PERFORMANCE DECAY IN A SINGLE SERVER EXPONENTIAL QUEUEING MODEL WITH LONG RANGE DEPENDENCE

Sidney Resnick and Gennady Samorodnitsky

Cornell University

Abstract. We discuss how long range dependence can influence the characteristics of a single server queue. We take the analogue of the G/M/1 queue except that the input stream is altered to exhibit long range dependence. The equilibrium queue size and equilibrium waiting time distributions each have heavy tails. By suitably selecting the parameters of the inputs, the queue size or waiting time can be made to possess infinite variance and even infinite mean. Some simulations dramatically illustrate the potential for undetected long range dependence to significantly alter the queueing behavior compared to what is anticipated with traditional inputs.

1. Introduction.

Long range dependence is a property of stationary time series models whose current state has a strong dependency on the remote past. Definitions vary from author to author but a commonly accepted definition in covariance stationary time series is that a process $\{X_n\}$ has long range dependence if

$$\sum_{j=1}^{\infty} |\mathrm{corr}(X_0, X_j)| = \infty$$

(cf. Brockwell and Davis, 1991). Here $corr(\cdot,\cdot)$ is the correlation coefficient, whose use is legitimate in the finite variance case. A very useful recent reference on long range dependence is Beran, 1992. In contexts outside traditional times series where correlations may not exist or be difficult to compute or to interpret, other measures of dependence may be more meaningful. In this case, long range dependence may refer to a slowly decreasing dependence between blocks of random variables as the time gap between the blocks grows. In our work, one interpretation of long range dependence in a stationary process $\{T_n\}$ is that the probability

$$P[T_i \le r, i = 1, \dots, n]$$

decreases slowly to 0 as $n \uparrow \infty$ for a suitable choice of constant r.

Statistical evidence is mounting that traffic on certain types of data networks may exhibit long range dependence. For instance, Beran et. al (1992) report on the analysis of several data sets representing the traffic seen as a result of transmitting video conference scenes. The data sets are large, sometimes in excess of 50,000 data points and consequently one may expect estimates of the autocorrelation function to be accurate to very large lags. Beran et. al. conclude that long range dependence is likely to be present in the underlying models which gave rise to many of these data sets. A time series plot of a sample video conference data set

1980 *Mathematics Subject Classification* (1985 *Revision*). 60K25, 62M10.

Key words and phrases. long range dependence, heavy tails, Poisson process, G/M/1 queue.

Sidney Resnick and Gennady Samorodnitsky were supported by NSF Grant DMS-94-00535 at Cornell University. Early work was supported by NSA Grant 92G–116. Gennady Samorodnitsky received some support from the United States-Israel Binational Science Foundation (BSF) Grant No. 92-00074/1 .

of size 48497 is given in Figure 1 and under it is a plot of the sample autocorrelation function to lag 1000. Notice the sample autocorrelation function does not go to zero rapidly as would be characteristic of short memory models. We remark that it is a matter of further research to understand the exact relationship between the long range dependence understood in terms of interarrival times (the point of view pursued in the present paper) and long range dependence between the numbers of packet/cells arriving per time unit, observed in the statistical studies of traffic measurements, even though on the intuitive level this relationship seems to be significant.

The presence of long range dependence is difficult to prove statistically. Indeed, several of the data sets analyzed in Beran et. al. (1992) have been previously analyzed by conventional time series means (cf. Heyman et. al. (1992)) and fitted by such garden variety time series models as a stationary autoregression of order 2. Although the autoregressive polynomial roots of such models are typically near the unit circle, which is necessitated by the requirement of modelling a slowly decreasing sample autocorrelation function, such classical time series models do not exhibit long range dependence. The fact that long range dependence is difficult to distinguish from time series models with autoregressive polynomial roots near the unit circle coupled with persistent engineering skepticism about whether long range dependence would materially affect system performance assuming it was indeed present has led us to inquire if the effects of long range dependence can be quantified at least in single server exponential models.

