Smoothing Point Processes As a Means to Increase Throughput

Cheng-Shang Chang
IBM Research Division
T.J. Watson Research Center
P.O. Box 704
Yorktown Heights, NY 10598

(May, 1991; revision: Dec., 1992)

Abstract

Unlike the leaky-bucket scheme which regulates the input rate, the filters we study in this paper reduce the variability of interarrival times (subject to a maximum delay constraint for each customer). These filters are called "smoothing" filters for point processes. By considering the output processes of various queueing systems, we show that (i) an infinite-server queue acts as a smoothing filter for a doubly stochastic Poisson process and (ii) a single-server queue with deterministic service times acts as a smoothing filter for a stationary and ergodic point process. Based on (ii), we provide a smoothing algorithm that satisfies a maximum delay constraint. The algorithm is shown to be robust and consistent. We then consider two examples where these filters can be applied to increase throughput: one is the single-server loss system with exponential service times, the other is the leaky-bucket scheme with token buffer size 1. Numerical examples and simulations are also given for comparing the trade-off between delay and throughput. Our approach is based on the variability ordering and a cut criterion for the majorization ordering. The criterion appears to be new and to be of independent interest.

1 Introduction

Since the advent of fiber-optic technology, there have been great changes in today's telecommunication networks. Traditional feedback flow control mechanisms are now difficult and inefficient to implement due to the large ratio of the propagation delay to message transmission time. As a major part of open loop control, filtering input point processes at local nodes becomes an increasingly important issue. Unlike the leaky-bucket scheme recently proposed (see [35, 30, 31] and their references) for input rate regulation, our objectives in this paper are (i) to design filters that reduce the variability of various point processes (with a maximum delay constraint for each customer) and (ii) to provide formal proofs for systems in which the throughputs can be increased by reducing the variability of the input processes.

Our comparison of two point processes is based on the comparison of their one-dimensional stationary distributions of interarrival times. We use the variability ordering, a commonly used partial ordering in stochastic comparison theorems ([32, 29]), to compare the distributions of two random variables. An random variable is said to be less variable than another random variable if all its expectations over the class of convex functions are not greater.

Analogous to the leaky-bucket scheme, the filters we propose are related to some well-defined queueing systems. We then compare the interdeparture times from queueing systems with the interarrival times of the input processes. A filter is called a smoothing filter if the one-dimensional stationary distribution of the interdeparture times is less variable than that of the input process. We consider two classes of input processes in this paper: (i) the doubly stochastic Poisson processes with stationary and ergodic intensities (a nonhomogeneous Poisson process with the intensity function itself being a stochastic process) and (ii) the class of stationary and ergodic processes. Note that the latter is richer than the former.

For a doubly stochastic Poisson process, we show that an infinite-server queue acts as a smoothing filter. It is known that the output process is also a doubly stochastic Poisson process ([36]). If the intensity of a doubly stochastic Poisson process is periodic with period p, then the departure process from an infinite-server queue with service times uniformly distributed over $[0, p]$ is a Poisson process. This filter, among the class of infinite-server queues, is called the matched filter for the doubly stochastic Poisson process with a periodic intensity since the Poisson process is the least variable process among all doubly stochastic Poisson processes. Furthermore, we show that the variability of the interdeparture processes among the family of infinite-server queues with exponential service times is decreasing in the mean of the exponential service times. The relationship between these filters and the doubly stochastic Poisson processes is similar to that between the RC low pass circuit and signals.

For a stationary and ergodic process, we show that a single-server queue with deterministic service times acts as a smoothing filter. Based on this smoothing filter, we then propose an algorithm that satisfies a maximum delay constraint for each customer. This algorithm is proved to be robust and consistent, i.e., the variability of the interdeparture processes from the algorithm is decreasing in the delay constraint for all stationary and ergodic processes. Our approach for the proof of this algorithm is based on a cut criterion for the majorization ordering ([22]) which in turn leads to the variability ordering. This cut criterion for the majorization ordering appears to be new and has independent interest.

One would expect that smooth point processes results in high throughput for systems with finite capacity. We justify the statement for two examples: one is the single-server loss system with exponential service times which might be used for modelling a wireless communication network, the other is the leaky-bucket scheme with token buffer size 1. We show that the throughput of these two examples only depends on the one-dimensional distributions of the interarrival times and that the throughput can be increased if smoothing filters are added before feeding the input processes to the systems. The gain of throughput is at the cost of delay. To gain some insight into the trade-off between delay and throughput, we obtain numerical results and simulations for the single-server loss systems using the smoothing algorithms.

2 Preliminaries

In this section, we introduce some notions which we need in the presentation of our main results. These notions include comparison of random variables and random vectors and stability of sequences.

2.1 Cut criteria for the variability ordering and the majorization ordering

We start from the comparison of random variables. A random variable (r.v.) X_1 with the distribution function $G_1(x)$ is less variable ($\leq_c$) than an r.v. X_2 with the distribution function $G_2(x)$ if

$$\int_{-\infty}^{\infty} h(x)dG_1(x) \leq \int_{-\infty}^{\infty} h(x)dG_2(x) \quad (Eh(X_1) \leq Eh(X_2)),$$

for all h convex. We note that $X_1 \leq_c X_2$ implies that $EX_1 = EX_2$ since both functions, $h(x) = x$ and $h(x) = -x$, are convex. Moreover, $X_1 \leq_c X_2$ implies that the variance of X_1 is not larger than that of X_2 since $(X_1 - EX_1)^2$ is convex. Thus, this ordering is referred to as the variability ordering in this paper(see Stoyan [32] or Ross [29]). We note that sometimes this ordering is also known as the convex ordering or the dilation of measures. In what follows, we use both notations $X_1 \leq_c X_2$ and $G_1(x) \leq_c G_2(x)$.

For comparison of vectors, we introduce the notion of majorization (see Marshall and Olkin [22]).

Definition 2.1 *An n-vector a is majorized by an n-vector b, denoted $a \prec b$, if (i) $\sum_{i=1}^{j} a_{[i]} \leq \sum_{i=1}^{j} b_{[i]}$, $j = 1, \ldots, n-1$, where $a_{[i]}$ (resp. $b_{[i]}$) is the i-th largest element in a (resp. b) and (ii) $\sum_{i=1}^{n} a_i = \sum_{i=1}^{n} b_i$.*

There are several ways to characterize the majorization ordering in [22], we propose a new one (to our best knowledge) in the following lemma.

Lemma 2.2 *(Cut criterion for the majorization ordering) Consider two n-vectors, a and b.*

(i) *If (c1) $\sum_{j=1}^{n} a_j = \sum_{j=1}^{n} b_j$ and (c2) there exists a constant z such that for each j, either $z \leq a_j \leq b_j$ or $z \geq a_j \geq b_j$ is satisfied, then $a \prec b$.*

(ii) *If $a \prec b$, then there exists a finite sequence of n-vectors, $a^0, a^1, \ldots, a^n$ with $a^0 = a$ and $a^n = b$ such that the vectors a_j and a_{j+1} satisfy (c1) and (c2).*

Note that (c2) implies that every element in a is closer to z than the corresponding element in b. Intuitively, the vector a is more "balanced" or less "variable" than b. Thus, the majorization ordering can be viewed as a partial ordering for the comparison of "balance" or "variability" of two vectors.

Proof. (i) Without loss of generality, we assume that under (c2), $z \leq a_j \leq b_j$ for $j = 1, 2, \ldots, m$ and $z \geq a_j \geq b_j$ for $j = m+1, \ldots, n$. Clearly, we have $a_{[j]} \leq b_{[j]}$, $j = 1, \ldots, m$ and $a_{(j)} \geq b_{(j)}$, $j = 1, \ldots, n-m$, where $a_{[j]}$ and $b_{[j]}$ (resp. $a_{(j)}$ and $b_{(j)}$) are the j-th largest (resp. smallest) elements in a and b. Thus, $\sum_{i=1}^{j} a_{[i]} \leq \sum_{i=1}^{j} b_{[i]}$ for $j = 1, \ldots, m$. Note that $a_{[i]} = a_{(n+1-i)}$ and that

$$\sum_{i=1}^{j} a_{[i]} = \sum_{i=1}^{n} a_i - \sum_{i=1}^{n-j} a_{(i)}.$$

Using (c1) and $a_{(j)} \geq b_{(j)}$, $j = 1, \ldots, n-m$, yields $\sum_{i=1}^{j} a_{[i]} \leq \sum_{i=1}^{j} b_{[i]}$ for $j = m+1, \ldots, n$. This completes the argument for (i).

