- \leq n\}}] + \mathbb{E}[M_n 1_{\{N_+\wedge N_- > n\}}]\end{aligned} \tag{14}$$

For additional insights into this crucial step see the follow-up discussion from § 4.2.

Let $n \to \infty$. From the Monotone Convergence Theorem we can interchange the limit with the first expectation, and by the finiteness of $N_+ \wedge N_-$ the first expectation converges to $\mathbb{E}[M_{N_+\wedge N_-}]$. From the Bounded Convergence Theorem we can interchange the limit with the second expectation (because $M_n < e^{\theta\sigma}$ on $N_+ \wedge N_- > n$) and thus the second expectation vanishes. Therefore

$$\mathbb{E}[h(a_0)] = \mathbb{E}[M_{N_+\wedge N_-}]\,, \tag{15}$$

which can be expanded as

$$\begin{aligned}\mathbb{E}[h(a_0)] &= \mathbb{E}[M_{N_+} 1_{\{N_+<N_-\}}] + \mathbb{E}[M_{N_-} 1_{\{N_-<N_+\}}] \\ &= \mathbb{E}[M_{N_+} 1_{\{N<\infty\}}] + \mathbb{E}[M_{N_-} 1_{\{N=\infty\}}] \\ &= \mathbb{E}[h(a_{N_+}) e^{\theta(A(N_+)-N_+C)} 1_{\{N<\infty\}}] \\ &\quad + \mathbb{E}[h(a_{N_-}) e^{\theta(A(N_-)-N_-C)} 1_{\{N=\infty\}}] \\ &\geq e^{\theta\sigma}\mathbb{E}[h(a_{N_+}) 1_{\{N<\infty\}}] \\ &\quad + e^{\theta(\sigma-K-C)}\mathbb{E}[h(a_{N_-}) 1_{\{N=\infty\}}] \\ &\geq e^{\theta\sigma} H_+ \mathbb{P}(N<\infty) \\ &\quad + e^{\theta(\sigma-K-C)} H_- \left(1-\mathbb{P}(N<\infty)\right)\,.\end{aligned} \tag{16}$$

In the second line we used the equivalence from (12). In the fifth and sixth lines we used, for the first term, the definition of N_+, and, for the second term, by (13), $A(n) - nC + K \geq \sigma$ for $n < N_-$, and hence

$$\begin{aligned}A(N_-) - N_-C &= A(N_- - 1) - (N_- - 1)C + a_{N_-} - C \\ &\geq \sigma - K + a_{N_-} - C \\ &\geq \sigma - K - C\,.\end{aligned} \tag{17}$$

In the last line we used Definition 2 with the fact that the last increment of the stopped process $A(n)-nC$ must be strictly positive at N_+ and strictly negative at N_-.

Now solve for $\mathbb{P}(N<\infty)$ to obtain:

$$\mathbb{P}(N<\infty) \leq \frac{\mathbb{E}[h(a_0)] - H_- e^{\theta(\sigma-K-C)}}{H_+ e^{\theta\sigma} - H_- e^{\theta(\sigma-K-C)}}\,.$$

Recall by the definition of N that $\mathbb{P}(N<\infty) = \mathbb{P}(Q \geq \sigma)$ and hence the derivation of the upper bound is complete.

For the lower bound we use a similar expansion, except for changing the inequality sign, i.e.,

$$\begin{aligned}\mathbb{E}[h(a_0)] &= \mathbb{E}[M_{N_+\wedge N_-}] \\ &= \mathbb{E}[M_{N_+\wedge N_-} 1_{\{N_+<N_-\}}] + \mathbb{E}[M_{N_+\wedge N_-} 1_{\{N_-<N_+\}}] \\ &= \mathbb{E}[M_{N_+} 1_{\{N<\infty\}}] + \mathbb{E}[M_{N_-} 1_{\{N=\infty\}}] \\ &\leq H'_+ e^{\theta(\sigma+a_{\max}-C)}\mathbb{P}(N<\infty) \\ &\quad + H'_- e^{\theta(\sigma-K)}\left(1-\mathbb{P}(N<\infty)\right)\,.\end{aligned}$$

We again used (12). For the term in the fourth line note that by the definition of N_+,

$$\begin{aligned}A(N_+) - N_+C &\leq A(N_+ - 1) - (N_+ - 1)\,C + a_{\max} - C \\ &< \sigma + a_{\max} - C\,.\end{aligned}$$

For the second term we used the definition of N_- from (10). Solving for $\mathbb{P}(N<\infty)$ completes the proof. □

4.2 Gist of the Technical Approach

Let us now provide some high-level insights into the martingale/stopping-times method at the core of the previous result from Theorem 3, as well as the other main results to follow. Recall the expression for the stationary queue size from (4)

$$Q := \max_{n\geq 0}\left\{A(n) - nC, \min_{0\leq m<n}\left\{A(m) - mC + K\right\}\right\}\,,$$

and visualize it for convenience by the diagonal matrix

$\mathbf{0}$				
$\mathbf{A(1) - C}$	K			
$\mathbf{A(2) - 2C}$	$A(1) - C + K$	K		
$\mathbf{A(3) - 3C}$	$\mathbf{A(2) - 2C + K}$	$A(1) - C + K$	K	
$A(4) - 4C$	$A(3) - 3C + K$	$\mathbf{A(2) - 2C + K}$	$A(1) - C + K$	K
$\vdots$	$\ddots$	$\ddots$	$\ddots$	$\ddots$

For each row a minimum is taken, and the maximum of all these yields Q.

Let us also restate N_+ and N_-

$$\begin{aligned}N_+ &:= \min\{n \geq 0 \mid A(n) - nC \geq \sigma\} \\ N_- &:= \min\{m \geq 0 \mid A(m) - mC < \sigma - K\}\,,\end{aligned}$$

and consider $\mathbb{P}(Q \geq \sigma)$ which is the central quantity to be estimated.

Suppose for instance that $N_- < N_+$, say at $N_- = 2$, which means that $A(2) - 2C + K < \sigma$. Additionally, $A(n) - nC < \sigma$ for $n = 0, 1, 2, 3$ because $N_+ > N_-$. In other words, the maximum of the minimums for the first four rows is smaller than σ; see the bold-font quantities, which are all smaller than σ. Because $A(2) - 2C + K$ will appear in all the rows from the forth one onwards, the queue size can never reach σ, i.e., $Q < \sigma$. The other case when $N_+ > N_-$ can be visualized similarly.

The gist of our main results is to essentially inspect the system at the two stopping times, N_+ and N_-, which essentially retain all the information about the event of interest $\{Q \geq \sigma\}$, in the sense that $\{Q \geq \sigma\} = \{N_+ < N_-\}$ a.s.

As natural as they appear, stopping times are misleading. For a quick illustration consider an iid Bernoulli process X_n and define the stopping time

$$N := \min\{n : X_n = 1\}\,.$$

Recall the definition of a stopping time, in particular that the event $\{N = n\}$ entirely depends on the history up to time n, i.e., the values $X_1, X_2, \ldots, X_n$ only. By inspecting the 'system' at the (random) time N, the iid property is lost because $E[X_N] = 1 \neq E[X_1]$ from the very definition of N.

A *counter-trick* is to consider martingales X_n, which do preserve the property of the system at stopping times, in the sense that $E[X_N] = E[X_1]$; this is essentially Doob's Optional-Stopping

Theorem (OST)[3] whose application was essential in the proof of Theorem 3; see (14).

In our queueing 'system' the driving factor is the cumulative drift $A(n)-nC$. In a underload regime ($E[A(1)] < C$), this process is in fact a supermartingale for which OST would hold as well, at the expense however of introducing a conceivably loose inequality which would severely weaken the accuracy of the bound for $\mathbb{P}(Q \geq \sigma)$ itself. To avoid this pitfall the key idea is to transform the cumulative drift process into a martingale, for which OST does hold with equality. This is essentially Kingman's original idea for bounding GI/G/1 queues ([34]), and which evolved as a simple technique for studying many queueing systems, including the case of Markovian arrivals (e.g., Chang [14] or Duffield [19]).

An important observation is on the robustness of the martingale-envelope model from Def. 1 in the sense that our main results (Theorem 3 and the later ones) apply to *any* arrival model for which martingale-envelope representations exist; several examples will be provided in § 5, § 6, and § B.1.

4.3 Queue Distribution. Overload ($\rho > 1$)

We now assume $\rho > 1$ which implies that arrival-envelopes have an exponent $-\theta$ instead of θ (where $\theta > 0$).

Theorem 4 (Queue distribution (overload)). *In the queueing scenario above, assume that the flow A admits a martingale-envelope M_n with parameters $-\theta$ and h, where $\theta > 0$. Then, the following upper bound holds for the queue size distribution Q, for $0 < \sigma \leq K$*

$$\mathbb{P}(Q \geq \sigma) \leq \frac{\mathbb{E}[h(a_0)] - H'_- e^{-\theta(\sigma-K-C)}}{H'_+ e^{-\theta\sigma} - H'_- e^{-\theta(\sigma-K-C)}} .$$

Further, if $a_n \leq a_{max}$ for some constant $a_{max} > 0$ and all $n \geq 0$, then additionally the following lower bound on Q holds

$$\mathbb{P}(Q \geq \sigma) \geq \frac{\mathbb{E}[h(a_0)] - H_- e^{-\theta(\sigma-K)}}{H_+ e^{-\theta(\sigma+a_{max}-C)} - H_- e^{-\theta(\sigma-K)}} .$$

Again, we tacitly assume that the denominators in the bounds are positive (all our examples satisfy this property); otherwise the inequality signs have to be reversed. Note that these bounds are similar to those in the underload case; the differences are θ vs. $-\theta$ and the 'H' parameters.

Proof. We only sketch the proof for the upper bound, which proceeds as the proof for the (underload) upper bound.

The first key difference is that N_+ rather than N_- is a.s. finite. It still holds however that

$$\{N < \infty\} = \{N_+ < N_-\} .$$

Proceeding further we have

$$\begin{aligned}\mathbb{E}[h(a_0)] &= \mathbb{E}[M_{N_+} 1_{\{N<\infty\}}] + \mathbb{E}[M_{N_-} 1_{\{N=\infty\}}] \\ &\leq H'_+ e^{-\theta\sigma} \mathbb{P}(N < \infty) \\ &\quad + H'_- e^{-\theta(\sigma-K-C)} (1 - \mathbb{P}(N < \infty)) .\end{aligned}$$

Here we used the negativity of θ and

$$\begin{aligned}A(N_-) - N_- C &= A(N_- - 1) - (N_- - 1)C + a_{N_-} - C \\ &\geq \sigma - K - C ,\end{aligned}$$

from the definition of N_- and $a_{N_-} \geq 0$. The proof is complete by solving for $\mathbb{P}(N < \infty) = \mathbb{P}(Q \geq \sigma)$. □

[3] See Williams [56], p. 100, for the precise technical conditions under which OST holds.