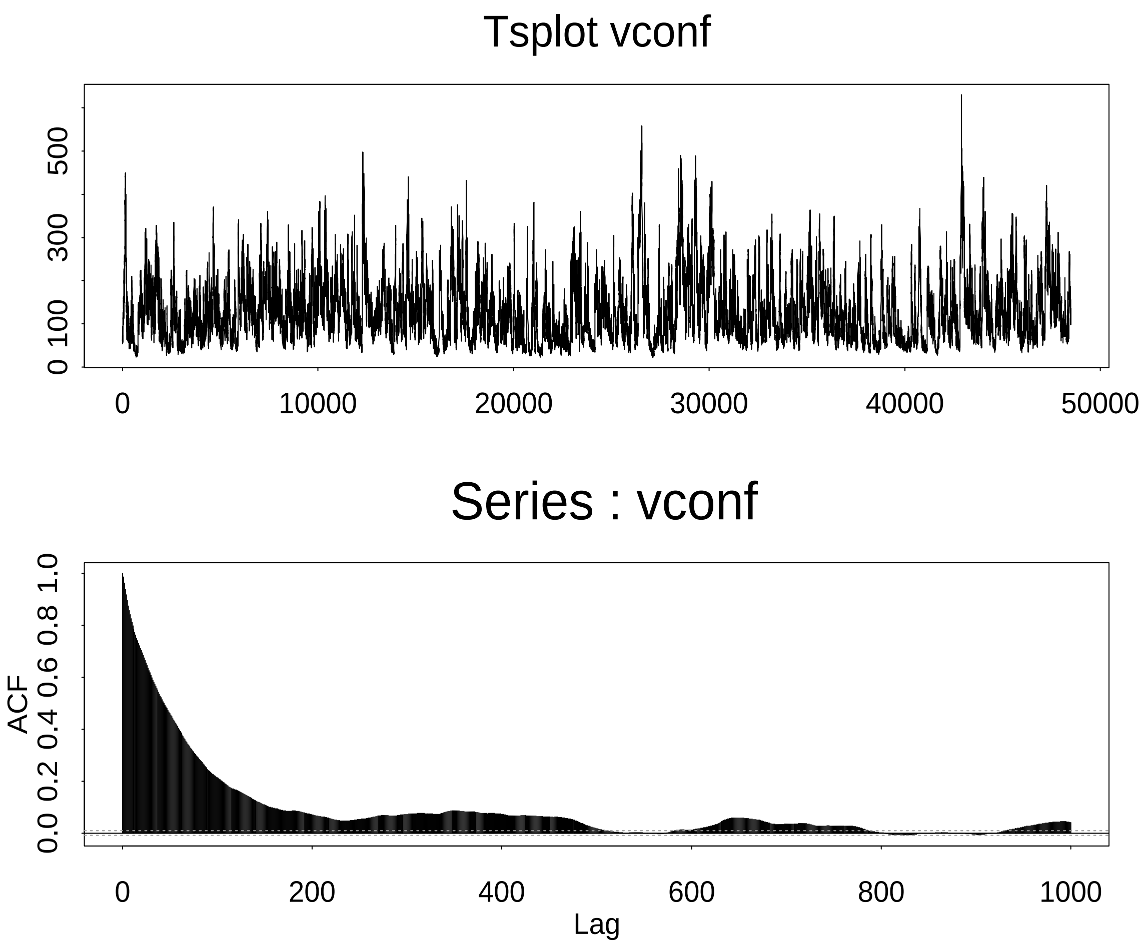

Figure 1.1

One of the difficulties in studying queueing systems with serially correlated input is that there are no commonly accepted paradigms for arrival processes. Specifying the autocorrelation function of a stationary sequence of interarrival times with given marginals is sufficient to fully describe the model only when this sequence is assumed to be Gaussian. Normal interarrival times are unsuitable because of the theoretical possibility of negative outcomes and also because of lack of skewness. Needless to say, the situation is even more difficult when one wants to model long range dependent interarrival times.

A recent paper of Livny, Melamed and Tsiolis (1993) studying the impact of serial correlation on queues, used two methods to simulate an input process. One is the TES method due to Melamed (1991) and studied by Jagerman and Melamed (1992a, b), while the other is the minification technique due to Lewis and McKenzie (1991) which simulates a min-moving average process of order 1 (see also Davis and Resnick, 1989, 1993 for background on such processes). Both techniques produce a sequence of uniform random variables on $(0,1)$ which can be then transformed to any given marginal distribution by the standard application of the inverse distribution function. The autocorrelation functions in both cases have one free parameter, which happens to be the same in the present context as saying that both stationary streams of uniform random variables can be parametrized by their correlations at lag 1 (say). TES and Minification methods produce very different autocorrelation functions, but both of them decrease to zero exponentially fast with the lag. For the same correlation at lag 1, the decrease is faster for the minification method and slower for the TES method.