To show (ii), we use T-transformations ([22], p. 107). Since $a \prec b$, the vector a can be derived from b by applications of a finite number of transformations T having for some $1 \leq i < j \leq n$ and some $\gamma \in [0,1]$, the form

$$T(b) = (b_1, \ldots, b_{i-1}, \gamma b_i + (1-\gamma) b_j, b_{i+1}, \ldots, b_{j-1}, (1-\gamma) b_i + \gamma b_j, b_{j+1}, \ldots, b_n).$$

Note that the difference between $T(b)$ and b is the i-th and j-th elements. It is then easy to see that $T(b)$ and b satisfy the cut criterion with $z = (b_i + b_j)/2$. ■

We will use the following two properties (see [22]) in this paper. Note that the second property is an application of the first one.

Proposition 2.3 *(i) (Hardy, Littlewood and Polya [13]) $a \prec b$ if and only if $\sum_{j=1}^{n} g(a_j) \leq \sum_{j=1}^{n} g(b_j)$ for all g convex.*

(ii) ([22], Proposition 5.A.7) If the four vectors, a^1, a^2, b^1 and b^2 satisfy $a^1 \prec b^1$ and $a^2 \prec b^2$, then the concatenated vectors satisfy $(a^1, a^2) \prec (b^1, b^2)$.

Analogous to the majorization ordering, there are many ways to characterize the variability ordering (see [32, 29]). One of them is the cut criterion by Karlin and Novikoff [18]. In the following lemma, we show how the cut criterion for the variability ordering is related to that for the majorization ordering and provide a converse statement of the cut criterion for the variability ordering. Our argument will be based on discrete distributions with rational and bounded supports. Possible extension to general distributions could be obtained by a limiting argument.

Lemma 2.4 *(Cut criterion for the variability ordering) Let X_1 and X_2 be two random variables with distributions, $G_1(x)$ and $G_2(x)$. Assume these two distributions have rational and finite supports, i.e., the sets $\{x : P(X_i = x) > 0\}$, $i = 1$ and 2, have finite numbers of elements and all these elements are rational numbers.*

(i) *(Karlin and Novikoff [18]) If (d1) $EX_1 = EX_2$ and (d2) there exist some finite z for which $G_1(x) \leq G_2(x)$ for $x \leq z$ and $G_1(x) \geq G_2(x)$ for $x \geq z$, then $X_1 \leq_c X_2$.*

(ii) *If $X_1 \leq_c X_2$, then there exists a finite sequence of distributions, $F_0(x), \ldots, F_n(x)$ with $F_0(x) = G_1(x)$ and $F_n(x) = G_2(x)$ such that $F_j(x)$ and $F_{j+1}(x)$ satisfy (d1) and (d2) for all j.*

Proof. Our proof is based on a relation between the variability ordering and the majorization ordering first recognized by Karamata [17]. Since we assume that the distributions have rational and finite supports, there exist an integer n and two n-vectors a and b such that $P(X_1 = a_j) = P(X_2 = b_j) = 1/n$ for all j. Without loss of generality, we assume that $a_1 \geq a_2 \geq \ldots \geq a_n$ and $b_1 \geq b_2 \geq \ldots \geq b_n$. Note that $Eg(X_1) = (1/n) \sum_{j=1}^{n} g(a_j)$ and $Eg(X_2) = (1/n) \sum_{j=1}^{n} g(b_j)$ for any function g. It then follows from the definition of the variability ordering and Proposition 2.3(i) that $X_1 \leq_c X_2$ if and only if $a \prec b$. Using this relationship and Lemma 2.2, it suffices to show that (d1) and (d2) are equivalent to (c1) and (c2) in Lemma 2.2. By taking $g(t) = t$, it is clear that $EX_1 = EX_2$ if and only if $\sum_{j=1}^{n} a_j = \sum_{j=1}^{n} b_j$. By taking $g(t) = \mathbf{1}_{\{t \geq x\}}$ and $g(t) =$

$\mathbf{1}_{\{t\le x\}}$, we have that (i) $P(X_1 \le x) = 1/n\sum_{j=1}^n \mathbf{1}_{\{a_j\le x\}}$, (ii) $P(X_2 \le x) = 1/n\sum_{j=1}^n \mathbf{1}_{\{b_j\le x\}}$, (iii) $P(X_1 \ge x) = 1/n\sum_{j=1}^n \mathbf{1}_{\{a_j\ge x\}}$ and (iv) $P(X_2 \ge x) = 1/n\sum_{j=1}^n \mathbf{1}_{\{b_j\ge x\}}$. From these four equations, it is clear that (c2) implies (d2). To show the reverse statement, we need to show that $z < a_j$ implies $a_j \le b_j$ and that $z > a_j$ implies $a_j \ge b_j$ if (d2) is given. Suppose that $z < a_j$ and $a_j > a_{j+1}$. Choose $\epsilon > 0$ small enough such that $z < a_j - \epsilon$ and $a_j > a_{j+1} - \epsilon$. Thus, $P(X_1 \le a_j - \epsilon) = (n-j)/n \le P(X_2 \le b_j)$. From the assumption that $G_1(x) \ge G_2(x)$ for $x \ge z$ and that $z < a_j - \epsilon$, it follows that

$$P(X_2 \le b_j) \ge P(X_1 \le a_j - \epsilon) \ge P(X_2 \le a_j - \epsilon).$$

Since the distribution function is nondecreasing, we have $a_j - \epsilon \le b_j$. Letting $\epsilon \to 0$ yields $a_j \le b_j$. For the case that $z < a_j$ and $a_j = a_{j+1}$, one can argue recursively that $b_j \ge a_j$ using the previous argument and $b_j \ge b_{j+1}$. The other inequality, i.e., $z > a_j$ imples $a_j \ge b_j$, can be shown by a similar argument. ■

2.2 Stability and variability of sequences

In this subsection, we introduce the notion of stability ([24, 1]).

Definition 2.5 *A sequence of not necessarily independent r.v's T=$\{T_1,T_2,\ldots\}$ is called stable if for all $m \ge 1$ and $x_1,\ldots,x_m$,*

$$\lim_{n\to\infty}\frac{1}{n}\Sigma_{i=1}^n P(T_i < x_1, T_{i+1} < x_2, \ldots, T_{i+m-1} < x_m) = P(T_1^0 < x_1, T_2^0 < x_2, \ldots T_m^0 < x_m),$$

for some sequence $T^0 = \{T_1^0, T_2^0, \ldots\}$. We call the sequence, T^0, a stationary version of T and the distribution of T^0 the stationary distribution of T. If, furthermore, T is ergodic, then the sequence T is said to be ergodically stable.

For instance, the one-dimensional (marginal) stationary distribution, $G(x)$, is defined to be $\lim_{n\to\infty} 1/n\sum_{i=1}^n P(T_i \le x)$. The ergodicity implies that for all bounded functions h

$$\lim_{n\to\infty}\frac{1}{n}\sum_{j=1}^n h(T_j) = \lim_{n\to\infty}\frac{1}{n}\sum_{j=1}^n Eh(T_j) = \int_{-\infty}^{\infty} h(x)dG(x). \tag{1}$$

In this paper, we will assume the limits in (1) exist and equal even though h is not bounded. Readers who require more details about the convergence should verify the uniform integrability of the sequence [7].

Definition 2.6 *Consider two stable sequences of r.v.'s T and $\tilde{T}$ with one-dimensional distributions $G(x)$ and $\tilde{G}(x)$. The sequence T is less variable than $\tilde{T}$ if $G(x)$ is less variable than $\tilde{G}(x)$.*

There are other definitions for comparing two random sequences (see [32]). In this paper, we will concentrate on the comparison for one-dimensional distributions.

3 Smoothing doubly stochastic Poisson processes via $\cdot/GI/\infty$ queues

In this section, we consider a doubly stochastic Poisson process (DSP) $N_{\text{in}}(t)$ with an intensity function $\lambda_{\text{in}}(t)$. One can view a DSP process as a nonhomogeneous Poisson process with the intensity itself being a stochastic process. In other words, when conditioning on a sample path of the intensity, the DSP process is a nonhomogeneous Poisson process. Thus, DSP processes are also known as conditional Poisson processes. For a formal definition of DSP processes, we refer to Bremaud [2]. For applications of DSP processes in telecommunication traffic, we refer to [14], where the class of Markov Modulated Poisson processes, a subclass of DSP processes, is used as a model for voice/data integration. We assume that

[A1] $\lambda_{\text{in}}(t)$ is stationary ergodic, which implies that

$$E\int_0^1 \lambda_{\text{in}}(s)ds = E\lambda_{\text{in}}(t) = E\lambda_{\text{in}}(0) = \lim_{t\to\infty}\frac{1}{t}\int_0^t \lambda_{\text{in}}(s)ds = \lambda$$

for some constant λ. The constant λ is then called the average arrival rate of the DSP process. The filter we consider in this section is an infinite-server queue. In other words, we feed the input process into an infinite-server queue. The operation is as follows: when there is an arrival, we generate a *nonnegative* random variable S according to a certain distribution $F(x)$. Then we delay the arrival by the amount of time S.

