Age of information without service preemption

George Kesidis[*] Takis Konstantopoulos[†] Michael A. Zazanis[‡]

15 April 2021

Abstract

When designing a message transmission system from the point of view of making sure that the information transmitted is as fresh as possible, two rules of thumb seem reasonable: use small buffers and adopt a last-in-first-out policy. In this paper, we measure freshness of information using the recently adopted "age of information" performance measure. Considering it as a stochastic process operating in a stationary regime, we compute not just the first moment but the whole marginal distribution of the age of information (something important in applications) for two well-performing systems. In neither case do we allow for preemption of the message being processed because this may be difficult to implement in practice. We assume that the arrival process is Poisson and that the messages have independent sizes (service times) with common distribution. We use Palm and Markov-renewal theory to derive explicit results for Laplace transforms which, in many cases can be inverted analytically. We discuss how well the systems we analyze perform and examine how close to optimality they are. In particular, we answer an open question that was raised in [9] regarding the optimality of the system denoted as $\mathcal{P}_2$.

Keywords and phrases. Age of information, Markov renewal, Palm probability, Laplace transform, queueing

AMS 2020 *subject classification.* 60K30, 60E10, 68M20, 60K25, 60K15

1 Introduction

Traditionally, networked systems performance is measured with respect to buffer sizes and delays. Relatively recently, there has been a shift in what is considered important both in terms of design and performance. The introduction of the so-called "age of information" (usually abbreviated as AoI), defined as the elapsed time since the information possessed by a monitor is generated and time stamped at the source, has received a lot of attention. The reason is simple: in several applications it is the freshness of information that is important rather than the correct transmission of all packets. Examples include virtual reality, online gaming, weather reports, autonomous driving, stock market trading, decision systems for an airplane, power systems, sensor/actuator systems and other other "cyber physical" systems.

We start by precisely defining the concept of AoI in general. Consider a message processing facility with one input stream of arriving messages. The facility can be a single queue or a complex network system. An arriving message has a certain positive "size" (expressed in time units and interpreted as processing or service time) and three things can happen: (i) the message is immediately rejected upon arrival; (ii) the message is accepted but rejected while in the system; (iii) the message is successfully transmitted as soon as it is processed in its entirety. We are interested in the time that the latter will happen in comparison to the time that the message arrives in the system. If messages are labeled by

[*] gik2@psu.edu; Univ. of Pennsylvania, USA

[†] takiskonst@gmail.com; Univ. of Liverpool, UK; research supported by Cast. Co. IIS-75

[‡] zazanis@aueb.gr; Athens Univ. of Economics and Business, Greece

integers in a way that the message with label $n \in \mathbb{Z}$ arrives at time $T_n \in \mathbb{R}$ and $T_m < T_n$ if $m < n$, if $T_n + \Delta_n$ denotes the time at which message labeled n leaves the system either by being rejected or successfully transmitted, and if ψ_n is a binary variable indicating the latter ($\psi_n = 1$ if message n is successful or 0 if not), then we let

$$D(t) := \sup\{T_n + \Delta_n : n \in \mathbb{Z}, T_n + \Delta_n \le t, \psi_n = 1\}, \tag{1}$$

$$A(t) := \sup\{T_n : n \in \mathbb{Z}, T_n + \Delta_n \le t, \psi_n = 1\}, \tag{2}$$

and define the AoI at time t by

$$\alpha(t) := t - A(D(t)). \tag{3}$$

Quite simply, $D(t)$ is the time of the last successful departure before t and $A(D(t))$ is the arrival time of the message that departed at time $D(t)$. This definition is quite general, that is, it does not depend on the details of the system design.

Typically, systems that adopt freshness of information as performance measure should be designed so that its AoI "be as small as possible". The last term can mean several things. For example, it can mean that the quantity $\alpha(t)$ is least for all t under identical traffic conditions. Or it could mean least in terms of an expectation or another functional of the process. Adopting AoI as a performance criterion immediately poses some simplifications over traditional queueing theory performance criteria but also presents some new challenges.

Suppose that the processing facility is a queue with one server and buffer consisting of a finite or infinite number of cells. One of the cells of the buffer is occupied by the message being processed and the rest of the cells are occupied by stored messages. If we are free to design the buffer and the service policy as we wish, how do we do this if we wish to keep the AoI "as small as possible"?

It is reasonable to conjecture that every time a message arrives we process it immediately (after all, we are not interested in obsolete information.) That is, even if the server is busy at the moment of arrival, the currently served message is immediately discarded and the new one starts being processed. Systems working in this manner are service-preemptive. It also seems reasonable to serve messages in reverse order of arrival: the most recent message must be served first (LIFO).

One may thus conjecture that LIFO-preemptive (meaning: service preemptive) is "best". But numerical examples and simulations show that this is false depending on the model assumptions. In particular, we showed in [9] that a single buffer system with no service preemption (called $\mathcal{B}_1$ below) has smaller AoI both in expectation and stochastically under particular assumptions on the message size distribution. In fact, we conjectured in [9] that the so-called $\mathcal{P}_2$ system (see below and Figure 1 for the definition) has lower AoI than $\mathcal{B}_1$. In this paper, we resolve this issue by analyzing under Poisson arrivals and generally distributed message sizes.

Note that in some cases, service preemption may not be technologically feasible. The simplest, small-size buffer, systems with no service preemption are defined next. One of them, denoted as $\mathcal{B}_2$, is nothing else but a single-server queue with buffer of size 2 and blocking. That is, an incoming message finding the buffer full is immediately discarded. The other one, denoted as $\mathcal{P}_2$, also works without service preemption. An arriving message in $\mathcal{P}_2$ finding the buffer full displaces or "pushes out" the stored message. See Figure 1 for a typical scenario in $\mathcal{P}_2$.

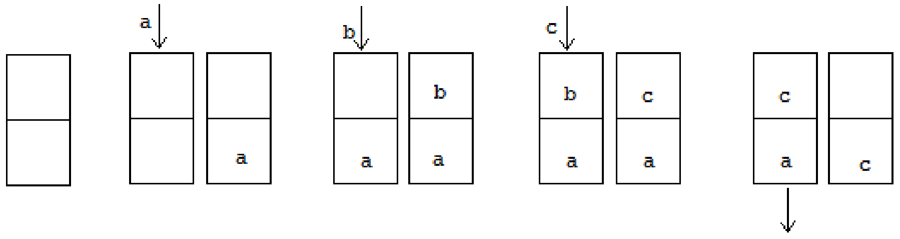

Figure 1: *The $\mathcal{P}_2$ system. The lower cell contains the message being processed (if any). Message a arrives at an empty system and is immediately processed. Message b arrives and is stored in cell the upper cell. While the buffer still contains a and b, a third message c arrives and immediately kicks b out. When a completes being processed it departs and c moves to the lower cell.*

More generally, we define $\mathcal{B}_n$ and $\mathcal{P}_n$ as follows. The $\mathcal{B}_n$ system is simply a single-server queue with buffer of size n, operating under FIFO (First In First Out) policy and blocking of any incoming message finding the buffer full. The $\mathcal{P}_n$ system, for $n \geq 2$, works as follows: messages are stored in an order that is reverse to their order of arrival; so if there is a message being processed in cell 1 at time t, the message in cell 2 arrived last before t while the message in cell n is the oldest; a new message arriving at a full buffer is always stored in cell 2, displacing the other messages upwards and pushing out the one sitting in cell 1 (oldest one). For $n \geq 2$, $\mathcal{P}_n$ has no service preemption. The $\mathcal{P}_1$ and $\mathcal{B}_1$ systems were analyzed under very general conditions in [9]. It is easy to see that, for if $\mathcal{P}_n$ and $\mathcal{B}_n$ have identical arrival processes and identical message sizes then

$$\alpha_{\mathcal{P}_n}(t) \leq \alpha_{\mathcal{B}_n}(t) \text{ for all } t, \text{ if } n \geq 2,$$

but, as observed in [10, 9], $\alpha_{\mathcal{P}_1}$ and $\alpha_{\mathcal{B}_1}$ are not comparable in this sense.

In this paper, we analyze the AoI processes and derive the stationary AoI distribution for the $\mathcal{P}_2$ and $\mathcal{B}_2$ cases under a Poisson arrival process but generally distributed message sizes. The technique used is different from that of [9] and is easily generalizable to $\mathcal{P}_n$ and $\mathcal{B}_n$ for arbitrary n. Indeed, we use the classical embedding technique, valid for queueing systems with Poisson arrivals (see, e.g. [4]), giving that the system sampled at certain epochs (the departures of successful messages in our case) has a Markovian property. This, of course, depends crucially on the Poisson property of the arrival process, so, for more general arrival processes (as in the case of [9]) we need a different method.

Under our assumptions, and because we consider finite buffers, it holds that there is a unique stationary version of the stochastic process α in all systems considered. We will not offer any reasons for this technical result, but only point out that even existence may not hold if the arrival and message size processes are neither independent nor renewal, and point out the difficulties by referring to [2]. We shall always be considering the stationary version. Hence $\alpha(t)$ has the same distribution for all t. This is precisely what we are interested in describing. We note that computing the expectation is, in general, not that much easier than deriving the whole distribution. We also note that deriving the distribution is essential in case that we are interested not just in maintaining a low AoI on the average but also in maintaining the tail of the probability distribution small.