4.4 Loss Distribution. Underload ($\rho < 1$)

To analyze the loss process we need the stationary distribution of the instantaneous arrivals. Assuming for convenience a discrete range of values b_i, denote

$$\pi_i = \mathbb{P}(a_1 = b_i)$$

over some countable set with index i.

The next result gives bounds on the loss distribution in the underload regime $\rho < 1$.

Theorem 5 (Loss distribution). *In the queueing scenario above, assume the flow A admits a martingale-envelope with parameters $\theta > 0$ and h. Then the following upper bound holds for the distribution of the loss process for $\sigma > 0$*

$$\begin{aligned}\mathbb{P}(L \geq \sigma) \leq &\sum_{\sigma+C \leq b_i < \sigma+C+K} \pi_i \frac{h(b_i) e^{\theta(b_i-C)} - H_- e^{\theta(\sigma-C)}}{H_+ e^{\theta(\sigma+K)} - H_- e^{\theta(\sigma-C)}} \\ &+ \mathbb{P}(a_1 \geq \sigma + C + K) .\end{aligned} \tag{18}$$

Further, if $a_n \leq a_{max}$ for some constant $a_{max} > 0$ and all $n \geq 0$, then additionally the following lower bound on L holds

$$\begin{aligned}\mathbb{P}(L \geq \sigma) \geq &\sum_{\sigma+C \leq b_i < \sigma+C+K} \pi_i \frac{h(b_i) e^{\theta(b_i-C)} - H'_- e^{\theta\sigma}}{H'_+ e^{\theta(\sigma+K+a_{max}-C)} - H'_- e^{\theta\sigma}} \\ &+ \mathbb{P}(a_1 \geq \sigma + C + K) .\end{aligned}$$

The explanation for the second term is that $L \geq \sigma$ on $a_1 \geq \sigma + C + K$ (a_1 is the last increment in reverse time). Note that the bounds match in a bufferless regime ($K = 0$), i.e., $\mathbb{P}(a_1 \geq \sigma + C)$.

The case when a_1 is a continuous random variable can be treated almost identically; the only difference is that the sums in Theorem 5 become integrals, and the π_i's are replaced by a_1's density function $f(x)$, provided that it exists. From a computational perspective, however, the integrals may not have a closed-form expression due to the factor $h(b_i)$ ($h(x)$ in continuous-time).

Proof. The proof is similar to the one for the queue size but with a fundamental difference. Define first the stopping time

$$\begin{aligned}N := \min \Big\{ n \geq 1 \;\Big|\; &\min \Big\{ A(n) - nC - K, \\ &\min_{1 \leq m < n} \{A(m) - mC\} \Big\} \geq \sigma \Big\} .\end{aligned} \tag{19}$$

The crucial observation is that in order to have a loss event then the last increment triggering the loss must satisfy

$$a_1 \geq \sigma + C .$$

(recall that a_1 plays the role of the last increment in reverse time).

Define now $\mathbb{P}_i$ as the underlying probability measure conditioned on $a_1 = b_i$, and also the stopping times

$$N_+ := \min \{n \geq 1 \mid A(n) - nC \geq \sigma + K\} \tag{20}$$

$$N_- := \min \{m \geq 1 \mid A(m) - mC < \sigma\} , \tag{21}$$

which are slightly different than those defined for the queue size process. Similarly, we next show that

$$\{L \geq \sigma\} = \{N < \infty\} = \{N_+ < N_-\} . \tag{22}$$

The first equality holds from the definition of N; this step was also used in the proof of Theorem 3. Assuming $N < \infty$ we have that $A(N) - NC - K \geq \sigma$ and hence $N_+ \leq N$. Also, $A(m) - mC \geq \sigma$ for all $m \leq N$ and hence $N_- > N$. Therefore $N_+ < N_-$; note that $P(N_+ = N_-) = 0$.

In the other direction, assume that $N_+ < N_-$. Because $A(m) - mC \geq \sigma$ for all $m < N_-$ it follows that

$$\min_{1 \leq m < N_+} \{A(m) - mC\} \geq \sigma \, .$$

Since $A(N_+) - N_+C \geq \sigma + K$ (from the definition of N_+) we obtain that $N \leq N_+$ and hence $N < \infty$ (note that N_- is finite and hence N_+ as well).

The next observation is that on $\mathbb{P}_i$ the process M_n starting at $M_1 = h(b_i)\, e^{\theta(b_i - C)}$ remains a martingale. The rest of the proof proceeds similarly as in Theorem 3 by invoking the OST for M_n, i.e.,

$$\mathbb{E}_i [M_1] = h(b_i)\, e^{\theta(b_i - C)} = \mathbb{E}_i \left[M_{N_+ \wedge N_- \wedge n} \right] \, ,$$

for any $n \geq 1$, where $\mathbb{E}_i$ is the expectation under $\mathbb{P}_i$.

Taking the limit in n we have

$$\begin{aligned} \mathbb{E}_i[M_1] &= \mathbb{E}[M_{N_+} 1_{\{N < \infty\}}] + \mathbb{E}[M_{N_-} 1_{\{N = \infty\}}] \\ &\geq H_+ e^{\theta(\sigma + K)} \mathbb{P}(N < \infty) \\ &\quad + H_- e^{\theta(\sigma - C)} \left(1 - \mathbb{P}(N < \infty)\right) \, . \end{aligned}$$

Here we used the positivity of θ and

$$\begin{aligned} A(N_-) - N_-C &= A(N_- - 1) - (N_- - 1)\, C + a_{N_-} - C \\ &\geq \sigma - C \, , \end{aligned}$$

The proof for the upper bound is complete by deconditioning on b_i; note in particular that

$$\mathbb{P}\left(N < \infty \mid b_i \geq \sigma + C + K\right) = 1 \, .$$

The proof for the lower bound proceeds similarly as in Theorem 3. □

4.5 Loss Distribution. Overload ($\rho > 1$)

The overload extension proceeds similarly as the overload queue distribution from § 4.3.

Theorem 6 (Loss distribution (overload)). *In the queueing scenario above, assume that the flow A admits a martingale-envelope M_n with parameters $-\theta$ and h, where $\theta > 0$. Then, the following upper bound holds for the distribution of the loss process for $\sigma > 0$*

$$\begin{aligned} \mathbb{P}(L \geq \sigma) \leq & \sum_{\sigma + C \leq b_i < \sigma + C + K} \pi_i \frac{h(b_i)\, e^{-\theta(b_i - C)} - H'_- e^{-\theta(\sigma - C)}}{H'_+ e^{-\theta(\sigma + K)} - H'_- e^{-\theta(\sigma - C)}} \\ & + \mathbb{P}(a_1 \geq \sigma + C + K) \, , \end{aligned} \tag{23}$$

Further, if $a_n \leq a_{max}$ for some constant $a_{max} > 0$ and all $n \geq 0$, then additionally the following lower bound on L holds

$$\begin{aligned} \mathbb{P}(L \geq \sigma) \geq & \sum_{\sigma + C \leq b_i < \sigma + C + K} \pi_i \frac{h(b_i)\, e^{-\theta(b_i - C)} - H_- e^{\theta\sigma}}{H_+ e^{-\theta(\sigma + K + a_{max} - C)} - H_- e^{-\theta\sigma}} \\ & + \mathbb{P}(a_1 \geq \sigma + C + K) \, . \end{aligned}$$

We note that the only changes from the underload bounds from Theorem 5 are the sign change of θ and the H'_+ and H'_- parameters, instead of H_+ and H_-. Concerning the proof itself, the only significant difference is that now N_+ is finite a.s.

5 CASE STUDY 1: IID ARRIVALS

We first address the underload regime, and then the overload and border regimes.

5.1 Underload regime ($\rho < 1$)

Assume that the process $(a_n)_{n \in \mathbb{Z}}$ is given by an iid family of random variables. The following lemma (see Lemma 14 in [44]) shows the existence of a corresponding martingale-envelope:

Lemma 7. *For iid instantaneous arrivals, let θ be defined by*

$$\theta := \sup \left\{ \theta \geq 0 \;\middle|\; E[e^{\theta a_1}] \leq e^{\theta C} \right\} \, .$$

Then the flow A admits a martingale-envelope with parameters θ and $h \equiv 1$, i.e.,

$$M_n = e^{\theta(A(n) - Cn)} \, .$$

The existence of θ is guaranteed under the tacit assumption $E[a_1] < C < \sup a_1$ to avoid the trivial scenario of an always empty queue.

With the observation that for the constant function h it clearly holds $H_+ = H_- = 1$, the results from Theorems 3 (queue distribution) and Theorem 5 (loss distribution) apply immediately.

For instance, an upper bound on the queue size distribution is

$$\mathbb{P}(Q \geq \sigma) \leq \frac{1 - e^{\theta(\sigma - K - C)}}{e^{\theta\sigma} - e^{\theta(\sigma - K - C)}} \, . \tag{24}$$

Improved bounds, relative to Theorem 3, can be obtained in the iid case using an idea from Ross [48] (see also [17])

Corollary 8 (Queue distribution; improved bounds). *In the queueing scenario above*

$$\frac{1 - \beta_+ e^{\theta(\sigma - K)}}{\alpha_+ e^{\theta\sigma} - \beta_+ e^{\theta(\sigma - K)}} \leq \mathbb{P}(Q \geq \sigma) \leq \frac{1 - \beta_- e^{\theta(\sigma - K)}}{\alpha_- e^{\theta\sigma} - \beta_- e^{\theta(\sigma - K)}} \, ,$$

where

$$\alpha_- = \inf_{x > C} \mathbb{E}\left[e^{\theta(a_1 - x)} \mid a_1 \geq x \right]$$

and

$$\beta_- = \inf_{0 \leq x < C} \mathbb{E}\left[e^{\theta(a_1 - x)} \mid a_1 < x \right] ,$$

whereas α_+ and β_+ are the same as α_- and β_- except for replacing the 'inf' by 'sup'.