Theorem 3.1 *The interdeparture sequence from an infinite-server queue is less variable than the interarrival sequence of the DSP process.*

As we will show later, the output process is still a DSP process. According to Rolski [25] and Chang and Pinedo [3], the least variable process among the family of DSP processes is the Poisson process. If the intensity process of the input process is periodic, i.e., $\lambda_{\text{in}}(t) = g(t+U)$ for some periodic function $g(t+p) = g(t)$ and U is a random variable uniformly distributed between 0 and p, then we have the following "matched filter" such that the output process from this filter is a Poisson process.

Theorem 3.2 *(Matched filter) For a DSP process with periodic intensity, the filter that minimizes the variability of the interdeparture sequences among the family of filters using infinite-server queues is the one with the distribution being uniformly distributed between 0 and the period p of the intensity function.*

For DSP processes with nonperiodic intensities, one still can smooth the input processes by feeding them into "low-pass filters." One of the most common low-pass filters is the RC circuit. Recall that the RC low-pass filter is a circuit connected by a resistor and a capacitor and that the impulse response of the RC circuit has an exponential tail. Thus, it is not surprising that the corresponding low-pass filters in our setting is an infinite-server queue with exponential service times. Moreover, the RC constant decides the sharpness of the low-pass filter. One expects that the variability of the interdeparture sequence decreases as the sharpness of the low-pass filter increases. The following theorem states the corresponding effect.

Theorem 3.3 *(Low-pass filters) The infinite-server queue with exponential service times acts as a RC low-pass filter for a DSP process, i.e., the variability of interdeparture sequences decreases as the mean of the exponential service times increases.*

We have a few remarks on the infinite-server queues as the smoothing filters for DSP processes:

(i) In order to satisfy a maximum delay constraint c for each customer, one can choose the service time distributions with supports bounded by c, i.e., $P(S > c) = 0$. However, it is not clear what distribution one should choose in order to minimize the variability of the departure process except the case that the input intensity is periodic.

(ii) Suppose we choose the uniform distribution between 0 and c to guarantee the maximum delay constraint c. We note that the variability of the departure process does not necessarily decrease as c increases. As shown in Theorem 3.2, the best filter for the DSP process with periodic intensity p is to choose $c = p$. In the next section, we will present filters that have this type of monotonic property in the maximum delay constraint.

(iii) Since the superposition of two DSP processes is still a DSP process, the superposition of two departure processes from infinite-server queues are less variable than the superposition of two input DSP processes.

(iv) Note that the departure process from an infinite-server queue may not be in the same order as the input process. Kurose [20] suggests the following resequencing algorithm such that the output is in same order as the input: The basic concept is relabelling the departure epochs. Let τ_n be the arrival epoch of the n-th customer and S_n be its random delay drawn from the distribution $F(x)$. Then $\{\tau_n + S_n, n \geq 1\}$ are departure epochs of the output process. Due to the randomness of S_n, $\tau_n + S_n$ may not be necessarily larger than $\tau_{n+1} + S_{n+1}$. Thus, let $\{\tau_n^1, n \geq 1\}$ be the sequence of $\{\tau_n + S_n, n \geq 1\}$ after sorting in the increasing order. Assign τ_n^1 as the departure epoch of the n-th customer. Clearly, the output now is in the same order as the input. Since τ_n^1 is the n-th largest element in the sequence $\{\tau_n + S_n, n \geq 1\}$, $\tau_n^1 \geq \tau_{n_i} + S_{n_i}$ for some indices, n_i, $i = 1, 2, \ldots, n$. This implies that $\tau_n^1 \geq \max_{1 \leq i \leq n}(\tau_{n_i}) \geq \tau_n$. Thus, the algorithm is realizable. In fact, it can be implemented by a customer queue and a token pool (a modified leaky bucket scheme). At the n-th arrival epoch τ_n, the n-th customer is added to (the back of) the customer queue. At the same time, a token is also generated and this token is scheduled to be released at $\tau_n + S_n$. When a token is released, a customer is also released from (the front of) the customer queue.

The rest of this section is devoted to the proofs of these three theorems.

The following proposition regarding the output process of $DSP/GI/\infty$ queue is adapted from Vere-Jones [36].

Proposition 3.4 *(Vere-Jones) The output process of a $DSP/GI/\infty$ queue with the density function of the service times $f(t)$ is still a DSP process. The intensity function of the output process $\lambda_{out}(t)$ is the convolution of the input intensity and the service time distribution, i.e.,*

$$\lambda_{out}(t) = \int_0^\infty \lambda_{in}(t-s) f(s) ds.$$

We note that the input-output relation for the intensities is of the same form as a signal passing through a linear time invariant filter with impulse response $f(t)$. Keilson and Servi [19] present a probabilistic argument for the input-output relation as follows: first, one observes that the output process of a $DSP/D/\infty$ queue with service times equal to z is merely the original process with a deterministic delay for the amount of time z. Thus, the output process is still a DSP process with intensity $\lambda_{\text{in}}(t-z)$. Now suppose the service time distribution is a mixture of a finite number of values, e.g. z_j with probability p_j, $j = 1, \ldots, k$. Since the sum of independent DSP processes is still a DSP process with the intensity being the sum of all the intensities. One can view the input DSP process as a superposition of k independent DSP processes with service times z_j, $j = 1, \ldots, k$. The output process is then again a superposition of independent DSP processes and its intensity is of the form of a convolution. The general distribution case is approximated by considering a sequence of distributions with a finite number of values and taking the limit.

Via the form of a convolution, it is clear that the intensity function of the output process is still stationary and ergodic. Since the intensity functions for both input and output processes are stationary and ergodic, both DSP processes are stationary and ergodic (see Grandell [11], Theorems 1.4 and 1.7). By the cross ergodic theorem (see (8.3.2) of Baccelii and Bremaud [1]), the sequences of interarrival times are ergodically stable. Thus, we can compare the input and output sequences via the stationary distributions. Chang and Pinedo [3], Theorem 3.2 provides the comparison of the one-dimensional interarrival times through intensity functions.

Proposition 3.5 *Consider two DSP processes with intensity functions, $\lambda(t)$ and $\tilde{\lambda}(t)$. Define the integral of intensity,*

$$\Lambda[t_1, t_2] \stackrel{\text{def}}{=} \int_{t_1}^{t_2} \lambda(s) ds.$$

$\tilde{\Lambda}_{in}[t_1, t_2]$ is defined similarly. If $\Lambda[0,t] \leq_c \tilde{\Lambda}[0,t]$ for all t, then the one-dimensional stationary distributions $G(x)$ and $\tilde{G}(x)$ are ordered in the variability ordering, i.e., $G(x) \leq_c \tilde{G}(x)$.

The detailed proof of Proposition 3.5 is given in Chang and Pinedo [3]. Intuitively, one can argue as follows: the expected interarrival times should be the same as the inverse of the average arrival rate, i.e., $\int_0^\infty x dG(x) = \int_0^\infty x d\tilde{G}(x) = \frac{1}{\lambda}$. The distribution of the first arrival of a stationary point *process* should be the same as that of the residual life of the stationary *sequence*. Thus, for any $a \geq 0$,

$$\lambda \int_a^\infty (1 - G(x)) dx = P(T_1 \geq a),$$

where T_1 is the first arrival epoch of the stationary point *process*. This equation is known as the inversion formula which we will use later in the proof for the leaky-bucket scheme. Since $P(T_1 \geq a)$ is the probability that there is no arrival between $[0, a]$. This probability is equal to $Ee^{-\Lambda[0,a]}$ for a DSP process with intensity $\lambda(t)$. The variability ordering between $\Lambda[0,a]$ and $\tilde{\Lambda}[0,a]$ implies $Ee^{-\Lambda[0,a]} \leq Ee^{-\tilde{\Lambda}[0,a]}$ since e^{-x} is convex. The inequality, $\int_a^\infty (1 - G(x)) dx \leq \int_a^\infty (1 - \tilde{G}(x)) dx$, implies the variability ordering between $G(x)$ and $\tilde{G}(x)$.