Throughout the paper, we let λ be the rate of the (Poisson) arrival process and G the distribution of a typical message size σ, a random variable that is positive with probability 1 and has finite expectation denoted by $1/\mu$. We thus only assume that $\lambda > 0$ and $\mu > 0$ (but σ may have infinite variance). It is assumed that the message sizes are i.i.d. copies of σ and independent of the arrival process. The ratio $\rho = \lambda/\mu$ is referred to as traffic intensity and is occasionally used.

The main results of the paper are Theorems 1 and 2. We here summarize some special case. For the case where the message sizes are i.i.d. exponential, having (for notational convenience) rate $\mu = 1$, we shall show as a corollary that, in steady-state, the value of AoI at some (and hence any) point of time, has density

$$f_{\mathcal{P}_2}(t) = c\,(q(t)e^{-t} + e^{-\lambda t}), \tag{4}$$

$$f_{\mathcal{B}_2}(t) = q_1(t)e^{-t} + q_2(t)e^{-(\lambda+1)t} - \frac{\lambda}{\lambda - 1}e^{-\lambda t}, \tag{5}$$

in the $\mathcal{P}_2$ and $\mathcal{B}_2$ cases, respectively, where

$$c = \frac{\lambda}{(\lambda^2+\lambda+1)(\lambda-1)^2}, \quad q(t) = \tfrac{1}{2}\lambda(\lambda-1)^2t^2 + \lambda(\lambda-1)t - 1,$$
$$q_1(t) = \frac{(\lambda^3+\lambda^2-2\lambda)t - \lambda^2+\lambda+3}{(\lambda^2+\lambda+1)(\lambda-1)}, \quad q_2(t) = \frac{(\lambda^2+\lambda)t+\lambda^2+3\lambda+3}{\lambda^2+\lambda+1},$$

when $\lambda \neq 1$; while, for $\lambda = 1$, the densities become

$$f_{\mathcal{P}_2}(t) = \tfrac{1}{3}(t^2+t)e^{-t}, \quad f_{\mathcal{B}_2}(t) = \tfrac{1}{3}(7+2t)e^{-2t} + \tfrac{1}{3}(6t-7)e^{-t}.$$

(For general μ, simply replace λ by λ/μ and t by μt in the foregoing expressions.) In particular, as $\lambda \to \infty$, we have convergence to $\sigma_1+\sigma_2+\sigma_3$ in the $\mathcal{B}_2$ system and to $\sigma_1+\sigma_2$ in the $\mathcal{P}_2$ system, where the σ_i are independent exponential random variables with mean 1 each. A moment of reflection shows that this is expected because, e.g., in the $\mathcal{B}_2$ system and in high traffic intensity, a message immediately enters the system as soon as there is a free space in the buffer. In fact, this observation is easily generalizes to $\mathcal{P}_n$ and $\mathcal{B}_n$. If $\alpha_{\mathcal{P}_n}$, $\alpha_{\mathcal{B}_n}$ are two random variables distributed as the AoI in steady-state for the two systems respectively, then,

$$\text{as } \lambda \to \infty,\ \alpha_{\mathcal{P}_n} \to \sigma_1 + \cdots + \sigma_n \text{ and } \alpha_{\mathcal{B}_n} \to \sigma_1 + \cdots + \sigma_{n+1}, \tag{6}$$

in distribution. In some sense then, for relatively high λ, $\alpha_{\mathcal{P}_n}$ and $\alpha_{\mathcal{B}_n}$ increase with n which suggests that studying a small n is enough. In fact, we can intepret $\mathcal{B}_\infty$ as a single server queue with infinite buffer operating under the FIFO discipline. From the point of view of AoI, this system is worst and should not be considered. Our practical rule of thumb says that one should store at most one message and discard everything beyond that, insofar as keeping AoI low is the goal. This rule of thumb may not be always optimal but it is frequently close to optimal.

Regarding related prior work on AoI for stationary queueing systems: The expectation of AoI for $\mathcal{B}_\infty$, a system with generally large AoI, was derived in [6] only in the case where arrivals are Poisson and message sizes exponential. Further work on the expectation of various systems was done in [10], again under Poisson and exponential assumptions. Other work for LIFO queues has been done in [7, 3, 8]. We derived probability distributions for $\mathcal{P}_1$ and $\mathcal{B}_1$ for renewal arrivals and generally distributed message sizes in [9]. In general, besides [9], most work has focused on expectations and not on distributions. The reason one is interested in the latter is that latency-sensitive information requires controlling the probability that the freshness measure is large and to keep it low, rather than keeping the expectation small.

The rest of the paper is organized as follows. We explain the basis of the technique used in Section 2 and see why it is absolutely general, as long as the system, possibly a network, has Poisson arrivals. The $\mathcal{B}_2$ system is considered in Section 3 and the $\mathcal{P}_2$ in Section 4. Comparisons and numerical results are given in Section 5 and some discussions on extensions and use of the results follow in Section 6.

2 Basic framework

We discuss the technique used in the analysis for all systems described in the introduction from the point of view of the distribution of the age of information. By this phrase, we will always mean that the age of information process $\alpha(t)$, $t \in \mathbb{R}$, is stationary and that we shall be interested in the distribution of $\alpha(t)$ for some, and hence all, t which will be taken to be the point $t = 0$. The goal is to derive a "fixed point equation" for $\alpha(0)$, see equation (12) below. We note right away, that the present analysis is different than that of paper [9] as we take advantage of the following assumptions. The arrival process is always taken to be Poisson process on $\mathbb{R}$ (=time) with rate λ. As mentioned above, T_n is the arrival time of message labeled n. Its size is σ_n. The collection of message sizes are i.i.d. and independent of the arrival process. Let

$$G(x) = \mathbb{P}(\sigma_1 \le x)$$

be the distribution function of the typical size and let $1/\mu$ be its expectation, assumed to be finite. Also assume that $G(0) = 0$. Abusing notation, we shall let G denote the probability measure defined by the function $G(x)$ and by $\hat{G}(s)$ its Laplace transform:

$$\hat{G}(s) = \int_0^\infty e^{-sx} dG(x).$$

Recall that $T_n + \Delta_n$ is defined as the time at which message n departs either because it was pushed out or rejected or because it was successfully processed ($\psi_n = 1$ in the latter case). See discussion around (1) and (2) where these symbols were introduced. Then the number of messages in the system at time $t \in \mathbb{R}$ is given by

$$\xi(t) = \sum_{n \in \mathbb{Z}} \mathbf{1}_{T_n \le t < T_n + \Delta_n}.$$

Note that if the message is immediately rejected then $\Delta_n = 0$ and so this message does not contribute to ξ. We let

$$\{S_m, m \in \mathbb{Z}\} := \{T_n + \Delta_n : \ n \in \mathbb{Z}, \ \psi_n = 1\},$$

and, thinking of the two sets as sequences, $\{S_m\}$ is a subsequence of $\{T_n + \Delta_n\}$ and is enumerated so that $S_{m_1} < S_{m_2}$ if $m_1 < m_2$. We note that ξ is right-continuous for all t. Recalling the notions of Markov renewal and semi-Markov processes, see, e.g., [1, VII.4], our first observation is:

Lemma 1. *For both $\mathcal{B}_2$ and $\mathcal{P}_2$ cases, the process $\xi(t)$, $t \in \mathbb{R}$, is a semi-Markov process [4, Ch.10] with respect to the points S_n, $n \in \mathbb{Z}$. Moreover, the distribution of ξ is the same in both $\mathcal{B}_2$ and $\mathcal{P}_2$ cases.*

This follows easily by standard arguments in queueing theory, for instance in the analysis of a queue with Poisson arrivals; see, e.g., [4, Ch. 6, Sec. 5]. Thus, ξ does not "see" the difference between $\mathcal{B}_2$ and $\mathcal{P}_2$. The distinction between these two will become important in the next section when we discuss the details about α in each case.

We further assume that the arrival process together with the process ξ are stationary under a probability measure $\mathbb{P}$. (This assumption is non-vacuous; we shall not elaborate on this further but refer the reader to [2] for an exposition of techniques used to establish it.)

We refer to the intervals $[S_n, S_{n+1})$ as *segments* and split the paths of ξ into union of paths over segments. See Figure 2. By convention, we assume that the segment labelled 0 contains the point $t = 0$. Denote by $\mathbb{P}^0$ the Palm probability of $\mathbb{P}$ with respect to the point process $\{S_n\}$. We refer to [2] for this concept. Intuitively, $\mathbb{P}^0$ is $\mathbb{P}$ conditional on the event that $0 \in \{S_n, n \in \mathbb{Z}\}$. Hence $\mathbb{P}^0(S_0 = 0) = 1$. Let

$$K_n := \xi(S_n), \quad n \in \mathbb{Z}.$$

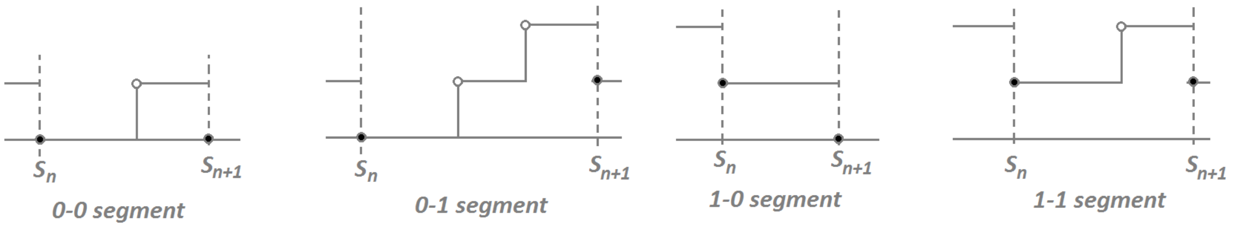

Figure 2: *What $\xi(t)$ looks like when $S_n \le t < S_{n+1}$, regardless of the policy used.*