In Section 2 we propose a simple modification of the G/M/1 queue. We assume we have a single server who serves according to a homogeneous Poisson process with rate μ. We feed this server an arrival stream modelled by a point process whose interpoint distances $\{T_n\}$ is a stationary process. Under a mild condition on the sequence $\{T_n\}$, the variable X_n representing the number in the system seen by the nth arriving customer just prior to the arrival. has a limit distribution as $n \to \infty$. We interpret this limit distribution as the equilibrium distribution and write $X_n \Rightarrow X_\infty$, where "$\Rightarrow$" denotes convergence in distribution and X_∞ represents a random variable with the equilibrium distribution.

In Section 3 we construct a class of examples where the interarrival times $\{T_n\}$ have long range dependence. For this model with a non–renewal input process, the distribution of X_∞ has a heavy right tail. Contrast this with the classical stable G/M/1 queue where the stationary queue length distribution is geometric (Asmussen, 1987, page 204). The presence of dependence can indeed have a dramatic effect on system performance and this is illustrated by the dramatic effects the long range dependence in the input has on the queue size distribution.

Section 4 presents a simulation of the $\{T_n\}$ process discussed in Section 3 and shows graphically the effect the long range dependent input process has by comparing system characteristics of a model with such an input compared with a standard model without the highly dependent input. Of course, the simultion results show us the combined effect of short and long range dependence. However, they do demonstrate how the heavy tails of the queue size, the presense of which is proven in Section 3, appear to the eye of an observer.

2. A reversible process with negative mean drift.

Consider a homogeneous Poisson process N with rate μ. We think of N as a random measure on the Borel subsets of $[0,\infty)$. For an interval I, $N(I)$ represents the number of service completions in time interval I by a single server serving at rate μ. Let $\{T_n, n \geq 1\}$ be a stationary sequence of non-negative random variables. Suppose $ET_n = \lambda^{-1}$ and set

$$S_0 = 0, \quad S_n = T_1 + \cdots + T_n, \quad n \geq 1$$

and think of S_n as the time of arrival of the nth customer. Setting $S_0 = 0$ is a convenience which could easily be eliminated if desired. We assume $\{S_n\}$ is independent of N. For convenience, set $X_0 = 0$ and define the process $\{X_n, n \geq 0\}$ by the Lindley recursion:

$$X_{n+1} = (X_n + 1 - N(S_n, S_{n+1}])^+ \,, n \geq 0, \tag{2.1}$$

where as usual $x^+ = x$, if $x \geq 0$ and $x^+ = 0$ if $x < 0$. We think of X_n as the number in the system seen by the nth arriving customer. Finally define

$$K_{n+1} = 1 - N(S_n, S_{n+1}], n \geq 0,$$

so that (2.1) can be rewritten as

$$X_{n+1} = (X_n + K_{n+1})^+, n \geq 0. \tag{2.1$'$}$$

Note that

$$EK_n = 1 - \mu ET_n = 1 - \mu\lambda^{-1},$$

and it is natural to set

$$\rho = \frac{\lambda}{\mu},$$

and think of ρ as the traffic intensity. Then

$$EK_n = 1 - \rho^{-1}$$

and

$$EK_n < 0 \text{ iff } \rho < 1.$$

Then paralleling the classical theory we have the following result.

Proposition 2.1. *If $\{T_n\}$ is a reversible, stationary, ergodic process and $\rho < 1$, we have as $n \to \infty$*

$$X_n \Rightarrow X_\infty \stackrel{d}{=} \bigvee_{j=0}^{\infty} \left(\sum_{i=1}^{j} K_i \right).$$

(The symbols $\vee$ and $\wedge$ stand, as usual, for the max and min operators and we follow the convention that $\sum_{i=1}^{0} K_i = 0$.) The limit random variable X_∞ is finite because of the assumption $EK_i < 0$.