Proof of Theorem 3.1. Define similarly $\Lambda_{\text{in}}[t_1, t_2]$ and and $\Lambda_{\text{out}}[t_1, t_2]$ as the integral of the intensities, $\lambda_{\text{in}}(t)$ and $\lambda_{\text{out}}(t)$. According to Proposition 3.5, it suffices to show that $\Lambda_{\text{out}}[0, t] \leq_c \Lambda_{\text{in}}[0, t]$ for all t. To show the variability ordering, we first consider the case that the service time distribution of the infinite-server queue takes a finite number of values on the set $\{z_j, j = 1, \ldots, k\}$ with probabilities $\{p_j, j = 1, \ldots, k\}$. Then the input-output relation of the intensities can be rewritten as $\lambda_{\text{out}}(t) = \sum_{j=1}^k p_j \lambda_{\text{in}}(t - z_j)$. Thus, for any convex function h

and any $a \geq 0$,

$$h(\Lambda_{\text{out}}[0,a]) = h(\int_0^a \lambda_{\text{out}}(t)dt) = h(\sum_{j=1}^{k} p_j \int_0^a \lambda_{\text{in}}(t - z_j)dt)$$

$$\leq \sum_{j=1}^{k} p_j h(\Lambda_{\text{in}}[-z_j, a - z_j])$$

Since we assume that $\lambda_{\text{in}}(t)$ is stationary, $Eh(\Lambda_{\text{in}}[-z_j, a - z_j]) = Eh(\Lambda_{\text{in}}[0,a])$. The proof for the variability ordering is completed by taking expectations on both sides of the inequality. For the case that the service times have a general distribution, we can approximate the distribution by considering a sequence of distributions that take a finite number of values and then take the limit. ∎

Proof of Theorem 3.2. Assume that the period of the input intensity is p. Thus, for any z

$$\int_0^p \lambda_{\text{in}}(t - z)dt = \lambda p.$$

Since the service time distribution is uniformly between $[0,p]$, it then follows from Proposition 3.4 that

$$\lambda_{\text{out}}(t) = \int_0^p \lambda_{\text{in}}(t - s)\frac{1}{p}ds = \lambda.$$

Thus, the output process is a Poisson process. This theorem is completed by applying Proposition 3.5 and $\lambda t \leq_c \Lambda_{\text{out}}[0,t]$. ∎

Proof of Theorem 3.3. To show the monotonic property of the low-pass filters, we consider two exponential service times of the infinite-server queue with means $1/\nu_1$ and $1/\nu_2$ $(1/\nu_1 \geq 1/\nu_2)$. Let $\lambda_{\text{out}}^{(j)}(t)$, $j = 1$ and 2 be the corresponding quantities of the output intensities. Define accordingly $\Lambda_{\text{out}}^{(j)}[0,t]$, $j = 1$ and 2. It suffices to show that for any convex function h and any $a \geq 0$,

$$Eh(\Lambda_{\text{out}}^{(1)}[0,a]) \leq Eh(\Lambda_{\text{out}}^{(2)}[0,a]).$$

Let X and Y be two exponentially distributed random variables with means $1/\nu_1$ and $1/\nu_2$. Since the geometric random sum of exponential random variables is still an exponential random variable, the random variable X has the same distribution as $\sum_{j=1}^{\tau} Y_j$, where Y_j's are independent random variables with the same distribution as Y and τ is a geometric random variable with the probability distribution, $P(\tau = l) = \frac{\nu_1}{\nu_2}(1 - \frac{\nu_1}{\nu_2})^{l-1}$, $l = 1, 2, \ldots$. We denote the expectation over a set of random variables $\{X_1, X_2, \ldots\}$ by $E_{\{X_1, X_2, \ldots\}}$. With the new notation, we write $\lambda_{\text{out}}^{(1)}(t) = E_{\{X\}}\lambda_{\text{in}}(t-X)$ and $\Lambda_{\text{out}}^{(1)}[0,a] = E_{\{X\}}\Lambda_{\text{in}}[-X, a-X]$. Thus, for any convex function h and any $a \geq 0$,

$$
\begin{aligned}
& h(\Lambda_{\text{out}}^{(1)}[0,a]) \\
& = h(E_{\{X\}}\Lambda_{\text{in}}[-X, a-X]) \\
& =_{st} h(E_{\{\tau, Y_1, Y_2, \ldots\}}\Lambda_{\text{in}}[-\sum_{j=1}^{\tau} Y_j, a - \sum_{j=1}^{\tau} Y_j]) \\
& = h(\sum_{l=1}^{\infty} P(\tau = l) E_{\{Y_1,\ldots,Y_l\}}\Lambda_{\text{in}}[-\sum_{j=1}^{l} Y_j, a - \sum_{j=1}^{l} Y_j]) \\
& \leq \sum_{l=1}^{\infty} P(\tau = l) h(E_{\{Y_1,\ldots,Y_l\}}\Lambda_{\text{in}}[-\sum_{j=1}^{l} Y_j, a - \sum_{j=1}^{l} Y_j]),
\end{aligned}
\tag{2}
$$

where $=_{st}$ denotes that both sides are identically distributed. Note that

$$E_{\{Y_1,\ldots,Y_l\}}\Lambda_{\text{in}}[-\sum_{j=1}^{l} Y_j, a - \sum_{j=1}^{l} Y_j]$$

is the integral of the intensity after passing through l stages of infinite-server queues with the service time at each stage being exponentially distributed with mean $1/\nu_2$. It then follows from (the proof for) Theorem 3.1 that the output process from l stages is less variable than that from a single stage. Thus, we have

$$Eh(E_{\{Y_1,\ldots,Y_l\}}\Lambda_{\text{in}}[-\sum_{j=1}^{l} Y_j, a - \sum_{j=1}^{l} Y_j]) \leq Eh(E_{\{Y\}}\Lambda_{\text{in}}[-Y, a-Y]). \tag{3}$$

Inserting (3) into (2) and taking expectations on both sides of (2) completes the proof. ■

4 Smoothing stationary and ergodic processes via $\cdot/D/1$ queues

One of the major drawbacks of having an infinite-server queue as a filter is that such queues do not necessarily smooth out the input processes except for DSP processes. For instance, consider a deterministic process with the same interarrival times as the input process. It is known that a deterministic process is less variable than any other processes with the same rate ([12]). However, if we feed this process into an infinite-server queue with random service times, the output process becomes more variable. In this section, we consider a stationary and ergodic process as the input process and a single-server queue with fixed deterministic service times as a filter. We also assume that the service discipline for the single-server queue is First-Come First-Served (FCFS) and that the buffer of the single-server queue is infinite. The following theorem shows a monotonic result for the single-server queue with fixed deterministic service times.

Theorem 4.1 *The variability of the interdeparture sequence decreases as the fixed service time of the single-server queue increases.*

In the case that the service times in the single-server queue is zero, the output process is the same as the input process. According to the monotonic result, the single-server queue actually acts as a smoothing filter.

We note that if the input process is Poisson, then a single-server queue with New Better than Used in Expectation (NBUE) service times acts as a smoothing filter (recall that a random variable X is NBUE if $E(X-a|X>a) \leq EX$ for all $a>0$ [29], pp. 273). This is due to the well known fact ([8], II.4.113) that the interdeparture time of an $M/GI/1$ queue with an arrival rate λ and service times distributed as a random variable S is distributed as S with probability λES, and $S+Y$ with probability $1-\lambda ES$, where Y is an exponential random variable with mean $1/\lambda$. Thus, the variability of the interdeparture sequences from $M/GI/1$ queues decreases as the variability of the service times decreases. Since the departure process of an $M/M/1$ queue is still a Poisson process with the same rate and an NBUE random variable is less variable than an exponential random variable with the same mean, the variability of a Poisson process can be reduced by feeding it into an $M/GI/1$ queue with NBUE service times.

Since the variability of the interdeparture times is reduced at the cost of the delay (in Theorem 4.1), in the following we propose a smoothing algorithm that reduces the expected delay in the $G/D/1$ queue by the fixed service time, z, and has the same joint distribution as that of the departure process.