The sequence $\{K_n\}$ is a Markov chain with state space $\{0,1\}$ while $\{(S_n,K_n)\}$ is the Markov renewal sequence [1, 4] associated to the semi-Markov process ξ. The latter has transition kernel

$$Q_{ij}(x) := \mathbb{P}^0(S_{n+1}-S_n \le x, K_{n+1}=j \,|\, K_n=i), \quad i,j\in\{0,1\},$$

explicitly given by

$$\begin{bmatrix} Q_{00}(x) & Q_{01}(x) \\ Q_{10}(x) & Q_{11}(x) \end{bmatrix} = \begin{bmatrix} \int_0^x \left(1-e^{-\lambda(x-u)}\right)e^{-\lambda u}dG(u) & \int_0^x \left(1-e^{-\lambda(x-u)}\right)\left(1-e^{-\lambda u}\right)dG(u) \\ \int_0^x e^{-\lambda u}dG(u) & \int_0^x \left(1-e^{-\lambda u}\right)dG(u) \end{bmatrix}, \tag{7}$$

as follows easily by considering the cases of Figure 2. Letting $x\to\infty$ in (7) we obtain the transition matrix for the Markov chain $\{K_n\}$,

$$\begin{bmatrix} Q_{00}(\infty) & Q_{01}(\infty) \\ Q_{10}(\infty) & Q_{11}(\infty) \end{bmatrix} = \begin{bmatrix} \hat{G}(\lambda) & 1-\hat{G}(\lambda) \\ \hat{G}(\lambda) & 1-\hat{G}(\lambda) \end{bmatrix},$$

from which it is evident that $K_n, n\ge 1$, is not just Markovian but also a sequence of independent Bernoulli random variables with

$$\mathbb{P}^0(K_n=0)=\hat{G}(\lambda)=1-\mathbb{P}^0(K_n=1). \tag{8}$$

Figure 2 shows the four different types of segments depending on the values of K_n and K_{n+1}. We next define

$$\Phi_i(s) := \mathbb{E}^0[e^{-s(S_1-S_0)}|K_0=i],$$

and, using the kernel (7), we obtain

$$\Phi_0(s) = \int_0^\infty e^{-sx}dQ_{00}(x) + \int_0^\infty e^{-sx}dQ_{01}(x) = \frac{\lambda}{\lambda+s}\hat{G}(s), \tag{9}$$

$$\Phi_1(s) = \int_0^\infty e^{-sx}dQ_{10}(x) + \int_0^\infty e^{-sx}dQ_{11}(x) = \hat{G}(s). \tag{10}$$

From (9), (10), and (8) we obtain the Laplace transform of the segment length:

$$\Phi(s) \;:=\; \mathbb{E}^0[e^{-s(S_1-S_0)}] \;=\; \left(1-\hat{G}(\lambda)+\hat{G}(\lambda)\frac{\lambda}{\lambda+s}\right)\hat{G}(s).$$

From this, we obtain the mean length of a segment as

$$\mathbb{E}^0[S_1-S_0] \;=\; \frac{1}{\mu}+\frac{\hat{G}(\lambda)}{\lambda}. \tag{11}$$

We shall henceforth use the abbreviation $\mathbb{E}(X;A)$ for the expectation of a random variable X on the event A, that is, the quantity $\mathbb{E}(X\mathbf{1}_A)$. The following result depends entirely on the semi-Markov property of ξ.

Proposition 1. *The random variable $\alpha(0)$ satisfies*

$$\mathbb{E}[e^{-s\alpha(0)}] \;=\; \frac{\lambda}{s}\cdot\frac{\mathbb{E}^0[e^{-s\alpha(0)};K_0=0]\left(1-\frac{\lambda}{\lambda+s}\hat{G}(s)\right)+\mathbb{E}^0[e^{-s\alpha(0)};K_0=1]\left(1-\hat{G}(s)\right)}{\frac{\lambda}{\mu}+\hat{G}(\lambda)}. \tag{12}$$

Proof. The Palm inversion formula [1, 2] applied to the $\mathbb{P}$-stationary process α gives

$$\mathbb{E}[e^{-s\alpha(0)}] = \frac{\mathbb{E}^0[\int_{S_0}^{S_1} e^{-s\alpha(t)}dt]}{\mathbb{E}^0[S_1 - S_0]}. \tag{13}$$

Take a look at (3) and notice that the process α is right-continuous. Its set of discontinuities is $\{S_n\}$. Moreover, it increases at unit rate on each segment:

$$\alpha(t) = \alpha(S_n) + t - S_n, \quad \text{for } t \in [S_n, S_{n+1}). \tag{14}$$

To see this, notice that, for $S_n \le t < S_{n+1}$, we have $D(t) = D(S_n) = S_n$, by the definition of D in (1), and so $A(D(t)) = A(D(S_n)) = A(S_n)$. Since, from the definition (3), $\alpha(t) = t - A(D(t))$ for all t, we have

$$\alpha(t) = t - A(S_n),$$

whenever $S_n \le t < S_{n+1}$. Writing this for $t = S_n$, we have

$$\alpha(S_n) = S_n - A(S_n),$$

and so (14) is obtained by subtracting the last two displays. In particular, $S_0 = 0$ and $\alpha(t) = \alpha(0) + t$ for $t \in [S_0, S_1)$, $\mathbb{P}^0$-a.s. Hence, for $i = 0, 1$,

$$\mathbb{E}^0\left[\int_{S_0}^{S_1} e^{-s\alpha(t)}dt; K_0 = i\right] = \mathbb{E}^0\left[e^{-s\alpha(0)}\,\frac{1 - e^{-sS_1}}{s}; K_0 = i\right] = s^{-1}\Phi_i(s)\mathbb{E}^0[e^{-s\alpha(0)}; K_0 = i], \tag{15}$$

where the last equality follows from the fact that $\alpha(0)$ and $S_1 - S_0$ are conditionally independent given $\{K_0 = i\}$, a consequence of the semi-Markov structure of the process $\{\xi(t)\}$, see Lemma 1. Using expressions (9) and (10) and adding the terms in (15) we obtain the numerator of (13). The denominator is given by (11). This shows the validity of (12). □

Remark 1. It should be clear that (12) holds for a much larger class of systems with one (or several independent) Poisson arrival process(es). For example, we may define $\mathcal{B}_n$ to be an extension of $\mathcal{B}_2$ when the buffer has n cells where messages are stored according to the order of their arrivals and a message arriving to a full buffer is immediately rejected (the so-called M/GI/1/n queue). On the other hand, we may define $\mathcal{P}_n$ to be an extension of $\mathcal{P}_2$: messages are stored in an order that is reverse to their order of arrival; so if there is a message being processed in cell 1 at time t, the message in cell 2 arrived last before t while the message in cell n is the oldest; a new message arriving at a full buffer is always stored in cell 2, displacing the other messages upwards and expels the one sitting in cell 1 (oldest one). In both $\mathcal{B}_n$ and $\mathcal{P}_n$, the process ξ is semi-Markov and thus Proposition 1, depending only on this semi-Markov property, applies and formula (12) holds. In fact, one can assert that proposition holds for networks with i.i.d. message sojourn times, e.g., due to a single bottleneck server. We shall not attempt to formalize this further in this paper.

3 The $\mathcal{B}_2$ system

Recall that the $\mathcal{B}_2$ system is the same as a single server queue with buffer size 2. Under our Poisson assumption for the arrival process and i.i.d. assumptions for message sizes, this is further denoted by M/GI/1/2 in standard queueing theory. We are, however, interested not in the number of messages in the system neither on message delays but, rather, on the age of information process α. Assuming that α is stationary, we compute the Laplace transform of $\alpha(0)$ under $\mathbb{P}$ by using (12) which requires knowledge of $\mathbb{E}^0[e^{-s\alpha(0)}; K_0 = j]$, $j = 0, 1$. To obtain the latter, we consider the segment $[S_{-1}, S_0)$ and further condition on K_{-1} and summarize the results in Lemma 2 below. In what follows, we let τ, σ be two independent random variables, where τ is exponential with rate λ and σ has distribution G.

Lemma 2. *For $\mathscr{B}_2$,*

$$\mathbb{E}^0[e^{-s\alpha(S_0)}; K_{-1} = 0, K_0 = 0] = \hat{G}(\lambda)\hat{G}(s+\lambda), \tag{16}$$

$$\mathbb{E}^0[e^{-s\alpha(S_0)}; K_{-1} = 0, K_0 = 1] = \hat{G}(\lambda)\left(\hat{G}(s) - \hat{G}(s+\lambda)\right), \tag{17}$$

$$\mathbb{E}^0[e^{-s\alpha(S_0)}; K_{-1} = 1, K_0 = 0] = \frac{\lambda}{\lambda - s}\left(\hat{G}(s) - \hat{G}(\lambda)\right)\hat{G}(s+\lambda), \tag{18}$$

$$\mathbb{E}^0[e^{-s\alpha(S_0)}; K_{-1} = 1, K_0 = 1] = \frac{\lambda}{\lambda - s}\left(\hat{G}(s) - \hat{G}(\lambda)\right)\left(\hat{G}(s) - \hat{G}(s+\lambda)\right). \tag{19}$$

Proof. Recall that the K_n are i.i.d. with distribution (8): $\mathbb{P}^0(K_n = 0) = \hat{G}(\lambda)$. We shall consider the four cases separately and, in each case, we shall be referring to the definition (3) to figure out what $\alpha(0)$ is.