Proof. As in the classical case (Asmussen, 1987, page 80; Resnick, 1992, page 514) induction yields

$$X_n = \max\{X_0 + \sum_{i=1}^{n} K_i, \sum_{i=2}^{n} K_i, \ldots, K_n, 0\}.$$

If the sequence $\{K_n\}$ is reversible, which means for all integers m

$$(K_1, \ldots, K_m) \stackrel{d}{=} (K_m, \ldots, K_1),$$

then with $X_0 = 0$

$$\begin{aligned} X_n &\stackrel{d}{=} \max\{\sum_{i=1}^{n} K_i, \sum_{i=1}^{n-1} K_i, \ldots, K_1, 0\} \\ &= \bigvee_{j=0}^{n} \left(\sum_{i=1}^{j} K_i \right). \end{aligned}$$

If $EK_i < 0$, then ergodicity of $\{T_n\}$ and the ergodic theorem implies $\sum_{i=1}^n K_i \sim nEK_1 \to -\infty$ and so as $n \to \infty$

$$\bigvee_{j=0}^{n} \left(\sum_{i=1}^{j} K_i \right) \uparrow \bigvee_{j=0}^{\infty} \left(\sum_{i=1}^{j} K_i \right) < \infty.$$

It only remains to easily verify that $\{T_n\}$ reversible implies $\{K_n\}$ is also reversible. We have for any m and non-negative integers $k_1, \dots, k_m$ that

$$\begin{aligned} P[K_i = 1 - k_i,\, i = 1, \dots, m] & \\ &= P[N(S_{i-1}, S_i] = k_i,\ i = 1, \dots, m] \\ &= EP[N(S_{i-1}, S_i] = k_i,\ i = 1, \dots, m | T_1, \dots, T_m] \\ &= E \prod_{i=1}^{m} \frac{e^{-\mu T_i} (\mu T_i)^{k_i}}{k_i!} \end{aligned}$$

and since $(T_1, \dots, T_m) \stackrel{d}{=} (T_m, \dots, T_1)$ we have this equal to

$$\begin{aligned} &= E \prod_{i=1}^{m} \frac{e^{-\mu T_{m-i+1}} (\mu T_{m-i+1})^{k_i}}{k_i!} \\ &= E \prod_{i=1}^{m} \frac{e^{-\mu T_i} (\mu T_i)^{k_{m-i+1}}}{k_{m-i+1}!} \\ &= P[K_i = 1 - k_{m-i+1},\ i = 1, \dots, m] \\ &= P[K_m = 1 - k_1, \dots, K_1 = 1 - k_m]. \quad \square \end{aligned}$$

In Section 3 we will analyze tail behavior of X_∞ and we will show that for positive constants c, r, it is possible to bound $P[X_\infty > cn]$ from below by probabilities of the form

$$\frac{1}{2} P[\frac{S_n}{n} < r] \geq \frac{1}{2} P[T_i \leq r, i = 1, \dots, n].$$

Thus the right tail probabilities of X_∞ can be bounded below by left tail probabilities of S_n/n which may be thought of as the probability that the sample mean is bounded away from the true mean. We will see that such probabilities can decrease to zero very slowly if the $\{T_n\}$ sequence has a long range dependence property.

3. A family of input processes with long range dependent interarrival times.

In this section we suggest a model for the input process that is parametrized by an infinite sequence of numbers and can have correlations that decrease as slowly or as quickly as one wishes. The model has a dependence structure that is fairly intuitive and is amenable to straightforward computer generation. The basic model generates stationary interarrival times with standard exponential marginals, and these can be transformed to any other desired marginals by the application of the inverse transform method.

The basic model is defined by a nonincreasing sequence of nonnegative numbers $\{a_i, i \geq 1\}$ such that $\sum_{i=1}^\infty a_i = 1$, and by a Gamma process $\{Y(t), -\infty < t < \infty\}$. This is a Lévy process, that is, a process with stationary, independent increments such that

$$E \left(e^{-\lambda (Y(t+h) - Y(t))} \right) = \left(\frac{1}{1+\lambda} \right)^h, \quad \lambda > 0,$$

so that $Y(t+h)-Y(t)$ has a Gamma density

$$\frac{e^{-x}x^{h-1}}{\Gamma(h)}, \quad x>0.$$

The basis for our long range dependent process is a stationary sequence $\{\xi_n\}$ defined as a moving average of increments of the Y process:

$$\xi_n = \sum_{j=1}^{\infty} Y(j-1-n, j-1-n+a_j] \tag{3.1}$$

where we use the notation $Y(a,b] = Y(b) - Y(a)$.