The proof is given in Appendix § A. An important remark is that, unlike Theorem 3, the lower bounds hold in the case when a_1 has unbounded support. When $K = \infty$ the bounds are exact in the case of exponential arrivals because $\alpha_- = \alpha_+$, as a direct consequence of the memoryless property.

In certain cases, the parameters $\alpha_-, \beta_-, \alpha_+, \beta_+$ can be easily computed. For instance, if a_1 has an increasing failure rate distribution then $\mathbb{E}\left[e^{\theta(a_1 - x)} \mid a_1 \geq x \right]$ and $\mathbb{E}\left[e^{\theta(a_1 - x)} \mid a_1 < x \right]$ are non-increasing (see Ross [48], Shaked and Shanthikumar [51] (Theorem 1.A.30), and Nanda *et al.* [42]). We also point out that, to

simplify notation, we tacitly consider the range of x to be included in the support of a_1.

Improved bounds on the loss distribution, relative to Theorem 5, can be obtained similarly:

Corollary 9 (Loss distribution; improved bounds). *In the queueing scenario above*

$$\mathbb{P}(L \geq \sigma) \leq \sum_{\sigma+C \leq b_i < \sigma+C+K} \pi_i \frac{h(b_i)\, e^{\theta(b_i-C)} - \beta_- e^{\theta\sigma}}{\alpha_- e^{\theta(\sigma+K)} - \beta_- e^{\theta\sigma}} + \mathbb{P}(a_1 \geq \sigma + C + K)\,. \tag{25}$$

and

$$\mathbb{P}(L \geq \sigma) \geq \sum_{\sigma+C \leq b_i < \sigma+C+K} \pi_i \frac{h(b_i)\, e^{\theta(b_i-C)} - \beta_+ e^{\theta\sigma}}{\alpha_+ e^{\theta(\sigma+K)} - \beta_+ e^{\theta\sigma}} + \mathbb{P}(a_1 \geq \sigma + C + K)\,. \tag{26}$$

where α_-, β_-, α_+, and β_+ are given in Corollary 8.

Alike for the queue distribution, the loss lower bounds are more general than those from Theorem 5 in that the arrivals are not restricted to finite support.

5.2 Overload regime ($\rho > 1$)

Assume that $\rho > 1$ and the additional constraint

$$\inf a_1 < C$$

to avoid the trivial scenario of an always full queue.

Lemma 10. *For iid instantaneous arrivals, and $\rho > 1$, let $\theta > 0$ be defined by*

$$\theta := \sup \left\{ \theta \geq 0 \;\middle|\; E[e^{-\theta a_1}] \leq e^{-\theta C} \right\}\,.$$

Then the flow A admits a martingale-envelope with parameters $-\theta$ and $h \equiv 1$, i.e.,

$$M_n = e^{-\theta(A(n)-Cn)}\,.$$

The proof is given in Appendix § A.

Upper and lower bounds follow directly from Theorem 4 by noting that $H'_+ = H'_- = 1$. Improved bounds (and more general lower bounds) follow as in the underload regime (see Corollary 8):

Corollary 11 (Queue distribution; improved bounds). *In the queueing scenario above*

$$\frac{1-\beta_- e^{-\theta(\sigma-K)}}{\alpha_- e^{-\theta\sigma} - \beta_- e^{-\theta(\sigma-K)}} \leq \mathbb{P}(Q \geq \sigma) \leq \frac{1-\beta_+ e^{-\theta(\sigma-K)}}{\alpha_+ e^{-\theta\sigma} - \beta_+ e^{-\theta(\sigma-K)}}\,,$$

where

$$\alpha_- = \inf_{x>C} \mathbb{E}\left[e^{-\theta(a_1-x)} \mid a_1 \geq x \right]$$

and

$$\beta_- = \inf_{0 \leq x < C} \mathbb{E}\left[e^{-\theta(a_1-x)} \mid a_1 < x \right],$$

whereas α_+ and β_+ are the same as α_- and β_- except for replacing the 'inf' by 'sup'.

Unlike Theorem 4, the lower bounds now hold in the case when a_1 has unbounded support. The proof is almost identical to that of Corollary 8. We also note that the conditional expectations in the expressions for α_- and β_- are non-decreasing in the case when a_1 has an increasing failure rate distribution.

Improved upper bounds, and more general lower bounds, can also be obtained for the loss distribution.

Corollary 12 (Loss distribution; improved bounds). *In the queueing scenario above*

$$\mathbb{P}(L \geq \sigma) \leq \sum_{\sigma+C \leq b_i < \sigma+C+K} \pi_i \frac{h(b_i)\, e^{-\theta(b_i-C)} - \beta_+ e^{-\theta\sigma}}{\alpha_+ e^{-\theta(\sigma+K)} - \beta_+ e^{-\theta\sigma}} + \mathbb{P}(a_1 \geq \sigma + C + K)\,. \tag{27}$$

and

$$\mathbb{P}(L \geq \sigma) \geq \sum_{\sigma+C \leq b_i < \sigma+C+K} \pi_i \frac{h(b_i)\, e^{-\theta(b_i-C)} - \beta_- e^{-\theta\sigma}}{\alpha_- e^{-\theta(\sigma+K)} - \beta_- e^{-\theta\sigma}} + \mathbb{P}(a_1 \geq \sigma + C + K)\,. \tag{28}$$

where α_-, β_-, α_+, and β_+ are given in Corollary 11.

Note that this result is the same as in Corollary 9 except for changing the sign of θ and interchanging α_- with α_+ and β_- with β_+; recall also the differences between Theorems 5 and 6.

5.3 Border regime ($\rho = 1$)

We now enforce the condition $\rho = 1$ and additionally

$$\inf a_1 < C < \sup a_1$$

to avoid trivial scenarios.

Corollary 13 (Queue distribution ($\rho = 1$)). *In the scenario above, the queue size distribution satisfies for $0 \leq \sigma \leq K$*

$$\mathbb{P}(Q \geq \sigma) \leq 1 - \frac{\sigma}{C+K}\,. \tag{29}$$

This bound is particularly interesting because it is insensitive to the arrivals' distribution; as simulations will show, the bound is quite accurate for several distributions.

Proof. The proof does not involve a martingale-envelope because the construction of θ from Lemma 7 would yield the trivial martingale $M_n = 1$. Instead we apply l'Hôpital rule in (24), i.e.,

$$\lim_{\theta \downarrow 0} \frac{1 - e^{\theta(\sigma-K-C)}}{e^{\theta\sigma} - e^{\theta(\sigma-K-C)}} = 1 - \frac{\sigma}{C+K}\,. \tag{30}$$

Note that $\theta \downarrow 0$ in the constructions from Lemma 7 when $\rho \uparrow 1$. However, because each θ is *implicitly* obtained from each ρ, the continuity/differentiability of the function

$$f(\rho) := \frac{1 - e^{\theta(\rho)(\sigma-K-C)}}{e^{\theta(\rho)\sigma} - e^{\theta(\rho)(\sigma-K-C)}}$$

is not guaranteed, where $\theta(\rho)$ is the corresponding value of θ from Lemma 7 for a specific ρ. The proof is complete by applying the sequential characterization of limits from real analysis: if $\lim_{x\to 0} g(x) = L$ exists then $\lim_{x_n \to 0} g(x_n) = L$ for any sequence $x_n \to 0$ with $x_n \neq 0$. In our case the values of x_n are taken by the values of $\theta(\rho)$ from Lemma 7, and therefore the limit from (30) applies to $f(\rho)$ when $\rho \uparrow 1$. □

An almost analogous result is that the number of jobs N in the M/M/1/K queue satisfies

$$\mathbb{P}(N \geq \sigma) = 1 - \frac{\sigma}{K+1}\,,$$

for $\sigma = 0, 1, \ldots, K$; for related results concerning insensitivity properties in queueing networks see Taylor [53]. Our queueing system can be viewed as a D/G/1/K queue, i.e., equally-spaced arrivals and general service times driven by the distribution of a_1 and the service rate C. However, its underlying dynamics (see the recursion from (3)) are slightly different from those of the D/G/1/K queue; one reason is that our queue process Q is measured in fluid arrivals rather than number of jobs (for a follow-up discussion see § 8).

Using the same limit argument we can derive a simplified upper bound on the loss distribution:

Corollary 14 (Loss distribution ($\rho = 1$)). *The loss distribution satisfies for $\sigma > 0$*

$$\mathbb{P}(L \geq \sigma) \leq \sum_{\sigma+C \leq b_i < \sigma+C+K} \pi_i \frac{b_i - \sigma}{K + C} + \mathbb{P}(a_1 \geq \sigma + C + K) \ . \quad (31)$$

Unlike the queue distribution from (29), the loss distribution does depend on the distribution of the increments a_n.

5.4 Bounds vs Simulations

5.4.1 Exact Bounds. We first consider a simple case in which the bounds are exact. Let a_n be Bernoulli random variables with

$$\begin{cases} \mathbb{P}(a_1 = 2) = \frac{\rho}{2} \\ \mathbb{P}(a_1 = 0) = 1 - \frac{\rho}{2} \ , \end{cases}$$

such that $\mathbb{E}[a_1] = \rho$; the capacity is $C = 1$ and the buffer size is $K = 10$ (or any integer value).

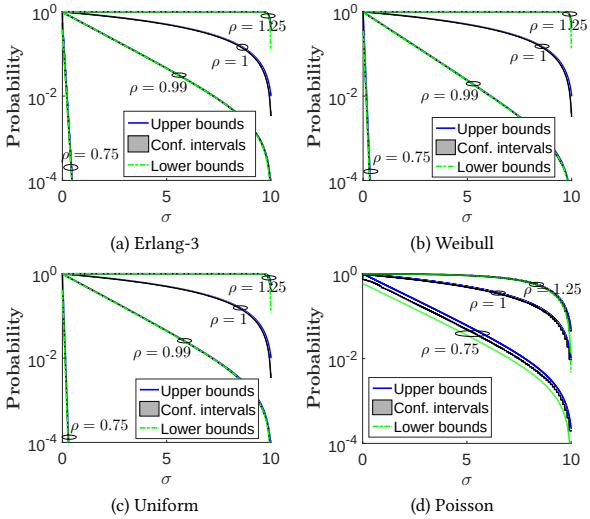
Consider the upper bounds on the queue distribution from Corollaries 8 (underload) and 11 (overload). The two bounds match those from Theorems 3 and 4 (e.g., $\alpha_- = 1$ and $\beta_- = e^{-\theta C}$ in Corollary 8). Moreover, the bounds are in fact exact, for integer values of σ. To see that, it is instructive to recall the proof of the upper bound, and in particular (16) where we made use of two inequalities:

$$\begin{cases} A(N_+) - N_+ C \geq \sigma \\ A(N_-) - N_- C \geq \sigma - K - C \ . \end{cases}$$

From the definitions of N_+ and N_-, and the parameters of a_1, these inequalities do hold as equalities, because the instantaneous drift $a_n - C$ is either -1 or 1.