Algorithm 4.2

1. *Initially, set* $s=0$.
2. *If the next arrival epoch is larger than* s,
   - *release the arrival when it arrives;*
   - *update* $s \leftarrow$ *the arrival epoch* $+z$;
3. *otherwise (the next arrival epoch is not larger than* s*),*

- *release the arrival at time s;*
- *update $s \leftarrow s + z$.*

To analyze the algorithm, first note that z is the minimum spacing between two departure epochs. Let τ_n, τ_n^1 and s_n be the n-th arrival epoch, the n-th departure epoch, the value of s at the n-th arrival. From the algorithm, it follows that $\tau_n^1 = \max(\tau_n, s_n)$ and $s_{n+1} = \max(\tau_n, s_n) + z$. Thus, we have $s_n = \tau_{n-1}^1 + z$ and

$$\tau_n^1 = \max(\tau_n, \tau_{n-1}^1 + z). \tag{4}$$

Rewrite (4) as

$$\tau_n^1 = \max(\tau_n - z, \tau_{n-1}^1) + z. \tag{5}$$

Note that (5) is the governing equation for the departure times of a $G/D/1$ queue with fixed service times z and arrival epochs, $\{\tau_n - z, n \geq 1\}$. Thus, the departure times from the filter using the algorithm would be the same as the departure times from a $G/D/1$ queue if the arrival process is shifted forward by z. Since we assume the input process is stationary, the shifted process still has the same joint distribution. This leads to the result that the joint distribution of the interdeparture times is the same as that of a $G/D/1$ queue. Moreover, the expected delay is reduced by z.

It is natural to ask if one could design a filter that minimizes the variance of interdeparture times with a guaranteed maximum delay for each customer. Though we are not able to solve this optimization problem, we propose a "good" filter that, at least, satisfies the following criteria.

(i) The delay for every customer is not longer than a constant c.

(ii) The interdeparture sequence from the filter is less variable than the input sequence.

(iii) The filter is universal, i.e., it "works" for all (stationary and ergodic) input processes.

(iv) The output from the filter is in the same order as the input.

(v) The filter is consistent, i.e., the variability of the interdeparture sequence decreases as the delay constraint c increases.

The essential concept is to add another filter to Algorithm 4.2 such that the delay constraint is satisfied (see Figure 1). We update step 3 in Algorithm 4.2 as follows:

Algorithm 4.3

3. *otherwise (the next arrival epoch is not larger than s),*

- *If s is not larger than the arrival epoch + c, release the arrival at time s; otherwise release the arrival at the arrival epoch +c.*
- *update* $s \leftarrow s + z$.

Since Algorithm 4.3 only needs two additions and two comparisons, it can be implemented efficiently. In the following theorem, we prove that it also satisfies the five criteria.

Theorem 4.4 *Algorithm 4.3 satisfies criteria (i)-(v).*

To gain more insights into the algorithm, let τ_n, τ_n^1 and τ_n^2 be the n-th arrival epoch, the n-th departure epoch from Algorithm 4.2, and the n-th departure epoch from Algorithm 4.3. Clearly, we have

$$\tau_n^2 = \min(\tau_n^1, \tau_n + c). \tag{6}$$

Unlike the monotonic property in z in Algorithm 4.2, Algorithm 4.3 does not have such a property. Moreover, the minimum spacing between two departure epochs is not guaranteed by z in Algorithm 4.3. If z in Algorithm 4.3 is larger than the expected interarrival time, then the event $\{\tau_n^1 \geq \tau_n + c\}$ occurs almost surely for every large n. In this case, $\tau_n^2 = \tau_n + c$ and the departure process is merely the arrival process delayed by c. Thus, increasing z does not necessarily reduce the variance of interdeparture times in Algorithm 4.3. The only question left is to find z that minimizes the variance. This problem seems to be difficult since the constant z depends on the input process. In the next section, we use a curve fitting technique to obtain suboptimal solutions. The constant z computed by the suboptimal solution is close to the optimal one when comparing with simulations subject to Poisson arrivals.

The rest of this section is devoted to the proofs of Theorems 4.1 and 4.4. The proofs are based on a sample path argument and the cut criterion for the majorization ordering.

Proof of Theorem 4.1. In order to show the monotonic result, we consider the departure processes of two $G/D/1$ queues with fixed deterministic service times, z and $\tilde{z}$ ($z \geq \tilde{z}$). Let $\{\tau_n, n \geq 1\}$ be the arrival epochs of the n-th customer to both $G/D/1$ queues and $\{\tau_n^1\}$, and $\{\tilde{\tau}_n^1\}$ be the departure epochs. Let T_n^1 and $\tilde{T}_n^1$ be the n-th interdeparture times of these two systems, i.e., $T_n^1 \stackrel{\text{def}}{=} \tau_n^1 - \tau_{n-1}^1$ and $\tilde{T}_n^1 \stackrel{\text{def}}{=} \tilde{\tau}_n^1 - \tilde{\tau}_{n-1}^1$ for all n. It suffices to show that for any convex function h

$$\lim_{n\to\infty} \frac{1}{n}\sum_{j=2}^{n} Eh(T_j^1) \leq \lim_{n\to\infty} \frac{1}{n}\sum_{j=2}^{n} Eh(\tilde{T}_j^1). \tag{7}$$

The key idea is to couple the sample paths in such a way that the majorization ordering between T^1 and $\tilde{T}^1$ can be established.

Let $\{n_k\}$ ($n_0 \stackrel{\text{def}}{=} 0$) be the first customer that arrives in the system after $k-1$ busy periods in the $G/D/1$ queue of the system with service times z. Now we will show that for all k

$$(T^1_{n_k+1}, \ldots, T^1_{n_{k+1}}) \prec (\tilde{T}^1_{n_k+1}, \ldots, \tilde{T}^1_{n_{k+1}}).$$

We denote the amounts of work in the $G/D/1$ queues at time t as V_t and $\tilde{V}_t$. Furthermore, we assume that both $G/D/1$ queues start from empty systems. It then follows from Lindley's equations:

$$V_t = \max(0, V_{\tau_n} + z - (t - \tau_n)), \quad \text{if } \tau_n < t \leq \tau_{n+1},$$

and $V_t = 0$ if $0 \leq t \leq \tau_1$. A similar equation holds for $\tilde{V}_t$. Since $z \geq \tilde{z}$, $V_t \geq \tilde{V}_t$. Thus, $V_{\tau_{n_k}} = \tilde{V}_{\tau_{n_k}} = 0$ for all t. From the fact that the departure time is the sum of the arrival time, waiting time and service time, i.e., $\tau^1_n = \tau_n + V_{\tau_n} + z$ and $\tilde{\tau}^1_n = \tau_n + \tilde{V}_{\tau_n} + \tilde{z}$, it follows that $\tau^1_n - \tilde{\tau}^1_n \geq z - \tilde{z}$ for all n and that $\tau^1_{n_k} - \tilde{\tau}^1_{n_k} = z - \tilde{z}$ for all k. Thus, we have for all k

$$T^1_{n_{k+1}} \leq \tilde{T}^1_{n_{k+1}}, \tag{8}$$

and

$$\sum_{j=n_k+1}^{n_{k+1}} T^1_j = \sum_{j=n_k+1}^{n_{k+1}} \tilde{T}^1_j. \tag{9}$$

Observe that the interdeparture time of the $G/D/1$ queue of the system with the service times z is the fixed constant z if the server is busy, i.e., $T^1_n = z$ for all $n_k < n < n_{k+1}$. Moreover, $T^1_{n_{k+1}} \geq z$ since the n_{k+1}-th customer enters a empty queue. In conjunction with (8) and (9), either $\tilde{T}^1_j \geq T^1_j \geq z$ or $\tilde{T}^1_j \leq T^1_j \leq z$ for $j = n_k + 1, \ldots, n_{k+1}$ and $\sum_{j=n_k+1}^{n_{k+1}} T^1_j = \sum_{j=n_k+1}^{n_{k+1}} \tilde{T}^1_j$. Applying the cut criterion for the majorization ordering in Lemma 2.2 (i) yields $(T^1_{n_k+1}, \ldots, T^1_{n_{k+1}}) \prec (\tilde{T}^1_{n_k+1}, \ldots, \tilde{T}^1_{n_{k+1}})$. Using the concatenation property for the majorization ordering in Proposition 2.3(ii), we have $(T^1_2, \ldots, T^1_{n_{k+1}}) \prec (\tilde{T}^1_2, \ldots, \tilde{T}^1_{n_{k+1}})$ for all k. From Proposition 2.3 (i), it follows that

$$\frac{1}{n_{k+1}} \sum_{j=2}^{n_{k+1}} h(T^1_j) \leq \frac{1}{n_{k+1}} \sum_{j=2}^{n_{k+1}} h(\tilde{T}^1_j).$$

The convergence of the limits are just due to the property of ergodicity. ■

Proof of Theorem 4.4. From (6), it follows that the delay for every customer is not larger than the constant c. In conjunction with (4), the output is in the same order as the input taking into account that the output from Algorithm 4.2 is in the same order as the input. Since the case $c = 0$ corresponds to the case without filtering, it suffices to show that for all stationary ergodic processes, the interdeparture sequence becomes less variable when the delay constraint c is increased.