Another way to describe $\{\xi_n\}$ is as follows: For $0 \le i \le j < n$ let

$$\gamma(i,j) = \begin{cases} a_{j+1}, & \text{if } i=0,\ j=0,1,\dots,n-2, \\ \sum_{k=n}^{\infty} a_k, & \text{if } i=0,\ j=n-1, \\ a_{n-i}, & \text{if } i=1,\dots,n-1,\ j=n-1, \\ a_{j-i}-a_{j-i+1}, & \text{if } 0<i\le j<n-1. \end{cases} \tag{3.2}$$

The distributional structure of $\xi_0,\dots,\xi_{n-1}$ is then explained by the fact that

$$\xi_k \stackrel{d}{=} \sum_{i=0}^{k}\sum_{j=k}^{n-1}\Gamma(i,j),\ \ k=0,\dots,n-1, \tag{3.3}$$

where $\Gamma(i,j)$, $1 \le i \le j \le n$ are independent Gamma random variables with scale parameter 1, and shape parameters given by (3.2).

More background on such moving average processes is given in the appendix which also gives additional perspective on what contributes to dependence of lagged variables in the $\{\xi_n\}$ sequence. We observe at this point that since $(\lambda > 0)$

$$\begin{aligned} E\left(e^{-\lambda\xi_k}\right) &= \prod_{j=1}^{\infty} E\left(e^{-\lambda Y(j-1-n,j-1-n+a_j]}\right) \\ &= \prod_{j=1}^{\infty}\left(\frac{1}{1+\lambda}\right)^{a_j} = \left(\frac{1}{1+\lambda}\right)^{\sum_1^\infty a_j} = \frac{1}{1+\lambda}, \end{aligned}$$

ξ_k has a standard, exponential distribution. Furthermore, the representation (3.1) shows immediately that $\{\xi_n\}$ is stationary and ergodic. Note, further, that if a_i's go to zero slowly, then the common term $\Gamma(0,n-1)$ that contributes to each $\xi_0,\xi_1,\dots,\xi_{n-1}$ in representation (3.3) is stochastically large, and so we expect the memory of the process to be long.

We now focus on the second order properties of the process. Even though second order properties in non-normal models are not as critical as in the normal case, it is still somewhat illuminating to see how variance-covariance relations behave here. Let $\rho_n = \operatorname{corr}(\xi_0,\xi_n)$, $n \ge 0$. Note that part (i) of Proposition 3.1 shows, in particular, that ρ_n can decrease to zero as slowly as we wish, if a_n's are chosen to go to zero slowly enough.

Proposition 3.1. *(i) For the correlations $\{\rho_n\}$ of the $\{\xi_n\}$ process we have*

$$\rho_n = \sum_{i=n+1}^{\infty} a_i. \tag{3.4}$$

In particular, if

$$a_n \sim cn^{-(1+\theta)},\ n \to \infty \tag{3.5}$$

for some $\theta > 0$ and $c > 0$, then

$$\rho_n \sim c\theta^{-1} n^{-\theta},\ n \to \infty. \tag{3.6}$$

(ii) For the variance of $\xi_0 + \ldots + \xi_{n-1}$, $n \geq 1$ we have

$$\begin{aligned} \mathrm{Var}(\sum_{j=0}^{n-1} \xi_j) =& n^2 \sum_{i=n}^{\infty} a_i + \sum_{i=1}^{n-1} (n-i)^2 \Big((i+1)a_{n-i} - (i-1)a_{n-i+1} \Big) \\ =& n \left(1 + 2 \sum_{h=1}^{n-1} (1 - \frac{h}{n}) \sum_{i=h+1}^{\infty} a_i \right). \end{aligned} \tag{3.7}$$