Case 1. $K_{-1} = 0, K_0 = 0$. Observe $\alpha(0) = S_0 - T_0$, see Figure 3. But

$$\mathbb{P}^0(S_0 - T_0 \in dx \mid K_0 = 0, K_{-1} = 0) \;=\; \mathbb{P}(\sigma \in dx \mid \sigma < \tau) \;=\; \frac{e^{-\lambda x} dG(x)}{\hat{G}(\lambda)},$$

and so

$$\mathbb{E}^0[e^{-s\alpha(S_0)}; K_{-1} = 0, K_0 = 0] \;=\; \int_0^\infty e^{-sx}\frac{e^{-\lambda x}}{\hat{G}(\lambda)} dG(x)\left(\hat{G}(\lambda)\right)^2 \;=\; \hat{G}(\lambda)\hat{G}(s+\lambda).$$

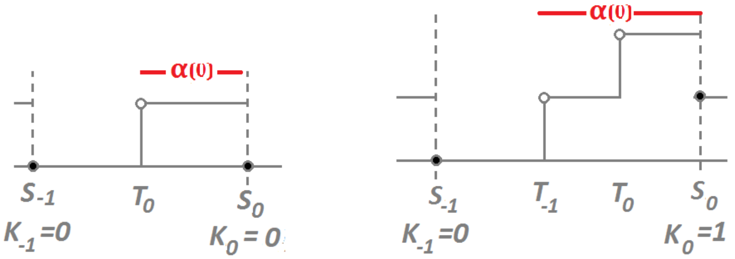

Figure 3: *The segment* $[S_{-1}, S_0)$ *when* $K_{-1} = K_0 = 0$ *(left) and* $K_{-1} = 0$, $K_0 = 1$ *(right), with* $S_0 = 0$.

Case 2. $K_{-1} = 0, K_0 = 1$. We have $\alpha(0) = S_0 - T_{-1}$, see Figure 3. Since

$$\mathbb{P}^0(S_0 - T_{-1} \in dx \mid K_0 = 1, K_{-1} = 0) \;=\; \mathbb{P}(\sigma \in dx \mid \tau < \sigma) \;=\; \frac{\left(1 - e^{-\lambda x}\right) dG(x)}{1 - \hat{G}(\lambda)}$$

we obtain

$$\begin{aligned}\mathbb{E}^0[e^{-s\alpha(S_0)}; K_{-1} = 0, K_0 = 1] \;&=\; \mathbb{P}^0(K_{-1} = 0, K_0 = 1)\mathbb{E}^0[e^{-s\alpha(S_0)} \mid K_{-1} = 0, K_0 = 1] \\ &=\; \hat{G}(\lambda)\int_0^\infty e^{-sx}\left(1 - e^{-\lambda x}\right)dG(x) \;=\; \hat{G}(\lambda)\left(\hat{G}(s) - \hat{G}(s+\lambda)\right).\end{aligned}$$

Case 3. $K_{-1} = 1, K_0 = 0$. To figure out $\alpha(0)$ we are here forced to consider two consecutive segments. We then have

$$\alpha(0) = (S_0 - S_{-1}) + (S_{-1} - T_0),$$

see Figure 4. Note that

$$S_0 - S_{-1} \text{ and } S_{-1} - T_0 \text{ are conditionally independent given } K_{-1} = 1$$

with $\mathbb{P}^0(S_0 - S_{-1} \in dx; K_0 = 0 \mid K_{-1} = 1) = \mathbb{P}(\sigma \in dx; \sigma < \tau)$ and $\mathbb{P}^0(S_{-1} - T_0 \in dx; K_{-1} = 1 \mid K_{-2} = 0) = \mathbb{P}(\sigma - \tau \in dx; \sigma > \tau)$, respectively. Thus

$$\begin{aligned} \mathbb{E}^0[e^{-s\alpha(0)}; K_{-1} = 1, K_0 = 0] &= \mathbb{E}[e^{-s\sigma}; \sigma < \tau]\,\mathbb{E}[e^{-s(\sigma-\tau)}; \sigma > \tau] \\ &= \mathbb{E}[e^{-s\sigma}e^{-\lambda\sigma}]\,\mathbb{E}[e^{-s\sigma}\int_0^\sigma \lambda e^{-(\lambda-s)t}dt] \;=\; \hat{G}(s+\lambda)\frac{\lambda}{\lambda-s}\big(\hat{G}(s) - \hat{G}(\lambda)\big). \end{aligned}$$

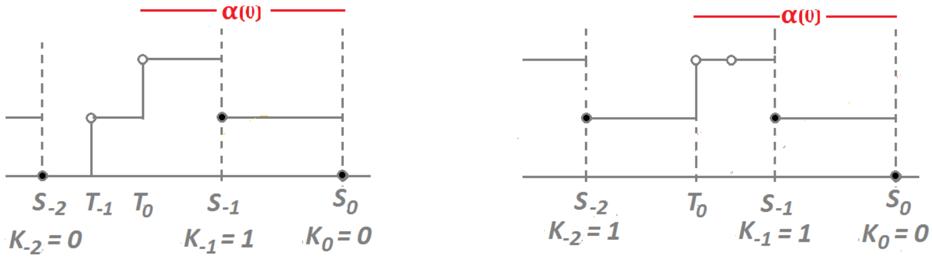

Figure 4: *The segments* $[S_{-2}, S_{-1}), [S_{-1}, S_0)$ *when* $K_{-1} = 1, K_0 = 0$ *in the case where* $K_{-2} = 0$ *(left) and* $K_{-2} = 1$ *(right).*

Case 4. Again, we have to consider two consecutive segments to realize that

$$\alpha(0) = S_0 - T_{-1} = (S_{-1} - T_{-1}) + (S_0 - S_{-1}),$$

see Figure 5. The two random variables

$$(S_{-1} - T_{-1}) \text{ and } (S_0 - S_{-1}) \text{ are conditionally independent given that } K_{-1} = 1$$

and thus

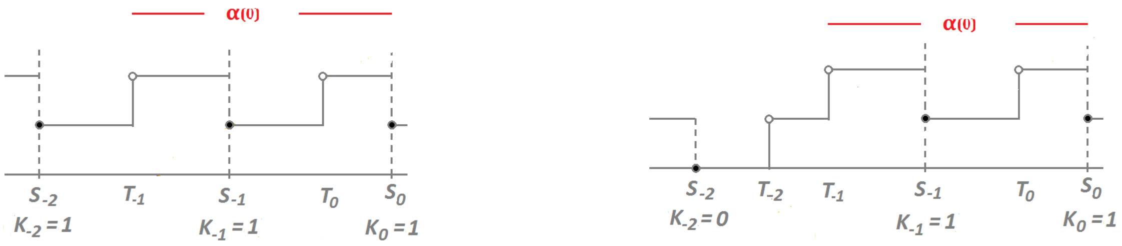

Figure 5: *The segments* $[S_{-2}, S_{-1}), [S_{-1}, S_0)$ *when* $K_{-1} = 1, K_0 = 1$ *in the case where* $K_{-2} = 0$ *(left) and* $K_{-2} = 1$ *(right).*

$$\begin{aligned} \mathbb{E}^0[e^{-s\alpha(0)}; K_{-1} = 1, K_0 = 1] &= \mathbb{E}^0[e^{-s(S_{-1}-T_{-1})-s(S_0-S_{-1})}; K_{-1} = 1, K_0 = 1] \\ &= \mathbb{E}[e^{-s(\sigma-\tau)}; \sigma > \tau]\,\mathbb{E}[e^{-s\sigma}; \sigma > \tau] \;=\; \mathbb{E}\left[\int_0^\sigma e^{-s(\sigma-t)}\lambda e^{-\lambda t}dt\right]\mathbb{E}\left[e^{-s\sigma}\left(1 - e^{-\lambda\sigma}\right)\right] \\ &= \frac{\lambda}{\lambda - s}\big(\hat{G}(s) - \hat{G}(\lambda)\big)\big(\hat{G}(s) - \hat{G}(s+\lambda)\big). \end{aligned}$$

This completes the proof. □

Define

$$\hat{G}_I(s) = \frac{1 - \hat{G}(s)}{s}\mu. \tag{20}$$

This is the Laplace transform of a probability measure G_I that is well-known in renewal theory: If we consider a renewal process with points, say, Z_n, $n \in \mathbb{Z}$, such that $Z_0 = 0$ and $Z_{n+1} - Z_n$ having distribution G, then there it has a stationary version (with no point at 0) and in such a way that Z_1 has distribution G_I.