5.4.2 Simulations. Next we consider several distributions for a_1 and compare the bounds against simulations shown in terms of Wilson confidence intervals, which are recommended for estimating the success probability of Binomial distributions [12]. Note that for any σ the empirical distribution $\mathbb{P}(\hat{Q} \geq \sigma)$ follows a binomial law $\mathbb{B}(n, p)$, where n is the number of samples and $p = P(Q \geq \sigma)$ is the success probability to be estimated. We use $n = 10^9$ in all our simulations; each sample is the queue/loss size experienced by the 10^5th packet, starting from an empty system.

In Fig. 2(a,b) we consider Erlang-3 and Weibull (with shape parameter 2) distributions for a_1; the corresponding rate and scale parameters (both denoted by λ) are determined from C and the utilization. For instance, in the Weibull case, $\lambda = \frac{2\rho C}{\sqrt{\pi}}$. We also consider the Uniform ($\mathcal{U}[0, 2\mathbb{E}[a_1]]$) and Poisson distributions in (c) and (d), with $\mathbb{E}[a_1]$ determined from C and ρ. In all figures we



(a) Erlang-3 (b) Weibull (c) Uniform (d) Poisson

Figure 2: Queue distribution for iid arrivals ($C = 0.1$, $K = 10$)

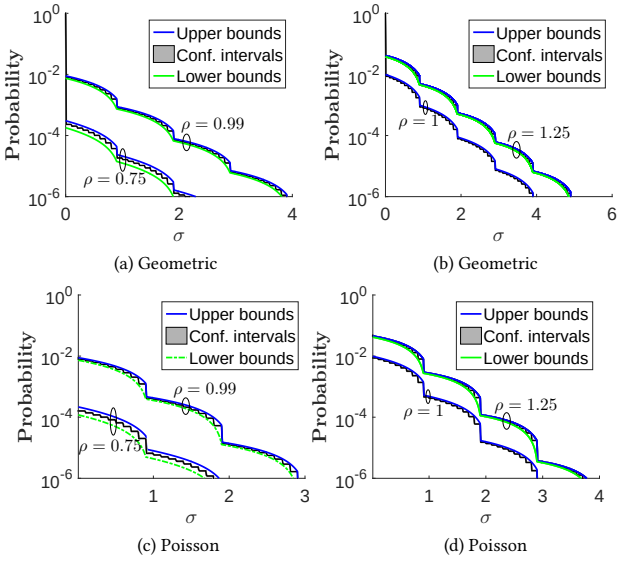
include the upper/lower bounds from Corollaries (8) and (11) for $\rho < 1$ and $\rho > 1$, respectively, and the simplified upper bound from (29) for $\rho = 1$. All the bounds can be derived in closed-form except for the parameter θ, and except those for the Poisson case which requires numerical procedures for estimating the α's and β's parameters.

In heavy-traffic the upper bounds, simulations, and lower bounds are visually almost indistinguishable. The shown upper bounds from Corollaries (8) and (11) only negligibly improve upon those from Theorems 3 and 4 (not shown here). We have also experienced a very slow convergence of the tails at utilization $\rho = 1.25$ when using fewer samples (e.g., 10^6 instead of 10^9). A possible explanation is that convergence is provably very slow in finite-buffer queues, more precisely it can have an order of $O(t^{-\gamma})$ for some parameter γ, where t is time (see Bratiichuk [11]); this slow convergence rate raises further computational concerns on existing recursive algorithms (recall the discussion from § 2).

An interesting observation is that there is a large gap between the plots for $\rho = 0.99$ and $\rho = 1$. This is not the case however in the Poisson case where we omitted the plot for $\rho = 0.99$ which almost overlaps with that for $\rho = 1$. The Poisson case further stands out because queues are significantly larger at $\rho = 0.75$ than in the other three cases; the reason lies in the magnitude of the coefficient of variation of the Poisson increments, i.e., about 6-fold larger than in the other cases for the given set of parameters.

We also note that the bounds for $\rho = 1$ slightly deteriorate in the tail. A possible reason is that the shape of the bound from (29) does not capture the fact that $\mathbb{P}(Q \geq K + \varepsilon) = 0$ for $\varepsilon > 0$.

The bounds for the loss distribution are illustrated in Fig. 3 for iid Geometric and Poisson arrivals with the same mean; we use the improved ones from this section and also the upper ones from (31) when $\rho = 1$. As it was the case for the queue distribution, there is only a negligible difference in the Poisson case (and also in the Geometric case) between $\rho = 0.99$ and $\rho = 1$, for which reason we split the figures into (a,b) and (c,d). The shown bounds only



(a) Geometric (b) Geometric

(c) Poisson (d) Poisson

Figure 3: Loss distribution for iid Geometric and Poisson ($C = 0.1$, $K = 10$)

marginally improve those from Theorem 5/(23); note that lower bounds are not available through Theorem 5 given the unbounded support of the two distributions.

6 CASE STUDY 2: MARKOVIAN ARRIVALS

6.1 Markov Modulated Processes; Constant Size Packets

Consider a Markov (modulating) chain X_n with state space $\{1, \ldots, S\}$, transition matrix $T \in \mathbb{R}^{S\times S}$, i.e., $T(i,j) := \mathbb{P}(X_n = j \mid X_{n-1} = i)$, and a rate function $r : \{1, \ldots, S\} \to \mathbb{R}$ such that

$$a_n := r(X_n) \, .$$

We assume that X_n has a steady-state distribution denoted by $\pi = (\pi_i)_{1\le i\le S}$; moreover, X_n starts in the steady-state.

Let the transform matrix $T_\theta \in \mathbb{R}^{S\times S}$ for $\theta > 0$ as

$$T_\theta(i,j) := T(i,j)e^{\theta r(j)} \, , \quad \text{for } 1 \le i,j \le S \, .$$

Moreover, let $\lambda(\theta) \in \mathbb{R}$ denote the spectral radius of

$$\Pi^{-1} T_\theta \Pi \, ,$$

where Π is the diagonal matrix formed from the π's, and let $\mathbf{v}_\theta = (v_1, v_2, \ldots, v_S) \in \mathbb{R}^S$ denote a corresponding (right) eigenvector. By the Perron-Frobenius theorem, $\lambda(\theta)$ is the maximal positive eigenvalue and $\mathbf{v}_\theta$ can be chosen to be positive.

The following lemma (see Lemma 16 in [44]) provides the martingale-envelope for the reversed process, denoted by $A(n)$ as usual:

Lemma 15. *For Markov-modulated arrivals, let*

$$\theta_{mar} := \max \left\{ \theta \ge 0 \;\middle|\; \lambda(\theta) \le e^{\theta C} \right\} \, ,$$

then the flow A admits the martingale-envelope

$$M_n := h(a_n) e^{\theta(A(n) - nC)}$$

for $n \ge 0$, with $h(r(i)) = v_i$ for $i = 1, 2, \ldots, S$.

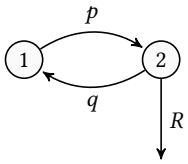


Figure 4: An MMOO process; R arrivals are produced in each time-slot while in state '2'

For the upper and lower thresholds H_+ and H_- from Definition 2 it holds

$$H_+ := \inf \{h(r(i)) \mid r(i) > C\}$$

and

$$H_- := \inf \{h(r(i)) \mid r(i) < C\} \, ,$$

and similarly for H'_+ and H'_-. The bounds on the queue size and loss from Theorems 3 and 5, respectively, apply immediately.

6.1.1 Example 1: MMOO. One of the simplest examples of a Markov-modulated process is the *Markov-Modulated On-Off* (MMOO), i.e., $S := 2$,

$$T := \begin{pmatrix} 1-p & p \\ q & 1-q \end{pmatrix} \, ,$$

for some probabilities p and q, and $r(1) := 0$, and $r(2) := R$, for a peak rate $R > 0$ (see Fig. 4). For the transformed matrix T_θ denote by v_1 and v_2 as the components of the eigenvector corresponding to the spectral radius $\lambda(\theta)$.

6.1.2 Example 2: An aggregate of MMOO's. Let us now consider the more general case of multiplexing N independent MMOO processes, each defined as earlier with identical parameters. Assume the stability condition $NR\frac{p}{p+q} < C$ and the non-trivial situation when $NR > C$ (otherwise the queue would always be empty).

Let us also assume the burstiness condition

$$p + q < 1 \, ,$$

under which it holds that $v_1 \le v_2$ (see [13]). Clearly, the transition matrix of the underlying Markov chain for the aggregate process can be computed, albeit in a quite cumbersome form, and one can further construct a corresponding martingale as in Lemma 15.

A numerically much more efficient technique is to use the statistical independence of the MMOO's (see [44]). By first constructing a martingale $M_i(n)$ for each individual arrival process $A_i(n) = \sum_k a_{i,k}$ as in Lemma 15 (with normalized capacity $c := \frac{C}{N}$), i.e.,

$$M_{i,n} = h_i(a_{i,n}) e^{\theta(A_i(n) - cn)}$$

the aggregate process $A(n) := \sum_i A_i(n)$ has the martingale

$$M_n = h(a_n) e^{\theta(A(n) - Cn)} \, ,$$

where $h()$ is the $(\min, \times)$ convolution of $h_i()'s$, i.e.,

$$h(r) = \min_{r_1 + r_2 + \cdots + r_N = r} h_1(r_1) h_2(r_2) \ldots h_N(r_N) \, .$$

Using the monotonicity property $v_1 \le v_2$ we obtain immediately that

$$h(iR) = v_1^{N-i} v_2^i$$

for $i = 0, 1, \ldots, N$ and also

$$H_- = v_1^N,\ H_+ = v_1^{N-\lceil \frac{C}{R} \rceil} v_2^{\lceil \frac{C}{R} \rceil},$$

$$H'_- = v_1^{N-\lfloor \frac{C}{R} \rfloor} v_2^{\lfloor \frac{C}{R} \rfloor},\ H'_+ = v_2^N .$$

(With abuse of notation $\lfloor \frac{C}{R} \rfloor$ denotes $\frac{C}{R} - 1$ if $\frac{C}{R}$ is an integer.)