In particular, if (3.5) holds, then as $n \to \infty$,

$$\mathrm{Var}(\sum_{j=0}^{n-1} \xi_j) \sim \begin{cases} 2cn^{2-\theta}/\theta(1-\theta)(2-\theta), & \text{if } 0 < \theta < 1, \\ 2cn \log n, & \text{if } \theta = 1, \\ n \sum_{n=1}^{\infty} (2i-1) a_i, & \text{if } \theta > 1. \end{cases} \tag{3.8}$$

Proof. (i) One can proceed directly from (3.3) or by noting that

$$(\xi_0, \xi_n) \stackrel{d}{=} (G_1 + G_2, G_1 + G_3),$$

where G_1, G_2 and G_3 are independent Gamma random variables with scale parameters equal to 1, and shape parameters being $\sum_{i=n}^{\infty} a_i$, $\sum_{i=1}^{n-1} a_i$ and $\sum_{i=1}^{n-1} a_i$ respectively, from which (3.4) follows. Under the assumption (3.5) the estimate (3.6) is an easy consequence of (3.4).

(ii) It is a standard fact (for example see Brockwell and Davis, 1991, page 219) that

$$\mathrm{Var}(\sum_{j=0}^{n-1} \xi_j) = n + 2 \sum_{i=1}^{n-1} (n-i) \rho_i.$$

Then (3.7) follows from (3.4) and (3.8) is a straighforward consequence of (3.4) and (3.5). $\square$

Observe that by (3.8) the variance of the partial sums $\sum_{j=0}^{n-1} \xi_j$ increase faster than linearly whenever, e.g., the sequence $\{a_n,\ n \geq 1\}$ satisfies (3.5) with $0 < \theta \leq 1$.

The process $\{\xi_n,\ n \geq 0\}$ defined by (3.3) has standard exponential marginals. We can transform these marginals to any desired marginal distribution by the inverse transform method. Given a distribution function F we define

$$Y_n^{(F,1)} = F^{\leftarrow}(1 - e^{-\xi_n}),\ n \geq 0 \tag{3.9}$$

and

$$Y_n^{(F,2)} = F^{\leftarrow}(e^{-\xi_n}),\ n \ge 0. \tag{3.10}$$

Since $1 - e^{-\xi_n}$ has uniform distribution on $(0,1)$, both $Y_n^{(F,1)}$ and $Y_n^{(F,2)}$ have the prescribed distribution F. Furthermore, as the processes $\{Y_n^{(F,1)},\ n \ge 0\}$ and $\{Y_n^{(F,2)},\ n \ge 0\}$ are each a pointwise monotone transformation of the process $\{\xi_n,\ n \ge 0\}$, they naturally have a dependence structure similar to that of $\{\xi_n\}$. It is also clear that these two processes have, in general, different multivariate distributions.

Recall from Section 2 that the tail properties of the stationary queue length distribution depend on the distribution of S_n/n and we now set $T_n = Y_n^{F,2},\ n \ge 1$ and discuss how dependence properties of $\{T_n\}$ influence the distribution of X_∞. It turns out that at least one of the processes defined in (3.9) and (3.10) is long-range dependent in the appropriate sense, as long as the sequence $\{a_n,\ n \ge 1\}$ decreases to zero slowly enough. This is discussed next.

Proposition 3.2. *For every $r \in \mathbb{R}$ such that $F(r) > 0$ there is a finite positive constant $c(r,F)$ such that for every $n \ge 1$*

$$\begin{aligned} P(Y_i^{(F,2)} \le r,\ i = 0,1,\dots,n-1) \ge & c(r,F)\Big(\Gamma(\sum_{j=n}^{\infty} a_j)\Big)^{-1} \\ \sim & c(r,F)\sum_{j=n}^{\infty} a_j = c(r,F)\rho_{n-1}, \quad \text{as } n \to \infty. \end{aligned} \tag{3.11}$$

Proof. Obviously, for any r with $F(r) > 0$ we have

$$P(Y_i^{(F,2)} \le r,\ i = 0,1,\dots,n-1) = P(\xi_i \ge \log(\frac{1}{F(r)}),\ i = 0,1,\dots,n-1),$$

$n = 1,2,\dots$. It follows from (3.3) that

$$\xi_i \ge \Gamma(0,n-1) = \sum_{j=0}^{\infty} Y(j, j+a_j], \quad i = 0,\dots,n-1.$$