Theorem 1. *For $\mathcal{B}_2$, the Laplace transform of the stationary Age of Information is given by*

$$\mathbb{E}[e^{-s\alpha(0)}] \;=\; \hat{G}(s)\left(\hat{G}(\lambda)+\lambda\frac{\hat{G}(s)-\hat{G}(\lambda)}{\lambda-s}\right)\left(\frac{\hat{G}(\lambda)}{\frac{\lambda}{\mu}+\hat{G}(\lambda)}\frac{\lambda}{\lambda+s}\frac{\hat{G}(s+\lambda)}{\hat{G}(\lambda)}+\frac{\frac{\lambda}{\mu}}{\frac{\lambda}{\mu}+\hat{G}(\lambda)}\hat{G}_I(s)\right). \tag{21}$$

Proof. Summing (16) and (17) we obtain

$$\begin{aligned}\mathbb{E}^0[e^{-s\alpha(S_0)};K_0=0] &= \hat{G}(s+\lambda)\left[\frac{\lambda}{\lambda-s}\hat{G}(s)-\frac{s}{\lambda-s}\hat{G}(\lambda)\right]\\ \mathbb{E}^0[e^{-s\alpha(S_0)};K_0=1] &= \left(\hat{G}(s)-\hat{G}(s+\lambda)\right)\left[\frac{\lambda}{\lambda-s}\hat{G}(s)-\frac{s}{\lambda-s}\hat{G}(\lambda)\right].\end{aligned}$$

Substituting the last two lines into the right hand side of (12) we obtain

$$\mathbb{E}[e^{-s\alpha(0)}] = \frac{\hat{G}(s)\left[\frac{\lambda}{\lambda-s}\hat{G}(s)-\frac{s}{\lambda-s}\hat{G}(\lambda)\right]\left(\frac{\hat{G}(s+\lambda)}{s+\lambda}+\frac{1-\hat{G}(s)}{s}\right)}{\frac{1}{\lambda}\left(\frac{\lambda}{\mu}+\hat{G}(\lambda)\right)}, \tag{22}$$

which gives (21) if we take into account the definition of $\hat{G}_I$. □

Corollary 1. *Expression (21) gives the stationary AoI as a sum of three independent random variables. In particular, the middle term in the right hand side of (21) corresponds to the Laplace transform of the random variable $(\sigma-\tau)^+$. Moreover, the expectation of $\alpha(0)$ is given by*

$$\mathbb{E}[\alpha(0)] \;=\; \frac{2}{\mu}-\frac{1-\hat{G}(\lambda)}{\lambda}+\frac{\hat{G}(\lambda)-\lambda\hat{G}'(\lambda)+\frac{1}{2}\lambda^2\int_0^\infty x^2dG(x)}{\lambda\left(\frac{\lambda}{\mu}+\hat{G}(\lambda)\right)}.$$

We obtained this corollary directly from the Laplace transform (21) where we recognize that $\frac{1}{2}\mu\int_0^\infty x^2dG(x) = \int_0^\infty x dB_I(x)$. Notice that if the message size has high variance then so does $\mathbb{E}\alpha(0)$. In particular, $\mathbb{E}\alpha(0)=\infty$ if $\int x^2dG(x)=\infty$. Rather than seeing this as a problem, one should change the point of view and adopt another function of α as a performance measure, for instance, $\mathbb{E}\alpha(0)^p$ for some $p<1$.

Corollary 2. *For $\mathcal{B}_2$, with G being exponential with mean $1/\mu$ we have*

$$\begin{aligned}\mathbb{E}[e^{-s\alpha(0)}] &= \left(\frac{\mu}{s+\mu}\right)^3\frac{\lambda}{s+\lambda}\frac{s^2+2s(\lambda+\mu)+\lambda^2+\lambda\mu+\mu^2}{(\lambda^2+\lambda\mu+\mu^2)},\\ \mathbb{E}[\alpha(0)] &= \frac{3\lambda^3+2\lambda^2\mu+2\lambda\mu^2+\mu^3}{\lambda\mu\left(\lambda^2+\lambda\mu+\mu^2\right)}.\end{aligned}$$

Indeed, if G is exponential, $\hat{G}(s)=\hat{G}_I(s)=\frac{\mu}{s+\mu}$. These expressions are consistent with [10] in this case. Inverting the last Laplace transform gives a measure with density equal to (5). In the limit as $\lambda\to\infty$ we see that $\alpha(0)$ is distributed as the sum of 3 i.i.d. exponentials. This is clear from the definition of $\mathcal{B}_2$: In the limit, the buffer always contains 2 messages and one is always available to be accepted once there is space available.

4 The $\mathcal{P}_2$ system

We remind the reader that $\mathcal{P}_2$ differs from $\mathcal{B}_2$ in that the arriving message is always admitted by replacing the message (if any) sitting in the second cell of the buffer, see Figure 1. Again, $\mathcal{P}_2$ is not service-preemptive: once a message starts being processed it will not be interrupted. The strategy for obtaining the Laplace transform of $\alpha(0)$ is the same as before. We make use of (12) of Proposition 1 which needs computation of the quantities involving $\alpha(0)$ in its right-hand side. The analog of Lemma 2 is Lemma 3 below which looks conspicuously the same. In fact, the first two formulas are identical. The last two differ.

Lemma 3. *For $\mathcal{P}_2$,*

$$\mathbb{E}^0[e^{-s\alpha(S_0)}; K_{-1} = 0, K_0 = 0] = \hat{G}(\lambda)\hat{G}(s+\lambda), \tag{23}$$

$$\mathbb{E}^0[e^{-s\alpha(S_0)}; K_{-1} = 0, K_0 = 1] = \hat{G}(\lambda)\left(\hat{G}(s) - \hat{G}(s+\lambda)\right), \tag{24}$$

$$\mathbb{E}^0[e^{-s\alpha(S_0)}; K_{-1} = 1, K_0 = 0] = \frac{\lambda}{\lambda+s}\left(1 - \hat{G}(s+\lambda)\right)\hat{G}(s+\lambda), \tag{25}$$

$$\mathbb{E}^0[e^{-s\alpha(S_0)}; K_{-1} = 1, K_0 = 1] = \frac{\lambda}{\lambda+s}\left(1 - \hat{G}(\lambda+s)\right)\left(\hat{G}(s) - \hat{G}(s+\lambda)\right) \tag{26}$$

Proof. 1) When $K_{-1} = 0, K_0 = 0$ or when $K_{-1} = 0, K_0 = 1$ the AoI same as in the $\mathcal{B}_2$ system, the reason being that the number of messages in the system is always at most 1, see Figure 3.

2) Suppose next that $K_{-1} = 1, K_0 = 0$. In Figure 6 we depict the two scenaria corresponding to the possible values of K_{-2}, namely $(K_{-2}, K_{-1}, K_0) = (0,1,0)$ or $(1,1,0)$. In both cases, (3) and the system dynamics imply that

$$\alpha(0) = S_0 - S_{-1} + V,$$

V is the time elapsed between the last arrival in the interval (S_{-2}, S_{-1}) and S_{-1}. If there is only one arrival in this interval then $V := S_{-1} - T_0$. In any case,

conditionally on $\{K_{-1} = 1, K_0 = 0\}$, the random variables V and $S_0 - S_{-1}$ are independent.

Therefore,

$$\mathbb{E}^0[e^{-s(S_0 - S_{-1} + V)} \mid K_{-1} = 0, K_0 = 1] = \mathbb{E}^0[e^{-s(S_0 - S_{-1})} \mid K_{-1} = 0, K_0 = 1]\, \mathbb{E}^0[e^{-sV} \mid K_{-1} = 0, K_0 = 1]$$

and the first factor on the right is easy:

$$\mathbb{E}^0[e^{-s(S_0 - S_{-1})} \mid K_{-1} = 0, K_0 = 1] = \int_0^\infty e^{-sx} \frac{e^{-\lambda x} dG(x)}{\hat{G}(\lambda)} = \frac{\hat{G}(s+\lambda)}{\hat{G}(\lambda)}.$$

To evaluate $\mathbb{E}^0[e^{-sV} \mid K_{-1} = 0, K_0 = 1]$ we note that, V is the distance of the last Poisson point inside the interval (T_{-1}, S_{-1}) (in the left scenario of Figure 6) or the interval (T_{-1}, S_{-1}) in the right scenario. (In both cases the length of the interval is that of a message size conditioned on containing at least one Poisson point.) To obtain the Laplace transform of V look backward in time starting from S_{-1} until the first Poisson point appears and condition on the event that this occurs between S_{-1} and S_{-2}. Thus, the density of V at $v > 0$ is

$$\frac{(1 - G(v))\lambda e^{-\lambda v}}{1 - \hat{G}(\lambda)},$$

which gives

$$\mathbb{E}^0[e^{-sV} \mid K_{-1} = 0, K_0 = 1] = \int_0^\infty e^{-sv} \frac{\lambda e^{-\lambda v}(1 - G(v))}{1 - \hat{G}(\lambda)} dv = \frac{\lambda}{\lambda+s} \frac{1 - \hat{G}(s+\lambda)}{1 - \hat{G}(\lambda)}. \tag{27}$$

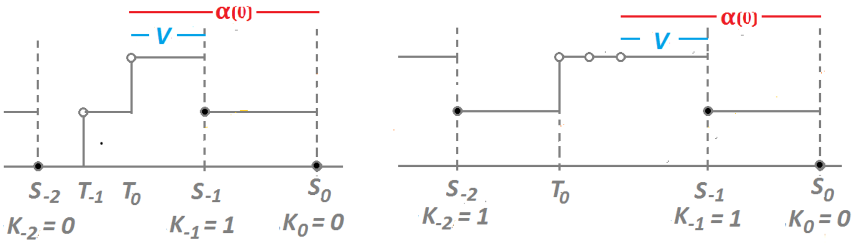

Figure 6: *Under $\mathcal{P}_2$, the segments* $[S_{-2}, S_{-1})$, $[S_{-1}, S_0)$ *when* $K_{-1} = 1, K_0 = 0$ *in the case where* $K_{-2} = 0$ *(left) and* $K_{-2} = 1$ *(right).*

Putting these together we obtain (25).