The steady-state probabilities of the aggregate chain are

$$\pi_i := \binom{N}{i} \left(\frac{q}{p+q}\right)^{N-i} \left(\frac{p}{p+q}\right)^i ,$$

for $i = 0, 1, \ldots, N$. Then, under these notations, the bounds from Theorems 3 and 5 hold. For instance, the upper bound on the backlog process is

$$\mathbb{P}(Q \geq \sigma) \leq \frac{\mathbb{E}[h(a_0)] - H_- e^{\theta(\sigma-K-C)}}{H_+ e^{\theta\sigma} - H_- e^{\theta(\sigma-K-C)}} .$$

Refined bounds can be immediately obtained as in the iid case, i.e.,

$$\mathbb{P}(Q \geq \sigma) \leq \frac{\mathbb{E}[h(a_0)] - H_- \beta_- e^{\theta(\sigma-K)}}{H_+ \alpha_- e^{\theta\sigma} - H_- \beta_- e^{\theta(\sigma-K)}} , \tag{32}$$

where $\alpha_- = \inf_{x>C,y} \mathbb{E}\left[e^{\theta(a_1-x)} \mid a_1 \geq x, a_0 = y\right]$ whereas $\beta_- = \inf_{0\leq x<C,y} \mathbb{E}\left[e^{\theta(a_1-x)} \mid a_1 < x, a_0 = y\right]$. All the improved bounds are identical as those from the iid case (Corollaries 8 and 9) except for the expanded conditional expectations in the expressions for α_- and β_- (and also of α_+ and β_+) to account for the Markov structure (see the free value y).

Fig. 5 compares the (refined) upper and lower bounds on the queue size distribution for an aggregate of MMOO's against simulations; as mentioned in the iid case as well, only the refined lower bounds significantly improve over those from Theorem 3. The MMOO's parameters are given in the caption; the utilizations $\rho = 0.99$ (heavy-traffic) and $\rho = 0.75$ (moderate) yield different capacities C. The figures indicate that the upper bounds are tight in heavy-traffic and in situations with larger buffers (e.g., (e) vs. (c)). Similar observations hold in Fig. 6 for the loss distribution. We note that $\mathbb{P}(L \geq \sigma) = 0$ when $N = 1$; also, in Fig. 6(b), $\mathbb{P}(L \geq 4) = 0$ due to the parameters' configuration; in (c) and (d) we omit the $\rho = 0.75$ case due to very small probability values.

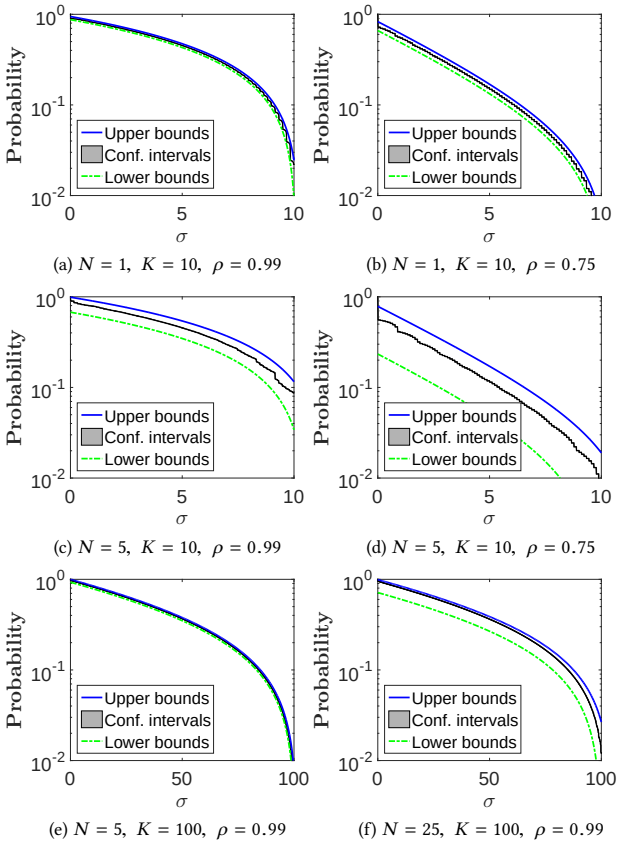
6.2 Markov Modulated Processes; Random Packet Size

Here we briefly consider a generalized version of the previous Markov Modulated Processes in the sense that while in state j a process generates packets of size $r(j)$ with probability p_j, instead of probability 1; with probability $1 - p_j$ no packet is generated. These processes are the discrete-time variant of Markov Modulated Poisson Processes (MMPP).

The martingale representation from Lemma 15 holds immediately. The only difference is that the transform matrix $T_\theta \in \mathbb{R}^{S\times S}$ is now defined as

$$T_\theta(i,j) := T(i,j)\left(1 - p_j + p_j e^{\theta r(j)}\right) , \quad \text{for } 1 \leq i,j \leq S .$$

Immediate extensions to other Markov Modulated processes are also possible. If the arrivals in each state are iid then their moment generating function would replace the factor $\left(1 - p_j + p_j e^{\theta r(j)}\right)$.



(a) $N = 1,\ K = 10,\ \rho = 0.99$ (b) $N = 1,\ K = 10,\ \rho = 0.75$

(c) $N = 5,\ K = 10,\ \rho = 0.99$ (d) $N = 5,\ K = 10,\ \rho = 0.75$

(e) $N = 5,\ K = 100,\ \rho = 0.99$ (f) $N = 25,\ K = 100,\ \rho = 0.99$

Figure 5: Queue distribution (i.e., $\mathbb{P}(Q \geq \sigma)$) for MMOO's ($p = 0.1$, $q = 0.5$, $R = 1$)

If they had a Markov structure then one would need to extend the dimension of T_θ to account for the bivariate (state + arrivals) Markov structure.

7 COMPARISON WITH RELATED WORK

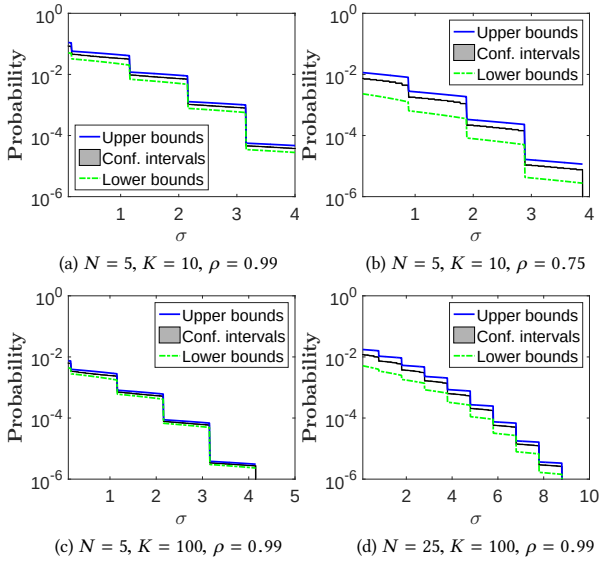
The discrete-time queueing model with finite buffer used in this paper appeared in Cruz and Liu [18]. While slightly different from the D/G/1/K queue, the advantage of this model is the non-recursive formulation for the backlog and loss (recall (4) and (6)). Next we compare our upper bounds on the loss distribution from (27) against related ones for the iid Geometric and Poisson setting from Fig. 3.

First we consider the bound of Cruz and Liu [18]; see Theorem 5.3 in Liu [37] for the actual result. In the Poisson case

$$\mathbb{P}(L \geq \sigma) \leq \inf_{\theta_0 < \theta} \frac{e^{-\theta_0(\sigma+K)}}{e^{\theta_0\left(C - \lambda \frac{e^{\theta_0}-1}{\theta_0}\right)} - 1} ,$$

with θ from Lemma 7. Slightly improved bounds appeared in Ghiassi-Farrokhfal and Ciucu [22]. What both methods have in common is the use of the Union Bound to upper bound $\mathbb{P}(\max_k X_k \geq \sigma)$, where X_k is some stochastic process.

A much improved bound recently appeared in Raeis *et al.* [45] by using an alternative 'min-max' non-recursive formulation for the loss (analogous to the 'max-min' one from (6)). By picking a single



(a) $N = 5$, $K = 10$, $\rho = 0.99$ (b) $N = 5$, $K = 10$, $\rho = 0.75$

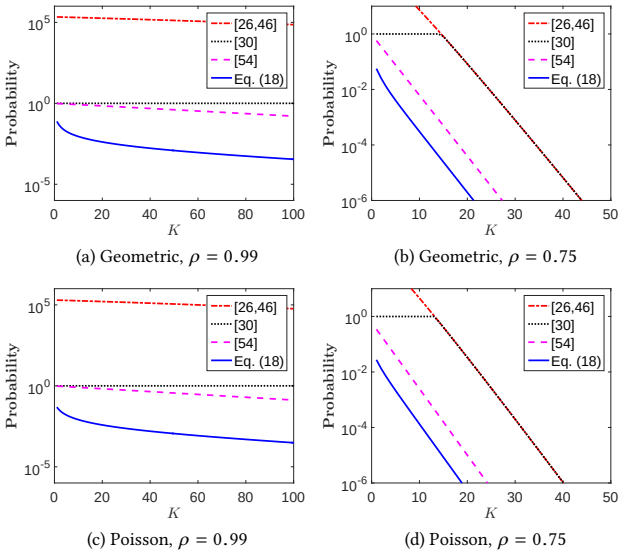
(c) $N = 5$, $K = 100$, $\rho = 0.99$ (d) $N = 25$, $K = 100$, $\rho = 0.99$

Figure 6: Loss distribution (i.e., $\mathbb{P}(L \geq \sigma)$) for MMOO's

point from the outer 'min' operator, Raeis *et al.* [45] deal with the remaining 'max' using the Kingman/Ross martingale methodologies from [34] and [48]. A bound on the overflow probability (see Theorem 4 therein) is

$$\mathbb{P}(L > 0) \leq \frac{1}{\alpha_-} e^{-\theta K} ,$$

with the same θ from Lemma 7 and α_- from Corollary 8.