Denoting $z_n = \sum_{k=n}^{\infty} a_k$, $\lambda = \log(\frac{1}{F(r)})$, and using (3.2) we conclude that

$$\begin{aligned} P(Y_i^{(F,2)} \le r,\ i = 0,1,\dots,n-1) &\ge P(\Gamma(0,n-1) \ge \lambda) \\ &= \int_\lambda^\infty e^{-t}\frac{t^{z_n - 1}}{\Gamma(z_n)}\,dt \ge c(r,F)\Big(\Gamma(z_n)\Big)^{-1} \end{aligned}$$

where we set

$$c(r,F) := e^{-2\lambda}(1 \wedge 2\lambda)/2,$$

from which the first part of (3.11) follows. The second part of (3.11) follows from $\Gamma(z) \sim z^{-1},\ z \to 0$. This completes the proof. □

We conclude from (3.11) that the probability on the left side of (3.11) can be made to decrease arbitrarily slowly to zero, provided a_j decreases to zero at the proper rate.

Now we recall the framework of Section 2. Let F be the distribution function of a random variable on $(0,\infty)$ with expected value λ^{-1}. Let

$$T_n = Y_n^{(F,2)},\ n \ge 1,$$

so that $\{T_n, n \geq 1\}$ is a stationary sequence of random variables with common distribution F. From the construction of the sequence $\{T_n\}$ or the general discussion in the Appendix, it is clear that $\{T_n\}$ is a reversible sequence. It is also ergodic. If we suppose that $\rho = \lambda/\mu < 1$, then Proposition 2.1 is applicable.

Assume that there is a $r < 1/\mu$ such that $F(r) > 0$ and set

$$S_0 = 0,\ S_n = T_1 + \ldots + T_n,\ n \geq 1.$$

As promised at the end of Section 2, we now bound the tail of the stationary distribution of the number X_∞ of customers in the queue seen by an arriving customer as follows:

$$\begin{aligned} P(X_\infty > (1-r\mu)k) =& P[\bigvee_{j=0}^{\infty}(j - N(0,S_j]) > (1-r\mu)k] \\ \geq& P[N(0,rk] \leq rk\mu, S_k < rk] \\ \geq& P\left[\frac{N(0,rk] - rk\mu}{\sqrt{rk\mu}} \leq 0\right] P[S_k < rk] \\ \sim& \frac{1}{2} P(S_k < rk) \\ \geq& \frac{1}{2} P[T_i \leq r,\, i = 1,\ldots,k] \\ \geq& c \sum_{j=k}^{\infty} a_j, \end{aligned} \tag{3.12}$$

where c is a finite positive constant, and we have applied (3.11). Thus

$$P(X_\infty > k) \geq c \sum_{j=\lfloor k/(1-r\mu)\rfloor}^{\infty} a_j, \tag{3.13}$$

for all k large enough.

One immediate conclusion is that, if the a_n's satisfy (3.5) with $\theta \leq 2$, then X_∞ has infinite second moment, whereas if $\theta \leq 1$, then even the mean of X_∞ is infinite! This is, of course, in a sharp contrast to what happens in the classical $G/M/1$ model.

Remarks: (i) The circumstance where the assumption that $F(r) > 0$ for some $r < 1\mu$ fails is somewhat degenerate since failure of the assumption means $F(\mu^{-1}) = 0$, in which case $P[T_1 > \mu^{-1}] = 1$ and *every* single interarrival time will be longer than the mean service time. The queue length will be bounded above by that of a $D/M/1$ queue and for such a queue, the stationary queue length distribution has tails which decay exponentially fast.