3) Finally, assume that $K_{-1} = 1, K_0 = 1$. This situation is similar to the previous one and thus will be treated succinctly. We are guided by Figure 7. Firstly, we have

$$\mathbb{E}^0[e^{-s(S_0-S_{-1})} \mid K_{-1} = 1, K_0 = 1] \;=\; \int_0^\infty e^{-sx} \frac{1-e^{-\lambda x} dG(x)}{1-\hat{G}(\lambda)} \;=\; \frac{\hat{G}(s)-\hat{G}(s+\lambda)}{1-\hat{G}(\lambda)}.$$

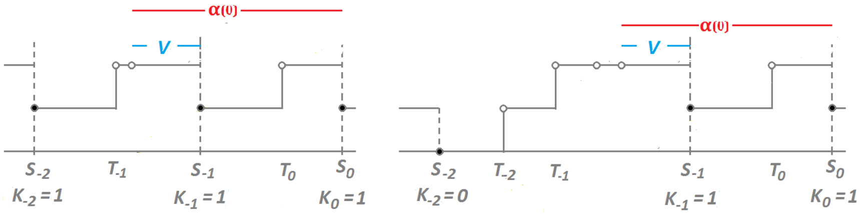

Figure 7: *Under $\mathcal{P}_2$, the segments* $[S_{-2}, S_{-1})$, $[S_{-1}, S_0)$ *when* $K_{-1} = 1, K_0 = 1$ *in the case where* $K_{-2} = 0$ *(left) and* $K_{-2} = 1$ *(right).*

Secondly, the argument used to derive (27) can be used here again with no changes to obtain

$$\mathbb{E}^0[e^{-sV} \mid K_{-1} = 1, K_0 = 1] \;=\; \int_0^\infty e^{-sv} \frac{\lambda e^{-\lambda v}(1-G(v))}{1-\hat{G}(\lambda)} dv \;=\; \frac{\lambda}{\lambda+s}\,\frac{1-\hat{G}(s+\lambda)}{1-\hat{G}(\lambda)}.$$

Putting these together we obtain (25) as well. □

The formula for $\mathbb{E}e^{-s\alpha(0)}$ now clear.

Theorem 2. *For $\mathcal{P}_2$, the Laplace transform of the stationary Age of Information is given by*

$$\mathbb{E}[e^{-s\alpha(0)}] \;=\; \hat{G}(s)\left(\hat{G}(\lambda) + \frac{\lambda}{\lambda+s}\left(1-\hat{G}(s+\lambda)\right)\right)\left(\frac{\hat{G}(\lambda)}{\frac{\lambda}{\mu}+\hat{G}(\lambda)}\,\frac{\lambda}{\lambda+s}\,\frac{\hat{G}(s+\lambda)}{\hat{G}(\lambda)} + \frac{\frac{\lambda}{\mu}}{\frac{\lambda}{\mu}+\hat{G}(\lambda)}\hat{G}_I(s)\right) \tag{28}$$

Proof. Adding up (23) and (24) of Lemma 3 and similarly (25) and (26) we obtain

$$\mathbb{E}[e^{-s\alpha(S_0)}; K_0 = 0] = \hat{G}(s+\lambda)\left[\hat{G}(\lambda) + \frac{\lambda}{\lambda+s}\left(1-\hat{G}(s+\lambda)\right)\right]$$
$$\mathbb{E}[e^{-s\alpha(S_0)}; K_0 = 1] = \left(\hat{G}(s)-\hat{G}(s+\lambda)\right)\left[\hat{G}(\lambda) + \frac{\lambda}{\lambda+s}\left(1-\hat{G}(\lambda+s)\right)\right].$$

Substituting these expressions in the numerator of (12), and recalling the definition (20) of G_I, we obtain (28). □

An alternative expression for (28) is:

$$\mathbb{E}[e^{-s\alpha(0)}] \;=\; \frac{\hat{G}(s)\left(\hat{G}(\lambda) + \frac{\lambda}{\mu}\hat{G}_I(s+\lambda)\right)\left(\frac{\lambda}{\lambda+s}\hat{G}(s+\lambda) + \frac{\lambda}{\mu}\hat{G}_I(s)\right)}{\frac{\lambda}{\mu} + \hat{G}(\lambda)}. \tag{29}$$

Corollary 3. *In expression (28) we recognize that $\alpha(0)$ is equal in distribution to the sum of three independent random variables, of which the middle one, $\hat{G}(\lambda) + \frac{\lambda}{\lambda+s}\left(1 - \hat{G}(s+\lambda)\right)$, is the Laplace transform of the random variable $\tau\mathbf{1}_{\tau>\sigma}$. Moreover,*

$$\mathbb{E}[\alpha(0)] \;=\; \frac{1}{\mu} + \frac{1}{\lambda}\left(1 - \hat{G}(\lambda) + \lambda\hat{G}'(\lambda)\right) + \frac{1}{\lambda}\,\frac{1}{\frac{\lambda}{\mu} + \hat{G}(\lambda)}\left(\hat{G}(\lambda) - \lambda\hat{G}'(\lambda) + \frac{1}{2}\lambda^2\hat{G}''(0)\right) \tag{30}$$

One should compare this to the expectation for the $\mathcal{P}_1$ system, [9],

$$\mathbb{E}[\alpha(0)] \;=\; \frac{1}{\lambda\hat{G}(\lambda)}.$$

Corollary 4. *For $\mathcal{P}_2$ with exponential message sizes,*

$$\mathbb{E}[e^{-s\alpha(0)}] = \frac{\mu}{\mu+s}\left(\frac{\mu}{\mu+\lambda} + \frac{\lambda}{\lambda+\mu+s}\right)\left(\frac{\mu^2}{\lambda^2+\lambda\mu+\mu^2}\,\frac{\lambda}{\lambda+s}\,\frac{\lambda+\mu}{\lambda+\mu+s} + \frac{\lambda^2+\lambda\mu}{\lambda^2+\lambda\mu+\mu^2}\,\frac{\mu}{\mu+s}\right),$$

$$\mathbb{E}[\alpha(0)] = \frac{2\lambda^5 + 7\lambda^4\mu + 8\lambda^3\mu^2 + 7\lambda^2\mu^3 + 4\lambda\mu^4 + \mu^5}{\lambda\mu\,(\lambda+\mu)^2\,(\lambda^2+\lambda\mu+\mu^2)},$$

and, with $\rho = \lambda/\mu$, the standard deviation of $\alpha(0)$ under $\mathbb{P}$ is

$$\mathrm{sd}_{\mathbb{P}}(\alpha(0)) = \frac{1}{\mu}\,\frac{\sqrt{2\rho^{10} + 12\rho^9 + 35\rho^8 + 60\rho^7 + 66\rho^6 + 56\rho^5 + 45\rho^4 + 34\rho^3 + 18\rho^2 + 6\rho + 1}}{\rho\,(\rho+1)^2\,(\rho^2+\rho+1)}.$$

The expectation (consistent with [10] in this case) and variance have been computed by summing up the expectations and variance of the three independent random variables comprising $\alpha(0)$. Inverting $\mathbb{E}e^{-s\alpha(0)}$ shows that $\alpha(0)$ has the density (4). As $\lambda \to \infty$, the distribution of $\alpha(0)$ tends to the distribution of the sum of 2 independent copies of σ. This is clear from the dynamics of $\mathcal{P}_2$.

5 Comparisons and numerics

To simplify life, we shall now use normalized units, assuming $\mu = 1$. We use [9] for the formulas (31) and (32) below. For $\mathcal{P}_1$ (buffer of size 1, where an arriving message immediately pushes out the message being processed), we have

$$\mathbb{E}[e^{-s\alpha(0)}] = \frac{\rho\hat{G}(s+\rho)}{s + \rho\hat{G}(s+\rho)}, \quad \mathbb{E}[\alpha(0)] = \frac{1}{\rho\hat{G}(\rho)}. \tag{31}$$

For $\mathcal{B}_1$ (buffer of size 1, where an arriving message is instantly rejected if a message is being processed), we have [9]:

$$\mathbb{E}[e^{-s\alpha(0)}] = \frac{\rho}{1+\rho}\cdot\frac{(s+\rho-\rho\hat{G}(s))\hat{G}(s)}{s(s+\rho)}, \quad \mathbb{E}[\alpha(0)] = 1 + \frac{1}{\rho} + \frac{\rho}{2}\cdot\frac{\mathbb{E}\sigma^2}{1+\rho}. \tag{32}$$

We now have information about all systems that we now compare. The comparisons depend on the message size distributions. We choose to consider two "extremes". First, exponentially distributed size; Second, deterministic.