(a) Geometric, $\rho = 0.99$ (b) Geometric, $\rho = 0.75$

(c) Poisson, $\rho = 0.99$ (d) Poisson, $\rho = 0.75$

Figure 7: Upper bounds on the overflow probability $P(L > 0)$ for iid Geometric and Poisson ($C = 0.1$)

In Fig. 7 we compare all the bounds on the overflow probability (the metric derived in [45]) as a function of the buffer size K. The bounds from [22] simply restrict the bounds from [18] to proper probability values. The improvement from [45] is significant, as a direct consequence of applying martingale-based techniques rather than the Union Bound (for a related discussion see Ciucu and Poloczek [17]). The additional improvement of our bound from (27) is also significant, especially in heavy-traffic. The reason is that we fully exploit the 'max-min' structure through the stopping times N_+ and N_-, unlike the approach from [45] which picks a single point from the 'min' operator.

8 IMPROVEMENTS AND EXTENSIONS

The source of possible inaccuracies of the stopping-times/martingale method lies in two inequalities

$$\begin{cases} A(N_+) - N_+ C \geq \sigma \\ A(N_-) - N_- C \geq \sigma - K - C \, . \end{cases}$$

The latter is also subject to the use of $a_{N_-} \geq 0$ (recall the discussion from § 5.4). A possible method for improvements would have to properly deal with overshoot probabilities, which essentially concern the last increment when a stopping-time occurs (in our case a_{N_+} and a_{N_-}, for which we used the immediate bounds $a_{N_+} \geq C$ and $a_{N_-} \geq 0$); see, e.g., Asmussen [2].

To further improve the bounds one could consider alternative continuous-time models. In this case, the second inequality above would be strengthen to

$$A(N_-) - N_- C \geq \sigma - K - \varepsilon \, ,$$

for infinitesimally small ε (recall the derivation of (17) in discrete-time). Such a continuous-time extension of Kingman's technique was considered by Palmowski and Rolski [43]; see [44] for related comments concerning continuous vs discrete-time models.

Our results could be extended to a slight variation of the G/G/1/K queue subject to the recursion

$$Q(n+1) = \max\{0, \min\{Q(n) + X_n, K\} - T_n\} \, .$$

T_n's are the jobs' interarrival times, whereas X_n's measure the job sizes (e.g., bits) to be served at rate 1. Unlike the standard G/G/1/K model whereby $Q(n)$ measures the number of jobs in the queue, in the modified model $Q(n)$ measures 'bits', or, equivalently, waiting times. Moreover, no arrivals are fully dropped, but only the overflowing fraction of X_n. The advantage of this fractional queueing model is that it would lend itself to a non-recursive representation similar to (4), which could be solved by adapting our stopping-times/martingale technique. An alternative challenge is to directly express the standard G/G/1/K queue size in a non-recursive manner and apply our technique.

9 CONCLUSIONS

In this paper we have analyzed finite-buffer queues with iid and Markovian arrivals. Using a non-elementary extension of Kingman's bounding approach for GI/G/1 queues, we have obtained bounds on the queue and loss distributions. The former retain the inherent truncated behavior characteristic to finite-buffer queues, thus departing from classical exponential tail approximations. In the iid case and at utilization $\rho = 1$, the upper bounds are insensitive to the arrivals distribution, whereas in heavy-traffic the bounds are numerically accurate and improve upon existing bounds by orders

of magnitude. A fundamental challenge is to extend our results in realistic feedback-based/closed-loop scenarios whereby the arrival model reacts to losses.

REFERENCES

[1] Vyacheslav M. Abramov. 1997. On a Property of a Refusals Stream. *Journal of Applied Probability* 34, 3 (March 1997), 800–805. https://doi.org/10.2307/3215106

[2] Søren Asmussen. 2000. *Ruin Probabilities*. World Scientific.

[3] François Baccelli and Pierre Brémaud. 2002. *Elements of Queueing Theory. Palm Martingale Calculus and Stochastic Recurrences*. Springer.

[4] Andrea Baiocchi, Nicola B. Melazzi, Marco Listanti, Aldo Roveri, and Roberto Winkler. 1991. Loss Performance Analysis of an ATM multiplexer Loaded with High-Speed ON-OFF Sources. *IEEE Journal on Selected Areas in Communications* 9, 3 (April 1991), 388–393. https://doi.org/10.1109/49.76637

[5] Alamin A. Belhaj and László Pap. 2000. An Efficient Bandwidth Assignment Algorithm for Real-Time Traffic in ATM Networks. In *Performance Analysis of ATM Networks: IFIP TC6 WG6.3 / WG6.4. Fifth International Workshop on Performance Modeling and Evaluation of ATM Networks*. 339–357. https://doi.org/10.1007/978-0-387-35353-1_17

[6] Chatschik Bisdikian, John S. Lew, and Asser N. Tantawi. 1992. On the Tail Approximation of the Blocking Probability of Single Server Queues with Finite Buffer Capacity. In *Proc. of the Second International Conference on Queueing Networks with Finite Capacity*. 267–280.

[7] Chris Blondia. 1989. The N/G/1 Finite Capacity Queue. *Communications in Statistics. Stochastic Models* 5, 2 (1989), 273–294. https://doi.org/10.1080/15326348908807110

[8] Chirs Blondia and Olga Casals. 1991. Cell Loss Probabilities in a Statistical Multiplexer in an ATM Network. In *Proc. of 6th GI/ITG-Fachtagung, Messung, Modellierung and Bewertung von Rechensystemen*. 121–136.

[9] Chris Blondia and Olga Casals. 1992. Statistical Multiplexing of VBR sources: A Matrix-Analytic Approach. *Performance Evaluation* 16, 1 (Nov. 1992), 5 – 20. https://doi.org/10.1016/0166-5316(92)90064-N

[10] Jean-Chrysotome Bolot. 1993. End-to-end Packet Delay and Loss Behavior in the Internet. In *ACM SIGCOMM*. 289–298. https://doi.org/10.1145/166237.166265

[11] A. M. Bratiichuk. 2007. Rate of Convergence to Ergodic Distribution for Queue Length in Systems of the Type M^{θ}/G/1/N. *Ukrainian Mathematical Journal* 59, 9 (01 Sept. 2007), 1300–1312.

[12] Lawrence D. Brown, Tianwen Tony Cai, and Anirban DasGupta. 2001. Interval Estimation for a Binomial Proportion. *Statist. Sci.* 16, 2 (2001), 101–117. http://www.jstor.org/stable/2676784

[13] Emmanuel Buffet and Nick G. Duffield. 1994. Exponential Upper Bounds via Martingales for Multiplexers with Markovian Arrivals. *Journal of Applied Probability* 31, 4 (Dec. 1994), 1049–1060.

[14] Cheng-Shang Chang and Jay Cheng. 1995. Computable Exponential Bounds for Intree Networks with Routing. In *Proc. of IEEE Infocom*. 197–204.

[15] Mohan L. Chaudhry, Umesh C. Gupta, and Manju Agarwal. 1991. On Exact Computational Analysis of Distributions of Numbers in Systems for M/G/1/N + 1 and GI/M/1/N + 1 Queues using Roots. *Computers & Operations Research* 18, 8 (1991), 679 – 694. https://doi.org/10.1016/0305-0548(91)90006-D

[16] Israel Cidon, Asad Khamisy, and Moshe Sidi. 1993. Analysis of Packet Loss Processes in High-Speed Networks. *IEEE Transactions on Information Theory* 39, 1 (Jan. 1993), 98–108. https://doi.org/10.1109/18.179347

[17] Florin Ciucu and Felix Poloczek. 2018. Two Extensions of Kingman's GI/G/1 Bound. *Proc. of the ACM on Measurement and Analysis of Computing Systems - SIGMETRICS* 2, 3 (Dec. 2018), 43:1–43:33.

[18] Rene L. Cruz and Haining N. Liu. 1993. Single Server Queues with Loss: A Formulation. In *Proc. of the 1993 Conference on Information Sciences and Systems (CISS)*.

[19] Nick G. Duffield. 1994. Exponential Bounds for Queues with Markovian Arrivals. *Queueing Systems* 17, 3-4 (Sept. 1994), 413–430.

[20] Josep M. Ferrandiz and Aurel A. Lazar. 1992. Monitoring the Packet Gap of Real-Time Packet Traffic. *Queueing Systems* 12, 3 (Sept. 1992), 231–242. https://doi.org/10.1007/BF01158800

[21] Dieter Fiems, Stijn De Vuyst, Sabine Wittevrongel, and Herwig Bruneel. 2008. Packet Loss Characteristics for M/G/1/N Queueing Systems. *Annals of Operations Research* 170, 1 (Sept. 2008), 149–154. https://doi.org/10.1007/s10479-008-0436-9

[22] Yashar Ghiassi-Farrokhfal and Florin Ciucu. 2012. On the Impact of Finite Buffers on Per-Flow Delays in FIFO Queues. In *24th International Teletraffic Congress (ITC)*.

[23] Frank N. Gouweleeuw and Henk C. Tijms. 1998. Computing Loss Probabilities in Discrete-Time Queues. *Operations Research* 46, 1 (Jan.-Feb. 1998), 149–154. http://www.jstor.org/stable/223070

[24] Umesh C. Gupta and T.S.S. Srinivasa Rao. 1996. Computing Steady State Probabilities in $\lambda(n)$/G/1/K Queue. *Performance Evaluation* 24, 4 (1996), 265 – 275. https://doi.org/10.1016/0166-5316(94)00035-2

[25] Omer Gurewitz, Moshe Sidi, and Israel Cidon. 2000. The Ballot Theorem Strikes Again: Packet Loss Process Distribution. *IEEE Transactions on Information Theory* 46, 7 (Nov. 2000), 2588–2595. https://doi.org/10.1109/18.887866

[26] Mark Handley. 1997. *An Examination of MBONE Performance*. Technical Report. University of Southern California / Information Sciences Institute, ISI/RR-97-450.

[27] Fumio Ishizaki and Tetsuya Takine. 1999. Loss Probability in a Finite Discrete-Time Queue in Terms of the Steady State Distribution of an Infinite Queue. *Queueing Systems* 31, 3 (July 1999), 317–326. https://doi.org/10.1023/A:1019170500574

[28] Predrag R. Jelenković and Petar Momčilović. 2003. Asymptotic Loss Probability in a Finite Buffer Fluid Queue with Heterogeneous Heavy-Tailed On-Off Processes. *The Annals of Applied Probability* 13, 2 (May 2003), 576–603. http://www.jstor.org/stable/1193160

[29] Wenyu Jiang and Henning Schulzrinne. 2000. Modeling of Packet Loss and Delay and their Effect on Real-Time Multimedia Service Quality. In *Proc. of NOSSDAV*.