Now consider two systems using the same parameters in Algorithm 4.3 except the delay constant c and $\tilde{c}$, $c \geq \tilde{c}$. We denote the system with the delay constant c as system A and the other as system B. Again, let τ_n be the n-th arrival epoch for both systems. Since Algorithm 4.2 does not use the parameter c, the departures from Algorithm 4.2 are still the same for both systems. We denote the n-th departure from Algorithm 4.2 as τ_n^1 for both systems. Let τ_n^2 be the n-th departure epoch from Algorithm 4.3 in system A. Let T_n^2 be the n-th interdeparture time from Algorithm 4.4 in system A, i.e., $T_n^2 = \tau_n^2 - \tau_{n-1}^2$. Let $\tilde{\tau}_n^2$ and $\tilde{T}_n^2$ be the corresponding quantities for system B. From (4) and (6), we write the three governing equations for these two systems:

$$\begin{aligned}
\tau_n^1 &= \max(\tau_n, \tau_{n-1}^1 + z), \\
\tau_n^2 &= \min(\tau_n^1, \tau_n + c), \\
\tilde{\tau}_n^2 &= \min(\tau_n^1, \tau_n + \tilde{c}), \quad c \geq \tilde{c}.
\end{aligned}$$

Analogous to the proof for Theorem 4.1, let $\{n_k, k \geq 1\}$ ($n_0 \stackrel{\text{def}}{=} 0$) be the sequence of customers that satisfies $\tau_{n_k}^1 = \tau_{n_k}$ or, equivalently, $\tau_{n_k} > \tau_{n_k-1} + z$. Also let $\{m_k, k \geq 1\}$ ($m_0 \stackrel{\text{def}}{=} 0$) be the sequence of customers that satisfy $\tilde{\tau}_{m_k}^2 = \tau_{m_k}^1$ or, equivalently, $\tau_{m_k}^1 < \tau_{m_k} + \tilde{c}$. Since we assume that $c \geq \tilde{c}$, $\tau_{m_k}^2 = \tilde{\tau}_{m_k}^2 = \tau_{m_k}^1$. Analogous to the proof of Theorem 4.1, we would like to show that for all k

$$(T_{m_k+1}^2, \ldots, T_{m_{k+1}}^2) \prec (\tilde{T}_{m_k+1}^2, \ldots, \tilde{T}_{m_{k+1}}^2)$$

via the cut criterion for the majorization ordering. Clearly, we have

$$\sum_{j=m_k+1}^{m_{k+1}} T_j^2 = \sum_{j=m_k+1}^{m_{k+1}} \tilde{T}_j^2 = \tau_{m_{k+1}}^1 - \tau_{m_k}^1. \tag{10}$$

Thus, we only need to show (c1) in the cut criterion, i.e., either $\tilde{T}_j^2 \geq T_j^2 \geq z$ or $\tilde{T}_j^2 \leq T_j^2 \leq z$ is satisfied for $j = m_k+1, \ldots, m_{k+1}$. Since $\tau_{n_k}^1 = \tau_{n_k} \leq \tau_{n_k}+c$, the sequence $\{n_k\}$ is a subsequence of the sequence $\{m_k\}$. From (4), it follows that $\tau_{j+1}^1 = \tau_j^1 + z$ for $j \neq n_k$ and $\tau_{n_k}^1 \geq \tau_{n_k-1}^1 + z$.

Using the fact that $\{n_k\}$ is a subsequence of $\{m_k\}$,

$$\tau^1_{j+1} = \tau^1_j + z \quad \text{for } j \neq m_k \tag{11}$$

and

$$\tau^1_{m_k} \geq \tau^1_{m_k-1} + z. \tag{12}$$

Moreover, according to the definition of m_k, we also know that $\tilde{\tau}^2_j = \tau_j + \tilde{c}$ for $j \neq m_k$. Thus,

$$\tilde{T}^2_j = \tau_j - \tau_{j-1} \tag{13}$$

for $j = m_k + 1, \ldots, m_{k+1} - 1$. In order to compare T^2_j and $\tilde{T}^2_j$ for $j = m_k + 1, \ldots, m_{k+1} - 1$, we consider the following four cases:

Case 1. $\tau^1_j < \tau_j + c$ and $\tau^1_{j-1} < \tau_{j-1} + c$.

In this case, $\tau^2_j = \tau^1_j$ and $\tau^2_{j-1} = \tau^1_{j-1}$. It then follows from (11) that $T^2_j = \tau^1_j - \tau^1_{j-1} = z$. Thus, we have either $z = T^2_j \leq \tilde{T}^2_j$ or $z = T^2_j \geq \tilde{T}^2_j$ in the case.

Case 2. $\tau^1_j < \tau_j + c$ and $\tau^1_{j-1} \geq \tau_{j-1} + c$.

In this case, $\tau^2_j = \tau^1_j$ and $\tau^2_{j-1} = \tau_{j-1}+c$. It then follows from (13) that $T^2_j = \tau^1_j-(\tau_{j-1}+c) \leq \tau_j + c - (\tau_{j-1}+c) = \tilde{T}^2_j$. Moreover, we have from (11) that $T^2_j = \tau^1_j - (\tau_{j-1}+c) \geq \tau^1_j - \tau^1_{j-1} = z$. Thus, we have $z \leq T^2_j \leq \tilde{T}^2_j$ in the case.

Case 3. $\tau^1_j \geq \tau_j + c$ and $\tau^1_{j-1} \geq \tau_{j-1} + c$.

In this case, $\tau^2_j = \tau_j + c$ and $\tau^2_{j-1} = \tau_{j-1} + c$. It then follows from (13) that $T^2_j = \tau_j + c - (\tau_{j-1} + c) = \tilde{T}^2_j$. Thus, we have either $z \leq T^2_j = \tilde{T}^2_j$ or $z \geq T^2_j = \tilde{T}^2_j$ in the case.

Case 4. $\tau^1_j \geq \tau_j + c$ and $\tau^1_{j-1} < \tau_{j-1} + c$.

In this case, $\tau^2_j = \tau_j + c$ and $\tau^2_{j-1} = \tau^1_{j-1}$. It then follows from (13) that $T^2_j = \tau_j + c - \tau^1_{j-1} \geq \tau_j + c - (\tau_{j-1} + c) = \tilde{T}^2_j$. Moreover, we have from (11) that $T^2_j = \tau_j + c - \tau^1_{j-1} \leq \tau^1_j - \tau^1_{j-1} = z$. Thus, we have $z \geq T^2_j \geq \tilde{T}^2_j$ in the case.

For the comparison of $T^2_{m_{k+1}}$ and $\tilde{T}^2_{m_{k+1}}$, we first note that

$$\begin{aligned} T^2_{m_{k+1}} &= \tau^2_{m_{k+1}} - \tau^2_{m_{k+1}-1} \\ &= \tau^1_{m_{k+1}} - \min(\tau^1_{m_{k+1}-1}, \tau_{m_{k+1}-1} + c) \\ &\geq \tau^1_{m_{k+1}} - \tau^1_{m_{k+1}-1} \geq z \end{aligned}$$

taking account of (6) and (12). Thus, we only need to show that $T^2_{m_{k+1}} \leq \tilde{T}^2_{m_{k+1}}$. From $\tau^2_n = \min(\tau^1_n, \tau_n + c)$, $\tilde{\tau}^2_n = \min(\tau^1_n, \tau_n + \tilde{c})$ and $c \geq \tilde{c}$, it follows that $\tau^2_n \geq \tilde{\tau}^2_n$ for all n. Since $\tau^2_{m_{k+1}} = \tilde{\tau}^2_{m_{k+1}} = \tau^1_{m_{k+1}}$, we have $T^2_{m_{k+1}} \leq \tilde{T}^2_{m_{k+1}}$. Thus, the cut criterion for the majorization ordering in Lemma 2.2 (i) is satisfied and we have that $(T^2_{m_k+1}, \ldots, T^2_{m_{k+1}}) \prec (\tilde{T}^2_{m_k+1}, \ldots, \tilde{T}^2_{m_{k+1}})$. Using the concatenation property for the majorization ordering in Proposition 2.3(ii), we have $(T^2_2, \ldots, T^2_{m_{k+1}}) \prec (\tilde{T}^2_2, \ldots, \tilde{T}^2_{m_{k+1}})$ for all k. The rest of the proof follows the same argument as in the proof of Theorem 4.1. ■

5 Applications

In this section, we consider two examples and show that the throughput in these two examples can be increased if smoothing filters are added.

5.1 Single-server loss systems

Now we consider a G/M/1 loss system with arbitrary arrivals and independent exponential service times with mean μ^{-1}. Since there is no buffer for customers, a customer when he arrives is either accepted and served immediately if he finds the system empty or lost if he finds the system busy.