5.1 Exponential message sizes

Using the notation $M(\rho)$ for $\mathbb{E}[\alpha(0)]$, recall that $\mu = 1$. We summarize the comparisons in a plot:

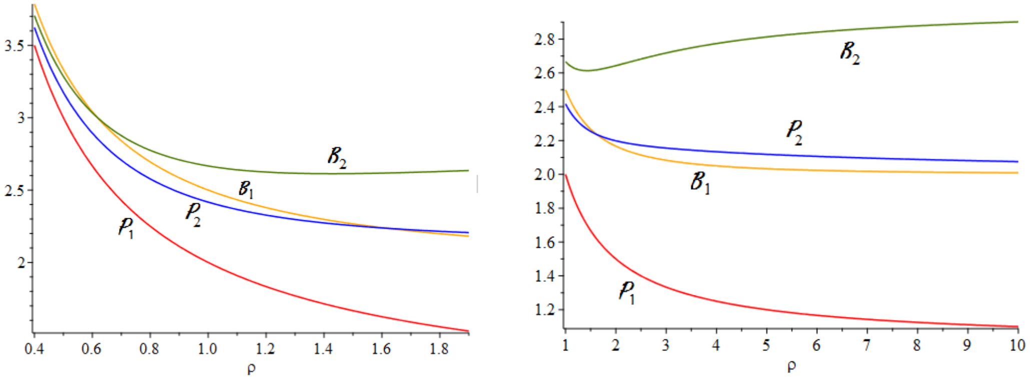

Figure 8: *Mean AoI as a function of ρ; right plot extends to high values of ρ*

Letting $M_{\mathcal{P}_1}$ to be the expectation for $\mathcal{P}_1$, etc., we see that

$$M_{\mathcal{P}_1}(\rho) < M_{\mathcal{P}_2}(\rho) < M_{\mathcal{B}_2}(\rho) \text{ for all } \rho.$$

The odd system is $\mathcal{B}_1$. For small ρ, $M_{\mathcal{B}_1}(\rho)$ is worst (highest); for large ρ, $M_{\mathcal{B}_1}(\rho)$ is between $M_{\mathcal{P}_1}(\rho)$ and $M_{\mathcal{P}_2}(\rho)$, whereas for intermediate ρ it is between $M_{\mathcal{P}_2}(\rho)$. The behavior for large ρ is easily explainable since it is easy to see that $\alpha(0)$ converges in distribution to e_1 when for $\mathcal{P}_1$, to $e_1 + e_2$ for $\mathcal{B}_1$ and $\mathcal{P}_2$ and to $e_1 + e_2 + e_3$ for $\mathcal{B}_2$, where e_1, e_2, e_3 are i.i.d. exponential random variables with unit expectation.

With some effort, we can also ask whether the comparisons above remain true in the sense of stochastic ordering. Recall that a real random variable X is stochastically smaller than Y, and write $X <_{\text{st}} Y$, if $\mathbb{P}(X > u) \leq \mathbb{P}(Y > u)$ for all $u \in \mathbb{R}$. Note that stochastic ordering is a partial order in the space of probability measures on the real line so two random variables may not be comparable at all. The information obtained from the Laplace transforms when message sizes are unit mean exponentials in Corollaries 2 and 4 as well as in equations (31) and (32) allows us to explicitly invert the Laplace transforms and then compare the tails of the distributions. We explain the observations in a couple of plots:

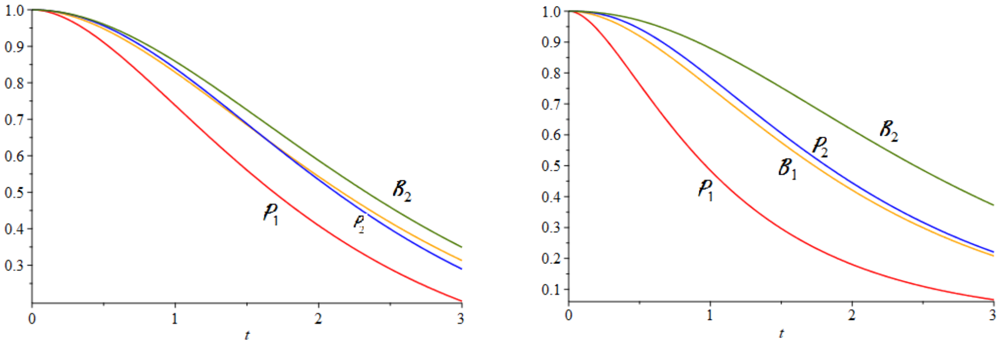

Figure 9: $\mathbb{P}(\alpha(0) > t)$ *plotted against t for small ρ on the left and high ρ on the right.*

Our observation is then that

$$\alpha_{\mathcal{P}_1} <_{\text{st}} \alpha_{\mathcal{P}_2} <_{\text{st}} \alpha_{\mathcal{B}_2} \text{ for all } \rho,$$

with the obvious understanding that $\alpha_{\mathcal{P}_1}$ stands for $\alpha(0)$ in stationarity for the system $\mathcal{P}_1$, etc. Moreover,

$$\alpha_{\mathcal{P}_1} <_{\text{st}} \alpha_{\mathcal{B}_1} <_{\text{st}} \alpha_{\mathcal{P}_2} \text{ for all sufficiently high } \rho.$$

The following figure gives plots of variances as functions of ρ. Note that they all converge to integers, as evidenced by (6) and the obvious uniform integrability.

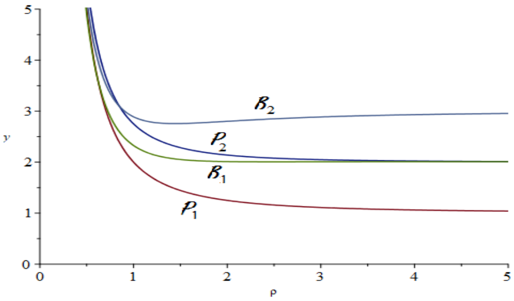

Figure 10: variances of all systems as a function of ρ

5.2 Deterministic message sizes

We now assume that $\mathbb{P}(\sigma = 1) = 1$: message sizes are all equal to 1 with probability 1. We can thus easily obtain $M(\rho)$ in all cases.

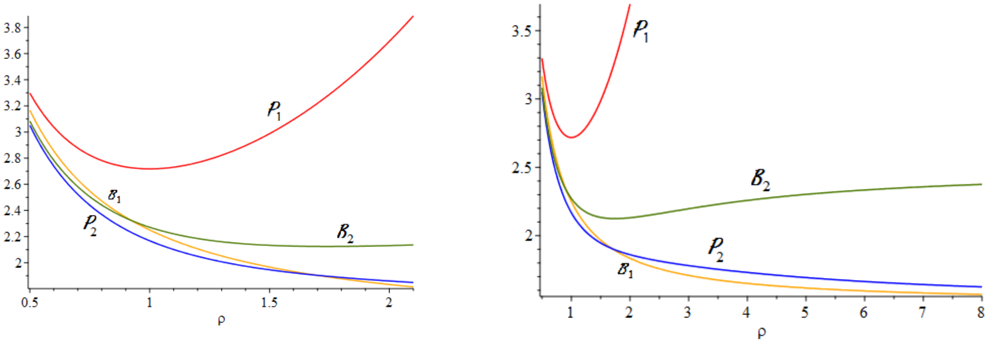

Figure 11: *Mean AoI as a function of ρ; right plot extends to high values of ρ*

We see that

$$M_{\mathcal{P}_2}(\rho) < M_{\mathcal{B}_2}(\rho) < M_{\mathcal{P}_1}(\rho) \text{ for all } \rho.$$

Whereas $\mathcal{P}_1$ was best in the exponential case, it is now worst. In fact, as $\rho \to \infty$, $M_{\mathcal{P}_1}(\rho) \to \infty$; moreover, $\alpha_{\mathcal{P}_1}$ tends to ∞ in distribution. This is natural: when ρ is high, the message being processed is constantly interrupted. The best system, from the point of view of expectation, seems to be $\mathcal{P}_1$. However, as in the exponential case, $\mathcal{B}_1$ is the odd system in that it is between $\mathcal{B}_2$ and $\mathcal{P}_1$ for small ρ, but $M_{\mathcal{B}_1}(\rho) < M_{\mathcal{P}_2}(\rho)$ for all large enough ρ. However, the difference between the two goes to 0 as $\rho \to \infty$. We can easily see that $\lim_{\rho\to\infty} M_{\mathcal{B}_1}(\rho) = \lim_{\rho\to\infty} M_{\mathcal{P}_2}(\rho) = 3/2$, while $\lim_{\rho\to\infty} M_{\mathcal{B}_2}(\rho) = 5/2$.

We again ask whether the comparisons in the mean translate to stochastic comparisons. To see this substitute $\hat{G}(s) = e^{-s}$ in the formulas for the Laplace transforms and try to invert them. We observe that all of them correspond to random variables with densities that can all be analytically computed with the exception of $\mathcal{P}_1$ that we discuss briefly in §5.2.1. For now, we summarize the comparisons in the plot below.

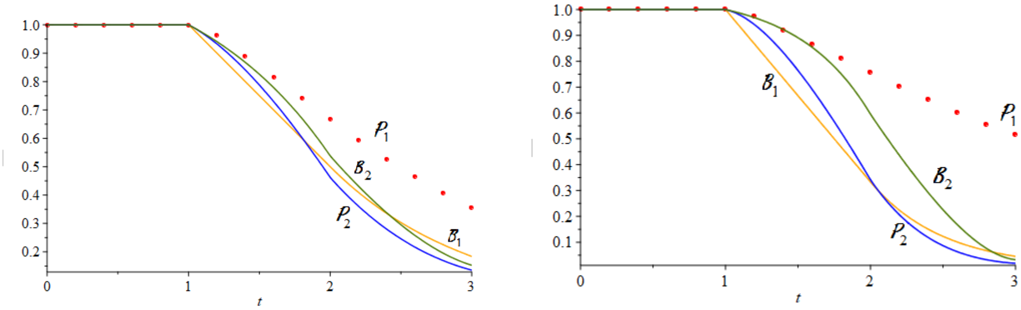

Figure 12: $\mathbb{P}(\alpha(0) > t)$ *plotted against t for small* ρ *on the left and high* ρ *on the right.*

Our observation is then that

$$\alpha_{\mathcal{P}_2} <_{\text{st}} \alpha_{\mathcal{B}_2} <_{\text{st}} \alpha_{\mathcal{P}_1}, \; \alpha_{\mathcal{B}_1} <_{\text{st}} \alpha_{\mathcal{P}_1} \text{ for all } \rho,$$

whereas $\alpha_{\mathcal{B}_1}$ is not comparable to any of the other three random variables.