[30] Julian Keilson. 1966. The Ergodic Queue Length Distribution for Queueing Systems with Finite Capacity. *Journal of the Royal Statistical Society. Series B* 28, 1 (1966), 190–201.

[31] Wojciech M. Kempa. 2017. A comprehensive Study on the Queue-Size Distribution in a Finite-Buffer System with a General Independent Input Flow. *Performance Evaluation* 108 (Feb. 2017), 1 – 15. https://doi.org/10.1016/j.peva.2016.11.002

[32] Han S. Kim and Ness B. Shroff. 2001. Loss Probability Calculations and Asymptotic Analysis for Finite Buffer Multiplexers. *IEEE/ACM Transactions on Networking* 9 (Dec. 2001), 755–768. Issue 6. https://doi.org/10.1109/90.974529

[33] Nam K. Kim and Kyung C. Chae. 2003. Transform-Free Analysis of the GI/G/1/K Queue Through the Decomposed Little's Formula. *Computers & Operations Research* 30, 3 (March 2003), 353 – 365. https://doi.org/10.1016/S0305-0548(01)00101-0

[34] John F. C. Kingman. 1964. A Martingale Inequality in the Theory of Queues. *Cambridge Philosophical Society* 60, 2 (April 1964), 359–361.

[35] Rajeev S. Koodli and Rayadurgam Ravikanth. 2002. One-Way Loss Pattern Sample Metrics. IETF RFC 3357.

[36] San-Qi Li. 1989. Study of Information Loss in Packet Voice Systems. *IEEE Transactions on Communications* 37, 11 (Nov. 1989), 1192–1202. https://doi.org/10.1109/26.46513

[37] Haining Liu. 1993. *Buffer Size and Packet Loss in Tandem Queueing Network*. Ph.D. Dissertation. University of California at San Diego, San Diego, CA.

[38] David M. Lucantoni. 1993. The BMAP/G/1 Queue: A Tutorial. In *Performance Evaluation of Computer and Communication Systems*, Lorenzo Donatiello and Randolph Nelson (Eds.). Springer, 330–358.

[39] Josée Mignault, Annie Gravey, and Catherine Rosenberg. 1996. A Survey of Straightforward Statistical Multiplexing Models for ATM Networks. *Telecommunication Systems* 5, 1 (March 1996), 177–208. https://doi.org/10.1007/BF02109733

[40] Masakiyo Miyazawa. 1987. A Generalized Pollaczek-Khinchine Formula for the GI/GI/L/K Queue and its Application to Approximation. *Communications in Statistics. Stochastic Models* 3, 1 (Jan. 1987), 53–65. https://doi.org/10.1080/15326348708807046

[41] Ramesh Nagarajan, Jim F. Kurose, and Don F. Towsley. 1991. Approximation Techniques for Computing Packet Loss in Finite-Buffered Voice Multiplexers. *IEEE Journal on Selected Areas in Communications* 9, 3 (April 1991), 368–377. https://doi.org/10.1109/49.76635

[42] Asok K. Nanda, Harshinder Singh, Neeraj Misra, and Prasanta Paul. 2003. Reliability Properties of Reversed Residual Lifetime. *Communications in Statistics - Theory and Methods* 32, 10 (2003), 2031–2042. https://doi.org/10.1081/STA-120023264

[43] Zbigniew Palmowski and Tomasz Rolski. 1996. A Note on Martingale Inequalities for Fluid Models. *Statistics & Probability Letters* 31, 1 (Dec. 1996), 13–21.

[44] Felix Poloczek and Florin Ciucu. 2014. Scheduling Analysis with Martingales. *Performance Evaluation (Special Issue: IFIP Performance 2014)* 79 (Sept. 2014), 56 – 72.

[45] Majid Raeis, Almut Burchard, and Jörg Liebeherr. 2017. Analysis of the Leakage Queue: A Queueing Model for Energy Storage Systems with Self-Discharge. *CoRR* abs/1710.09506 (2017). arXiv:1710.09506 http://arxiv.org/abs/1710.09506

[46] Vaidyanthan Ramaswami. 1988. Traffic Performance Modeling for Packet Communication Whence, Where and Whiter (Keynote Address). In *Proc. of 3rd Australian Teletraffic Research Seminar*.

[47] Rhonda Righter. 1999. A Note on Losses in M/GI/1/n Queues. *Journal of Applied Probability* 36, 4 (Dec. 1999), 1240–1243. https://doi.org/10.1239/jap/1032374770

[48] Sheldon M. Ross. 1974. Bounds on the Delay Distribution in GI/G/1 queues. *Journal of Applied Probability* 11, 2 (June 1974), 417–421.

[49] Henning Schulzrinne, Jim F. Kurose, and Don F. Towsley. 1992. Loss Correlation for Queues with Bursty Input Streams. In *IEEE International Conference on Communications (ICC)*. 219–224.

[50] Bruno Sericola. 2001. A Finite Buffer Fluid Queue Driven by a Markovian Queue. *Queueing Systems* 38, 2 (June 2001), 213–220. https://doi.org/10.1023/A:1010962516045

[51] Moshe Shaked and J. George Shanthikumar. 2007. *Stochastic Orders*. Springer.

[52] Tetsuya Takine, Tatsuya Suda, and Toshiharu Hasegawa. 1995. Cell Loss and Output Process Analyses of a Finite-Buffer Discrete-Time ATM Queueing System with Correlated Arrivals. *IEEE Transactions on Communications* 43, 2/3/4 (Feb.

1995), 1022–1037. https://doi.org/10.1109/26.380134
[53] Peter G. Taylor. 2011. Insensitivity in Stochastic Models. In *R. Bourcherie and N. van Dijk (Eds.), Queueing Networks: A Fundamental Approach*. Springer, 121–140.
[54] Roger C. F. Tucker. 1988. Accurate Method for Analysis of a Packet-Speech Multiplexer with Limited Delay. *IEEE Transactions on Communications* 36, 4 (April 1988), 479–483. https://doi.org/10.1109/26.2773
[55] Ward Whitt. 2004. A Diffusion Approximation for the G/GI/n/m Queue. *Operations Research* 52, 6 (Nov-Dec 2004), 922–941.
[56] David Williams. 1991. *Probability with Martingales*. Cambridge University Press.
[57] Maya Yajnik, Jim F. Kurose, and Don F. Towsley. 1996. Packet Loss Correlation in the MBone Multicast Network. In *IEEE Global Telecommunications Conference (GLOBECOM)*. 94–99. https://doi.org/10.1109/GLOCOM.1996.586133
[58] Tao Yang and Danny H. K. Tsang. 1995. A Novel Approach to Estimating the Cell Loss Probability in an ATM Multiplexer Loaded with Homogeneous On-Off Sources. *IEEE Transactions on Communications* 43, 1 (Jan. 1995), 117–126. https://doi.org/10.1109/26.385936

A ADDITIONAL PROOFS

Proof of Corollary 9. The proof is similar to that of Theorem 3 except for the evaluations of $\mathbb{E}\left[M_{N_+}1_{N_+<N_-}\right]$ and $\mathbb{E}\left[M_{N_-}1_{N_-<N_+}\right]$. Expanding

$$\mathbb{E}\left[M_{N_+}1_{N_+<N_-}\right] = \sum_{k\geq 1}\mathbb{E}\left[M_k 1_{N_+=k}1_{N_->k}\right] ,$$

and denoting the partial sums $S_k := U_1 + \cdots + U_k$, $U_k := a_k - C$, and the density of U_1 by $f(x)$, we can further expand each term as

$$\begin{aligned}\mathbb{E}\left[M_k 1_{N_+=k}1_{N_->k}\right] &= \mathbb{E}\big[e^{\theta S_k}1_{U_1<\sigma}1_{U_1\geq\sigma-K}\cdots 1_{S_{k-1}<\sigma} \\ &\qquad 1_{S_{k-1}\geq\sigma-K}1_{S_k\geq\sigma}1_{S_k\geq\sigma-K}\big] \\ &= \int_{\sigma-K}^{\sigma} e^{\theta x_1}f(x_1)\cdots\int_{\sigma-K-s_{k-2}}^{\sigma-s_{k-2}} e^{\theta x_{k-1}}f(x_{k-1})E\left[e^{\theta U_k}1_{U_k\geq\sigma-s_{k-1}}\right] \\ &\qquad dx_{k-1}\ldots dx_1 ,\end{aligned}$$

where $s_k := x_1 + \cdots + x_k$. Rewriting the inner expectation as

$$E\left[e^{\theta(U_k-(\sigma-s_{k-1}))} \mid U_k \geq \sigma - s_{k-1}\right]\mathbb{P}(U_k \geq \sigma - s_{k-1})$$

and simplifying terms we obtain

$$\mathbb{E}\left[M_k 1_{N_+=k}1_{N_->k}\right] \geq \alpha_- e^{\theta\sigma}\mathbb{P}(N_+ = k, N_- > k) ,$$

and hence

$$\mathbb{E}\left[M_{N_+}1_{N_+<N_-}\right] \geq \alpha_- e^{\theta\sigma}\mathbb{P}(N_+ < N_-) .$$

Proceeding similarly for $\mathbb{E}\left[M_{N_-}1_{N_-<N_+}\right]$ we can write

$$\begin{aligned}\mathbb{E}\left[M_k 1_{N_-=k}1_{N_+>k}\right] &= \mathbb{E}\big[e^{\theta S_k}1_{U_1<\sigma}1_{U_1\geq\sigma-K}\cdots 1_{S_{k-1}<\sigma} \\ &\qquad 1_{S_{k-1}\geq\sigma-K}1_{S_k<\sigma}1_{S_k<\sigma-K}\big] \\ &= \int_{\sigma-K}^{\sigma} e^{\theta x_1}f(x_1)\cdots\int_{\sigma-K-s_{k-2}}^{\sigma-s_{k-2}} e^{\theta x_{k-1}}f(x_{k-1})E\left[e^{\theta U_k}1_{U_k<\sigma-K-s_{k-1}}\right] \\ &\qquad dx_{k-1}\ldots dx_1 .\end{aligned}$$

Bounding as above and simplifying terms yields

$$\mathbb{E}\left[M_{N_-}1_{N_-<N_+}\right] \geq \beta_- e^{\theta(\sigma-K)}\mathbb{P}(N_- < N_+) ,$$

and the proof for the upper bound is complete. The proof for the lower bound proceeds similarly except for reversing the inequalities above and replacing the 'inf' by 'sup'.