The single-server loss system might be used for modelling a wireless communication network in a building ([6]). We assume that the network is connected by a built-in token ring and that an adapter, which is attached to the ring, is built in each room. The communication link between a lap-top terminal and an adapter is through an infrared link. Since a large number of adapters needs to be built, one must keep the cost of the adapter low. We might assume that the buffer size in each adapter is only for one (incoming) packet. The packet in the source adapter is transferred to the destination adapter when the token arrives at the source adapter. For an active terminal, the system might be modelled as a single-server loss system by considering the cycle time of the token ring network as the service time of the customer in the loss system. Since we assume that the adapters are built at low cost, there is no feedback mechanism between the adapter and the active terminal. The objective is to smooth the traffic generated by the terminal such that the throughput between the adapter and the terminal can be increased.

Using Rolski [25], Remark 2.14, Chang and Pinedo [3] states the following result for the loss probability of a $G/M/1$ loss system.

Proposition 5.1 *If the interarrival sequence, $T = \{T_1, T_2, \ldots\}$, for a $G/M/1$ loss system is ergodically stable, then the loss probability is given by*

$$P_l = \int_0^\infty G(x)\mu e^{-\mu x} dx = \int_0^\infty e^{-\mu x} dG(x) \ ,$$

where $G(x)$ is the one-dimensional stationary distribution of the interarrival times T.

The argument for Proposition 5.1 is as follows: Based on the memoryless property of exponential distributions, an arrival starts an exponentially distributed busy period no matter whether he is accepted or not. The probability that the next customer is lost is just the probability that the interarrival time is shorter than the exponential busy period. Taking the limit and using Definition 2.5 completes the argument.

Since $e^{-\mu x}$ is convex, Chang and Pinedo [3] uses Proposition 5.1 to compare two $G/M/1$ loss systems with different arrival processes.

Proposition 5.2 *If T and $\tilde{T}$ are two ergodically stable input sequences of G/M/1 loss systems with the one-dimensional stationary distributions of the interarrival times $G(x)$ and $\tilde{G}(x)$, then $G(x) \leq_c \tilde{G}(x)$ implies $P_l \leq \tilde{P}_l$.*

Since the throughput of the system is $1 - P_l$, the throughput is increased if the variability of the input sequence of a single-server loss system is reduced. Thus, we can apply the smoothing filters in the previous sections to increase the throughput of single-server loss systems at the cost of increasing delay.

Now we provides some numerical examples to shed some light on the trade-off between delay and throughput. Let the input process be Poisson with rate λ. Let the service times at the single-server loss system be exponentially distributed with μ. The filter in front of the loss system runs Algorithm 4.2. We note that the case $z = 0$ corresponds to the system without filtering. Since the interdeparture times from Algorithm 4.2 are stochastically identical to those from an $M/D/1$ queue, the interdeparture time is z with probability λz and $z + Y$ with probability $1 - \lambda z$, where Y is an exponential random variable with mean $1/\lambda$ ([8], II.4.113). Using Proposition 5.1, we can compute the loss probability as follows:

$$P_l = e^{-\mu z}\lambda z + (1 - \lambda z)e^{-\mu z}\frac{\lambda}{\lambda + \mu} \quad \lambda z < 1. \tag{14}$$

The equation shows two extreme cases. In the case $z = 0$, the loss probability is $\frac{\lambda}{\lambda+\mu}$. In the case $z \to \frac{1}{\lambda}$, the loss probability is $e^{-\frac{\mu}{\lambda}}$. The gain of using the filter is to have the loss probability decrease *exponentially* in the service rate μ instead of decreasing *inversely*. To compute the expected delay for a randomly chosen customer who passes through the loss system, we observe that the interdeparture time from the filter between the previous customer and this randomly chosen one is either the constant z or the sum of the constant z and an exponentially distributed random variable with mean $1/\lambda$. These correspond to the cases when the randomly chosen customer sees a *busy* or *empty* "queue" at the filter. The probability for the first case is $\frac{\lambda z(1-e^{-\mu z})}{1-P_l}$. The expected delay in the filter in the first case is the same as the expected waiting time for a customer in a $M/D/1$ queue given that he sees a busy queue when he arrives. Using the Pollaczek-Khintchine formula and conditioning on the probability that a customer sees a busy $M/D/1$ queue, the conditioned delay is $\frac{z}{2}\frac{1}{1-\lambda z}$. In the second case, the delay in the filter is zero. Thus, the expected delay for a randomly chosen customer who passes through the system is

$$\frac{\lambda z(1-e^{-\mu z})}{1-P_l}\frac{z}{2}\frac{1}{1-\lambda z}. \tag{15}$$

Using (14) and (15), we plot two examples: (i) $\lambda = \mu = 1$ and (ii) $\lambda = 1$ and $\mu = 10$. These are shown in Figures 2-5. In the first case, the throughput, $1 - P_l$, is increased from 50% to 55% by using $z = 0.53$. The expected delay is merely 0.22. The second case is even better. The throughput is increased from 91% to 99% by using $z = 0.38$. The expected delay is 0.11.

In order to achieve a guaranteed maximum delay c, we apply Algorithm 4.3. Since the minimum and maximum functions in Algorithm 4.3 are nonlinear, we believe that the output process is analytically intractable. We use simulations to measure the performance of this algorithm. In Figure 6, we show the results for the throughput for the second example above, i.e., $\lambda = 1$ and $\mu = 10$ with z in the x-axis and the loss probability in the y-axis. The eight curves correspond to $c = 0.125, 0.25, 0.375, 0.5, 0.625, 0.75, 0.875$ and 1.0. We use the regenerative method to generate the confidence interval for the simulations [21] and halt the simulation when the ratio of the width of the 99% confidence interval to the estimated loss probability is less than 0.005. A curve fitting technique, based on an approximation of the distribution of the delay through the filter, is used to find the optimal z's in these curves. Though Algorithm 4.3 does not have a monotone property in z, one still expects that the variability of the output process decreases in z when a large proportion of customers pass through the filter without being affected by the delay constraint, i.e., $\tau_n^2 = \tau_n^1$ for most n. Thus, the constant z should be chosen when the monotonicity property begins to fail. Assume that the monotonicity property begins to fail when δ% of customers are affected by the delay constraint. Thus, the constant z is obtained by solving $P(\tau_n^1 - \tau_n > c) \leq \frac{\delta}{100}$. Note that $\tau_n^1 - \tau_n$ is the waiting time of a $M/D/1$ queue with arrival rate λ and service times z. In order to obtain a closed-form expression, we approximate the waiting time of the $M/D/1$ queue by an exponential distribution with the same mean. Note that the expected waiting time of a $M/D/1$ queue with arrival rate λ and service times z is $\frac{\lambda z^2}{2}(1-\lambda z)^{-1}$. Thus, we need to solve z in the following equation:

$$P(\tau_n^1 - \tau_n > c) \approx e^{-\frac{2}{\lambda z^2}(1-\lambda z)c} = \frac{\delta}{100}.$$

This leads to

$$z = \frac{\sqrt{(c\lambda)^2 + 2c\lambda \log(\frac{100}{\delta})} - c\lambda}{\lambda \log(\frac{100}{\delta})}. \tag{16}$$

Choosing $\delta = 1$, we mark in Figure 6 the constant z suggested by (16). We note that these marks are close to the optimal z in our simulation. This suggests that Algorithm 4.3 should be operated in the way that most of customers are not affected by the delay constraint c.

5.2 The leaky-bucket scheme

The leaky-bucket scheme proposed by Turner [35] has been widely studied in the input rate control of B-ISDN/ATM networks (see [31, 30] and references therein). We assume that the input process is stationary and ergodic and that the token is coming at every d seconds to a token pool with buffer size 1. At the instant of an arrival, the packet is admitted to the network if there is a token in the token pool. After the packet is admitted to the network, the number of tokens in the pool is then reduced by 1. If the token pool is empty when a packet arrives, this packet is lost. An arriving token is lost if the token pool is full and it is put into the token pool if the token pool is empty. We compute the throughput in the following theorem.