We finally state without proof that, as $\rho \to \infty$, we have that $\alpha_{\mathcal{P}_2} \to U + 1$, while $\alpha_{\mathcal{B}_2} \to U + 2$, in distribution, where U is a uniform random variable on the interval $[0, 1]$. Figure 13 shows the densities for various values of ρ.

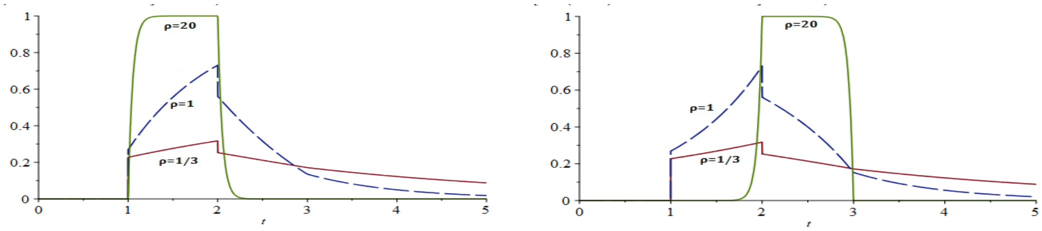

Figure 13: Densities of $\alpha(0)$ for $\mathcal{P}_2$ and $\mathcal{B}_2$ for various traffic intensities.

This, together with the evident uniform integrability, shows that $\lim_{\rho\to\infty} \text{var}_{\mathbb{P}}(\alpha_{\mathcal{P}_2}) = \lim_{\rho\to\infty} \text{var}_{\mathbb{P}}(\alpha_{\mathcal{B}_2}) = 1/12$.

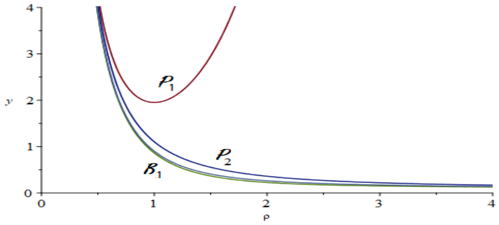

Figure 14: variances all systems as a function of ρ

5.2.1 The distribution of $\alpha_{\mathcal{P}_1}$

Let $\alpha_{\mathcal{P}_1}$ stand for a random variable that is distributed as $\alpha(t)$ for some (and hence all) t, where $\alpha(t)$, $t \in \mathbb{R}$, is the stationary version of the age of information process for the $\mathcal{P}_1$ system where $\sigma = 1$, almost surely. The reason we devote a little more ink on this is that it seems to be an upper bound on the age of information so long as message sizes are close to being deterministic. From (31) with $\hat{G}(s) = \exp(-s)$, we have

$$\mathbb{E}[e^{-s\alpha_{\mathcal{P}1}}] = \frac{1}{mse^s + 1} =: L_m(s),$$

with

$$m := e^{\rho}/\rho = \mathbb{E}[\alpha_{\mathcal{P}_1}].$$

This Laplace transform cannot be analytically inverted. However, since $L_m(s)$ as a function of a complex variable s has no singularity on the closed right-half plane $\mathrm{Re}(s) \geq 0$, we used the inversion formula [11, §8.2]

$$f_m(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \frac{e^{iy}}{1 + imye^{iy}} dy$$

together with a numerical approximation of this extended Riemann integral in order to obtain information about the tail of the distribution of $\alpha_{\mathcal{P}_1}$ which is plotted in Figure 12 as a dotted line. We can make a few remarks about f_m.

First, it is obviously a probability density function for all $m \geq e$ since we know that $\inf_{\rho>0} e^{\rho}/\rho = e$. One might conjecture that f_m remains a probability density function for all $m \geq 0$. However, L_m seizes to be completely monotone [5, XIII.4] for m small enough and thus, the conjecture is false. To see this, it is easy to see that $L''_m(s) < 0$ in a neighborhood of zero if (and only if) $m < 1$. We strengthen the conjecture by claiming that L_m is the Laplace transform of some random variable if and only if $m = 0$ (corresponding to a trivial random variable) or $m \geq e$.

Second, if since $L_m(s)$ is infinitely differentiable at all $s \geq 0$, we have that all moments of $\alpha_{\mathcal{P}_1}$. Using a formal series expansion, we can easily express the p-th moment in terms of the polynomial

$$Q_p(z) := \sum_{k=1}^{p} (p)_k k^{p-k} z^k, \quad p = 1, 2, \ldots$$

as follows

$$\mathbb{E}[\alpha^p_{\mathcal{P}_1}] = (-1)^p Q_p(-m).$$

with $m = e^{\rho}/\rho$, and then show that this is correct. It is easy to see that $Q_p(z)$ has a combinatorial interpretation. It is the generating function of the sequence $(p)_k\, k^{p-k}$, $1 \leq k \leq p$, where $(p)_k = p(p-1)\cdots(p-k+1)$, counting the number of ways to form k labeled groups, each with a distinct leader, using p different people. For values and other properties, see [12]. One can also see that $\mathbb{E}[\alpha^p_{\mathcal{P}_1}] \approx p! m^p$, that is, the p-th moment of an exponential random variable with the same mean.

6 Some final words

Using the Markov embedding and Palm inversion formula, we derived the stationary AoI distribution of $\mathcal{B}_2$ and $\mathcal{P}_2$ under Poisson message arrivals and generally distributed message sizes (processing times). The AoI of these systems and two other systems with low AoI, $\mathcal{B}_1$ and $\mathcal{P}_1$, were compared for exponential and deterministic service times, which, in some sense, are two extremes. We paid particular importance to their complementary distribution functions which play a role in important performance requirements. $\mathcal{P}_1$ has generally lower AoI for exponentially distributed message sizes, while $\mathcal{P}_2$ had lower AoI for deterministic message sizes when the traffic load was sufficiently low. The performance of $\mathcal{B}_1$ is interesting, having smallest AoI among these four considered policies in some cases. The $\mathcal{B}_2$ system may need to be used instead of these other three owing to technological constraints.

The rules of thumb derived can roughly be summarized as follows: If the message sizes are deterministic or nearly so then it is best to use $\mathcal{B}_1$ or $\mathcal{P}_2$, which is pretty close to $\mathcal{B}_1$. In particular, $\mathcal{P}_1$ seems to be an upper bound for AoI and this is the reason that we paid some special attention to it in §5.2.1 (where we had to resort to analytical tricks to approximate it, as there is no closed-form formula for the distribution). On the other hand, if message sizes are "very random", we expect the opposite: $\mathcal{P}_1$ performs best stochastically so. Granted, the study in this paper has been done only for Poisson arrivals and so, to be able to reach the generality of [9] one should use a different method which we leave for future work. We should also mention that dynamic policies should also be studied, that is, policies that decide on rejection or not of a message based on information such as the time the message has spent in the system, the arrival process up to the current time, etc.

Reiterating Remark 1, after Proposition 1, the approach employed above can also be applied to study the AoI distribution of other queueing systems with renewal points. In particular, further clarity is needed regarding the AoI for larger buffer sizes, particularly under LIFO queuing policies without service preemption.

Finally, we also mention that the AoI α defined in (3) may not be the most appropriate measure of freshness as it incorporates information about the arrival process as well. A different measure [9] is $\beta(t) = A(t) - A(D(t))$, in the notation of the processes introduced in (1) and (2), the distribution of which may differ significantly from that of α.

References

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

[2] Baccelli, F., and Brémaud. P. (2003). *Elements of Queueing Theory: Palm Martingale Calculus and Stochastic Recurrences*, 2nd Ed. Springer-Verlag, Berlin.

[3] Bedewy, A.M., Sun, Y. and Shroff, N.B. (2019). Minimizing the age of the information through queues. *IEEE Trans. Info. Theory* **65**, no. 8, 5215-5232.

[4] Cinlar, E. (1975). *Introduction to Stochastic Processes.* Prentice-Hall.

[5] Feller, W. (1971). *An Introduction to Probability Theory and its Applications*, Vol. 2, second edition, John Wiley.

[6] Kaul, S., Yates, R. and Gruteser, M. (2012). Real-time status: How often should one update? *Proc. 31st IEEE INFOCOM*, Orlando, Florida, pp. 2731-2735.

[7] Kaul, S., Yates, R. and Gruteser, M. Status updates through queues. In *Proc. Conf. on Information Sciences and Systems (CISS)*, March 2012.

[8] Kavitha, V., Altman, E. and Saha, I. (2018). Controlling packet drops to improve freshness of information. arXiv:1807.09325

[9] Kesidis, G., Konstantopoulos, T. and Zazanis, M.A. (2020). The new age of information: a tool for evaluating the freshness of information in bufferless processing systems. *Queueing Systems* **95**, 203-250, June 2020; http://arxiv.org/abs/1904.05924; https://arxiv.org/abs/1808.00443

[10] Kosta, A., Pappas, N. and Angelakis, V. (2017). Age of information: a new concept, metric, and tool. *Foundations and Trends in Networking* **12**, No. 3, 162-259.

[11] Marsden, J.E. and Hoffman, M.J. (1999). *Basic Complex Analysis,* third edition. Freeman, New York.

[12] The On-Line Encyclopedia of Integer Sequences; sequence A199673; https://oeis.org/A199673

George Kesidis
Computer Science Dep, The Pennsylvania State University, University Park, PA 16802, USA; gik2@psu.edu

Takis Konstantopoulos
Department of Mathematical Sciences, The University of Liverpool, Liverpool L69 7ZL, UK; takiskonst@gmail.com

Michael A. Zazanis
Department of Statistics, Athens University of Economics and Business, 76 Patission St., Athens 104 34, Greece; zazanis@aueb.gr