Lastly, we note that if a_1 does not have a density (e.g., it is a discrete random variable) then the proof can be slightly adapted by replacing the integrals by sums and; nonetheless, the results from Corollary 8 hold as stated. □

Proof of Lemma 10. The proof proceeds similarly as the proof of Lemma 7 (see [44]); the main difference is the sign change. Let the functions $\phi_1(\theta) = \mathbb{E}\left[e^{-\theta a_1}\right]$ and $\phi_2(\theta) = e^{-\theta C}$ for $\theta \geq 0$. Because

$$\phi_1'(\theta)|_{\theta=0} = -E[a_1] < -C = \phi_2'(\theta)|_{\theta=0}$$

and $\phi_1(0) = \phi_2(0)$, it follows that there exists $\varepsilon > 0$ such that

$$\phi_1(\theta) < \phi_2(\theta)$$

in $(0, \varepsilon)$. Moreover, because $\inf a_1 < C$, there exists θ' such that $\phi_1(\theta') \geq \phi_2(\theta')$; (if a_1 is a discrete r.v. then $\theta' = -\frac{\log\mathbb{P}(a_1=\inf a_1)}{C-\inf a_1}$). Therefore, θ is well-defined and the rest of the proof proceeds as in Lemma 7. □

B ADDITIONAL CASE-STUDIES

B.1 Autoregressive Arrival Processes

We consider *autoregressive* (AR) processes which belong to a subclass of Markovian arrivals whose instantaneous process $(a_k)_{k\in\mathbb{Z}}$ is defined recursively as follows

$$a_n = \varphi a_{n-1} + (1-\varphi)\mu + (1-\varphi)\sigma Z_n , \tag{33}$$

where $\varphi \in (0,1)$, $\mu, \sigma > 0$, and $(Z_k)_{k\in\mathbb{Z}}$ is an independent family of $\mathcal{N}_{0,1}$-distributed random variables.

We use the following result from [44] for the construction of the martingale-envelope.

Lemma 16. *Let $\theta = 2\frac{C-\mu}{\sigma^2}$ and $h(x) = e^{\theta\frac{\varphi}{1-\varphi}x}$. Then the autoregressive flow A admits a martingale-envelope with parameters θ and h.*

Next we give (*approximate*) bounds on the queue distribution; bounds on the loss distribution can be obtained in a similar manner. The lack of rigorousness is due to the fact that the increments a_n can potentially be negative, as they follow a normal distribution, which contradicts our basic assumption of positive arrivals. In general, however, depending on the AR model, negative arrivals can occur with negligible probabilities only.

Corollary 17 (Queue distribution - AR Case). *In the scenario above, the queue size distribution Q satisfies*

$$\mathbb{P}(Q \geq \sigma) \lessapprox \frac{e^{\theta\frac{\varphi}{1-\varphi^2}(\varphi C+\mu)} - e^{\theta(\sigma-K-C)}}{e^{\theta\frac{\varphi}{1-\varphi}C}e^{\theta\sigma} - e^{\theta(\sigma-K-C)}} . \tag{34}$$

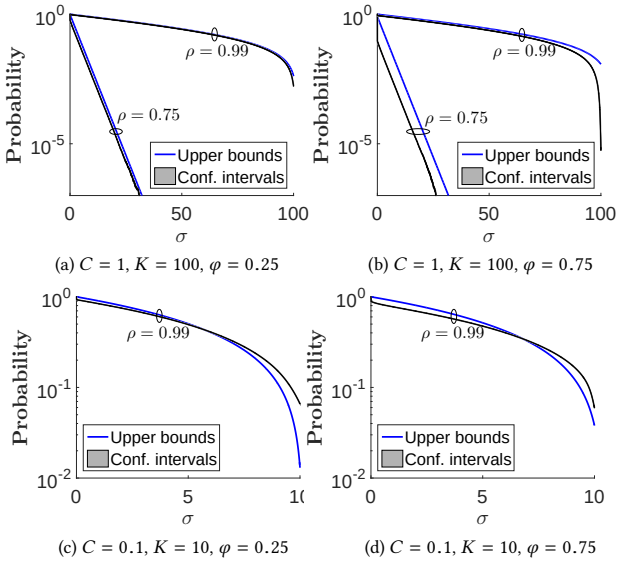
While AR processes have a Markovian structure, the key observation is that the parameter θ has now an explicit expression.

Proof. First we compute the mean and variance of a_n. By stationarity,

$$\begin{aligned}E[a_n] &= E[\varphi a_{n-1} + (1-\varphi)\mu + (1-\varphi)\sigma Z_n] \\ &= \varphi\mathbb{E}[a_{n-1}] + (1-\varphi)\mu + 0 \\ &= \varphi\mathbb{E}[a_n] + (1-\varphi)\mu ,\end{aligned}$$

and hence $E[a_n] = \mu$. Similarly,

$$\begin{aligned}\mathrm{Var}[a_n] &= \mathrm{Var}[\varphi a_{n-1} + (1-\varphi)\mu + (1-\varphi)\sigma Z_n] \\ &= \varphi^2\mathrm{Var}[a_{n-1}] + (1-\varphi)^2\sigma^2 \\ &= \varphi^2\mathrm{Var}[a_n] + (1-\varphi)^2\sigma^2 ,\end{aligned}$$



(a) $C = 1$, $K = 100$, $\varphi = 0.25$ (b) $C = 1$, $K = 100$, $\varphi = 0.75$

(c) $C = 0.1$, $K = 10$, $\varphi = 0.25$ (d) $C = 0.1$, $K = 10$, $\varphi = 0.75$

Figure 8: Queue distribution for AR processes

and hence

$$\mathrm{Var}[a_n] = \sigma^2 \frac{(1-\varphi)^2}{1-\varphi^2} = \sigma^2 \frac{1-\varphi}{1+\varphi} .$$

Therefore, a_n is normally distributed with mean μ and variance $\sigma^2 \frac{1-\varphi}{1+\varphi}$. The expectation $\mathbb{E}[h(a_0)]$ can now be written as

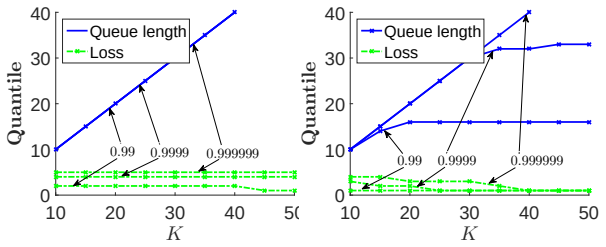
$$\begin{aligned}\mathbb{E}[h(a_0)] =& \mathbb{E}\left[e^{\theta \frac{\varphi}{1-\varphi} a_1}\right] = \mathbb{E}\left[e^{\theta \frac{\varphi}{1-\varphi}\left(\mu+\sigma\sqrt{\frac{1-\varphi}{1+\varphi}} Z_1\right)}\right] \\ =& e^{\theta \frac{\varphi}{1-\varphi}\mu} e^{\frac{\theta^2\varphi^2}{2(1-\varphi)^2}\sigma^2 \frac{1-\varphi}{1+\varphi}} = e^{\theta \frac{\varphi}{1-\varphi}\left(\mu+\frac{\theta\varphi}{2}\sigma^2 \frac{1}{1+\varphi}\right)} \\ =& e^{\theta \frac{\varphi}{1-\varphi}\left(\mu+\frac{(C-\mu)\varphi}{1+\varphi}\right)} = e^{\theta \frac{\varphi}{1-\varphi^2}(\varphi C+\mu)}\end{aligned}$$

For the computation of H_+ and H_- we observe from Def. 2 that

$$H_+ = h(C) = e^{\theta \frac{\varphi}{1-\varphi} C} \quad \text{and} \quad H_- = h(0) = 1 .$$

The rest follows by Theorem 3. □

In Fig. 8 we compare the approximate AR bounds against simulations, for two values of the weighting parameter φ (larger values of φ correspond to stronger correlation structures of the increments a_n). We let $\sigma = 1$, whereas μ is implicitly computed from the utilization ρ and C. Despite not being rigorous, the bounds are still accurate at high utilization in (a) and (b), which indicates a negligible effect of the negative increments. By scaling the (arrival) units and reducing the mean by a factor of 10 (in (c) and (d)), the adverse effects of the negative increments arise. We also note that the approximation errors in (c) and (d) are not uniform in σ (i.e., the 'upper bounds' can be either below or above simulations). The reason is that the 'true' result would include a prefactor L for the term $e^{\theta(\sigma-K-C)}$ in (34), both in the numerator and denominator; see the derivation of (17) from Theorem 3.



(a) $\rho = 0.99$ ($C \approx 0.84$) (b) $\rho = 0.75$ ($C \approx 1.11$)

Figure 9: Queue length and loss quantiles for MMOO's ($N = 5$, $p = 0.1$, $q = 0.5$, $R = 1$)

B.2 Queue-Size vs. Loss

Here we use our main results to study the relationship between queue-size and loss in the context of dimensioning buffer sizes subject to Quality-of-Service constraints of the form

$$\mathbb{P}\left(Q \geq \sigma_Q\right) \leq \varepsilon \text{ and } \mathbb{P}\left(L \geq \sigma_L\right) \leq \varepsilon ,$$

for some *target values* σ_Q, σ_L, and ε. Intuitively, larger buffer sizes K imply larger queues (and hence waiting-times/delays) and fewer losses; in turn, smaller values of K imply smaller queues (delays) but more losses. Depending on the parameters σ_Q, σ_L, and σ, an 'optimal' buffer size may not exist.

In Fig. 9 we show the queue and loss quantiles for a range of buffer sizes K, while keeping the utilization ρ constant; for instance, the values corresponding to the 0.99 quantiles translate into a value $\varepsilon = 10^{-2}$. At high utilization ($\rho = 0.99$), the key insight is that buffers should be small, as otherwise delays increase sharply while losses only reduce negligibly; however, depending on the application (e.g., involving forward error correction schemes) small gains in the loss can have a significant impact on the application's performance, and hence larger delays may be more desirable. At smaller utilization ($\rho = 0.75$), the underlying delays vs. losses tradeoff becomes more subtle, as it additionally depends on the 'confidence level' ε.