Theorem 5.3 *If the input process $N(t)$ is stationary and ergodic with rate λ and the one-dimensional stationary distribution of interarrival times $G(x)$, then the loss probability*

$$P_l = \frac{\lambda d - 1 + \lambda \int_d^\infty (1 - G(x))dx}{\lambda d}.$$

Proof. Let N_j be the number of packets that arrives in the interval $[(j-1)d, jd)$. Since the token arrives every d seconds, the number of packets lost within $[(j-1)d, jd)$ is $(N_j - 1)^+$, where $x^+ = max(0, x)$. Thus, the loss probability is

$$P_l = \lim_{n\to\infty} \frac{\sum_{j=1}^n (N_j - 1)^+}{\sum_{j=1}^n N_j}.$$

By the ergodicity and stationarity of the input process,

$$P_l = \frac{E(N_1 - 1)^+}{EN_1}.$$

Observe that $EN_1 = \lambda d$ and that $E(N_1 - 1)^+ = E(N_1) - 1 + P(N_1 = 0)$. From the inversion formula (see (4.2.4b) of Baccelli and Bremaud [1] or the Palm-Khintchine equations of Cox and Isham [9]), it follows that

$$P(N_1 = 0) = \lambda \int_d^\infty (1 - G(x))dx.$$

This completes the proof. ∎

Using the property that $\int_a^\infty (1 - G(x))dx \leq \int_a^\infty (1 - \tilde{G}(x))dx$ if $G(x) \leq_c \tilde{G}(x)$, we have the following corollary for comparing two input sequences for the leaky-bucket scheme with token buffer size 1.

Corollary 5.4 *If T and $\tilde{T}$ are two ergodically stable input sequences with the one-dimensional stationary distributions of interarrival times $G(x)$ and $\tilde{G}(x)$, then $G(x) \leq_c \tilde{G}(x)$ implies $P_l \leq \tilde{P}_l$.*

In other words, the throughput is increasing as the variability of the input sequence decreases. Similar to the conclusion in the single-server loss systems, we can apply the smoothing filters in the previous sections to increase the throughput of the leaky-bucket scheme with token buffer size 1.

6 Conclusions and future research

In this paper, we provided various filters for Double Stochastic Poisson processes and stationary and ergodic processes for reducing the variability of the one-dimensional distributions of interarrival times (subject to a maximum delay constraint). We also showed that the throughput of the single-server loss system and the leaky-bucket scheme with token buffer size 1 can be increased by smoothing the input processes since the throughput of these two systems only depend on the one-dimensional distributions of interarrival times. In general, we expect that one might be able to increase the throughput of a system with a finite capacity by smoothing the input processes. For instance, the buffer sizes of the two examples in section 6 is larger than 1. The proof for such systems appears to be difficult since the throughput depends not only on the one-dimensional distributions of interarrival times but also on the *joint* distributions. The conjecture is related to Ross's conjecture ([28, 10, 15, 24, 25, 26, 33, 34, 4, 5], among many others) and Niu's conjecture ([23, 16]) in the literature. Some of them remain unsolved ([27]).

Acknowledgments: The author would like to thank Kuang-Cheng Chen and Gee-Gao Mei for discussing the applicability of communication networks in Section 5.1, and Randolph Nelson and Khosrow Sohraby for discussing the leaky bucket scheme in Section 5.2, and Jim Kurose for suggesting the resequencing algorithm for infinite-server queues. The author would also like to thank Randolph Nelson for his careful comments which greatly enhanced the presentation of this work.

References

[1] F. Baccelli and P. Bremaud, *Palm Probabilities and Stationary Queues.* New York: Springer-Verlag, 1987.

[2] P. Bremaud, *Point Processes and Queues, Martingale Dynamics.* New York: Springer-Verlag, 1981.

[3] C.S. Chang and M. Pinedo, "Bounds and Inequalities for Single Server Loss Systems," *Queueing Systems* 6, pp. 425-436, 1990.

[4] C.S. Chang, X.L. Chao and M. Pinedo, "Integration of Discrete-Time Correlated Markov Processes in a TDM System: Structural Results," *Probability in the Engineering and Informational Sciences*, vol. 4, pp. 29-56, 1990.

[5] C.S. Chang, X.L. Chao and M. Pinedo, "Monotonicity Results for Queues with Doubly Stochastic Poisson Arrivals: Ross's Conjecture," *Adv. Appl. Prob*, vol. 23, pp. 210-228, 1991.

[6] K.C. Chen, IBM T.J. Watson Research Center, personal communication.

[7] Y.-S. Chow and H. Teicher, *Probability Theory, Independence, Interchangeability, Martingales*, New York: Springer-Verlag, 1988.

[8] J.W. Cohen, *The Single Server Queue.* New York: North-Holland, 1981.

[9] D.R. Cox and V. Isham, *Point Processes.* New York: Chapman and Hall, 1980.

[10] S. Fond and S.M. Ross, "A Heterogeneous Arrival and Service Queueing Loss Model," *Naval Research Logistics Quarterly*, Vol. 25, 483-488, 1978.

[11] J. Grandell, *Doubly Stochastic Poisson Processes.* New York: Springer-Verlag, 1976.

[12] B. Hajek, "The Proof of a Folk Theorem on Queueing Delay with Applications to Routing Networks," *J.A.C.M.* 30, pp. 834-851, 1983.

[13] G.H. Hardy, J.E. Littlewood and G. Polya, *Inequalities.* Cambridge Univ. Press, London and New York, 1952.

[14] H. Heffes and D.M. Lucantoni, "A Markov-modulated Characterization of Packetized Voice and Data Traffic and Related Statistical Multiplexer Performance," *IEEE Journal of Selected Areas in Communication* SAC-4(6), pp. 856-867, 1986.

[15] D.P. Heyman, "On Ross's Conjecture about Queues with Nonstationary Poisson Arrivals," *J. Appl. Prob.* 19, 245-249, 1982.

[16] B.S. Greenberg, "On Niu's Conjecture for Tandem Queues," *Adv. Appl. Prob.*, Vol. 19, pp. 751-753, 1987.

[17] J. Karamata, "Sur une inégalité rélative aux fonctions convexes," *Publ. Math. Univ. Belgrade* 1, 145-148, 1932.

[18] S. Karlin and A. Novikoff, "Generalized Convex Inequalities," *Pacific J. Math.* 13, 1251-1279, 1963.

[19] J. Keilson and L.D. Servi, *Networks of Non-homogeneous $M/G/\infty$ Systems*, submitted to *Performance Evaluation*, 1989.

[20] J. F. Kurose, IBM T.J. Watson Research Center, personal communication.

[21] S.S. Lavenberg, "Computer Performance Modeling Handbook," *Academic Press*, 1983.

[22] A.W. Marshall and I. Olkin, *Inequalities: Theory of Majorization and Its Applications.* New York: Academic Press, 1979.

[23] S.-C. Niu, "Bounds for the Expected Delays in Some Tandem Queues," *J. Appl. Prob.*, vol. 17, pp. 831-838, 1980.

[24] T. Rolski, "Queues with Non-stationary Input Stream: Ross's Conjecture," *Adv. Appl. Prob.*, vol. 13, pp. 603-618, 1981.

[25] T. Rolski, "Comparison Theorems for Queues with Dependent Inter-Arrival Times," Modelling and Performance Evaluation Proceeding of the International Seminar, Paris, France (Jan 24-26, 1982) 42-67.

[26] T. Rolski, "Upper Bounds for Single Server Queues with Doubly Stochastic Poisson Arrivals," *Math. of Oper. Res.*, vol. 11, pp. 442-450, 1986.

[27] T. Rolski, "Queues with Nonstationary Arrivals," *Queueing Systems* 5, pp. 113-130, 1989.

[28] S.M. Ross, "Average Delay in Queues with Non-stationary Poisson Arrivals," *J. Appl. Prob.* 15, 602-609, 1978.

[29] S.M. Ross, *Stochastic Processes.* New York: J. Wiley & Sons, 1983.

[30] M. Sidi, W.Z. Liu, I. Cidon and I. Gopal, "Congestion Control Through Input Rate Regulation," *GLOBECOM'89*, pp. 1764-1768, 1989.

[31] K. Sohraby, "On the Performance of Bursty and Correlated Sources Subject to Leaky Bucket Rate-Based Access Control Schemes," *IBM RC-16116*, 1990.

[32] D. Stoyan, *Comparison Methods for Queues and Other Stochastic Models*, Berlin: J. Wiley & Sons, 1983.

[33] A. Svoronos and L. Green, "The N-Seasons S-Servers Loss System," *Naval Research Logistics* 34, 579-591, 1987.

[34] A. Svoronos and L. Green, "A Convexity Result for Single Server Exponential Loss Systems with Nonstationary Arrivals," *J. Appl. Prob.* 25, 224-227, 1988.

[35] J.S. Turner, "New Directions in Communications (or Which Way to the Information Age?)," *IEEE Communication Magazine*, Vol. 24, No. 10, Oct. 1986.

[36] D. Vere-Jones, "Some Applications of Probability Generating Functions to the Study of Input-Output Streams," *Journal of the Royal Statistical Society, Series B* 30, pp. 321-333, 1968.