(ii) The more customers in the system, the longer a new arrival will spend in the system. So a lower bound on the tail of X_∞ as in (3.13) should imply a corresponding lower bound on the tail of the time in the system of a typical customer. Indeed, let W_n be the waiting time in the system of the nth customer arriving to the queue so that W_n satisfies

$$W_n \stackrel{d}{=} \sum_{i=1}^{X_n+1} Q_i, \quad n \geq 1, \tag{3.14}$$

where $Q_1, Q_2, \ldots$ are iid exponentially distributed random variables with common parameter μ and independent of X_n. It follows from Proposition 2.1 that

$$W_n \Rightarrow W_\infty \stackrel{d}{=} \sum_{i=1}^{X_\infty+1} Q_i, \tag{3.15}$$

where $\{Q_i\}$ are independent of X_∞. From (3.15) we get the tail behavior of W_∞:

$$\begin{aligned}
P[W_\infty > \lambda] \geq & P[X_\infty \geq \lfloor \mu\lambda \rfloor] P[\sum_{i=1}^{\lfloor \mu\lambda \rfloor + 1} Q_i > \lambda] \\
\geq & P[X_\infty \geq \lfloor \mu\lambda \rfloor] P[\sum_{i=1}^{\lfloor \mu\lambda \rfloor + 1} Q_i - \left(\frac{\lfloor \mu\lambda \rfloor + 1}{\mu}\right) > \lambda - \left(\frac{\lfloor \mu\lambda \rfloor + 1}{\mu}\right)] \\
\geq & P[X_\infty \geq \lfloor \mu\lambda \rfloor] P[\sum_{i=1}^{\lfloor \mu\lambda \rfloor + 1} Q_i - \left(\frac{\lfloor \mu\lambda \rfloor + 1}{\mu}\right) > 0]
\end{aligned}$$

and by applying the central limit theorem we get

$$\begin{aligned}
\sim & \frac{1}{2} P[X_\infty \geq \lfloor \mu\lambda \rfloor] \\
\geq & \frac{1}{2} c \sum_{j=\lfloor \lfloor \mu\lambda \rfloor / (1-r\mu) \rfloor}^{\infty} a_j,
\end{aligned}$$

as $\lambda \to \infty$.

As with the number in the system, we see that it is quite possible that W_∞ might have infinite variance or even infinite mean if $a_n \downarrow 0$ slowly enough.

(iii) It is impossible at present to precisely relate a given degree of dependence in the interarrival times with a specified heaviness of the tail of the waiting time or queue length distribution. There are several reasons for this. First, as discussed in the introduction, while it is clear how to measure the heaviness of a probability distribuion tail, it is by no means clear how to measure long range dependence. In particular, it is by no means clear what is the relationship (if any) between the often used Hurst parameter (Hurst 1951, 1955) and the "quadrant long range dependence" paradigm used here. Secondly, one simply needs to better understand the behavior of queues with a long range dependent input. The present work is but a small step in this direction.

4. Simulation results.

We simulated the behavior of the single server queue described in the previous sections when the input process of interarrival times $\{T_n,\ n \geq 1\}$ is given by (3.10). We have selected $F(x) = 1 - e^{-x},\ x \geq 0$, so that the interarrival times have marginally the standard exponential distribution. The service rate has been chosen to be $\mu = 10/7$, so that the traffic intensity in the system is

$$\rho = .7. \tag{4.1}$$

We have run the simulation with two choices of a_n's satisfying (3.5): first with

$$a_n = c(1+n)^{-2},\ n \geq 1, \tag{4.2}$$

(here the θ introduced in (3.5) is 1) and then with

$$a_n = c(1+n)^{-1.5},\ n \geq 1, \tag{4.3}$$

where $\theta = .5$. The constants c in both cases are chosen in such a way that (3.1) holds. Note that it follows from (3.12) that X_∞ has in both cases infinite mean, and that the lower bound on the probability tail of X_∞ is asymptotically bigger under (4.3) than under (4.2). For the purpose of comparison we present also simulations of the same single server queue when the input process is actually given by (3.9) with the same choice of the marginal distribution function F, and the same two choices of the interval lengths given by (4.2) and (4.3) accordingly. All the queues were simulated up to the time of the arrival of the 10,000th customer. The results are presented below. The long range dependent input process $\{T_n,\, n \geq 1\}$ has been simulated on a SPARCstation 10 using the representation (3.3). The results were then fed into the package *Sigma* (cf. Schruben, 1994) to produce the queueing characteristics. The simulation algorithm was based directly on the representation (3.3). That is, to generate n observations of the input process one has to generate $(n^2+n)/2$ Gamma random variables. Of course, a drawback of the method is that one needs to know the sample size n prior to the beginning of the simulation. We generated 10,000 observations for each value of the parameters. Each run took about 5 hours.

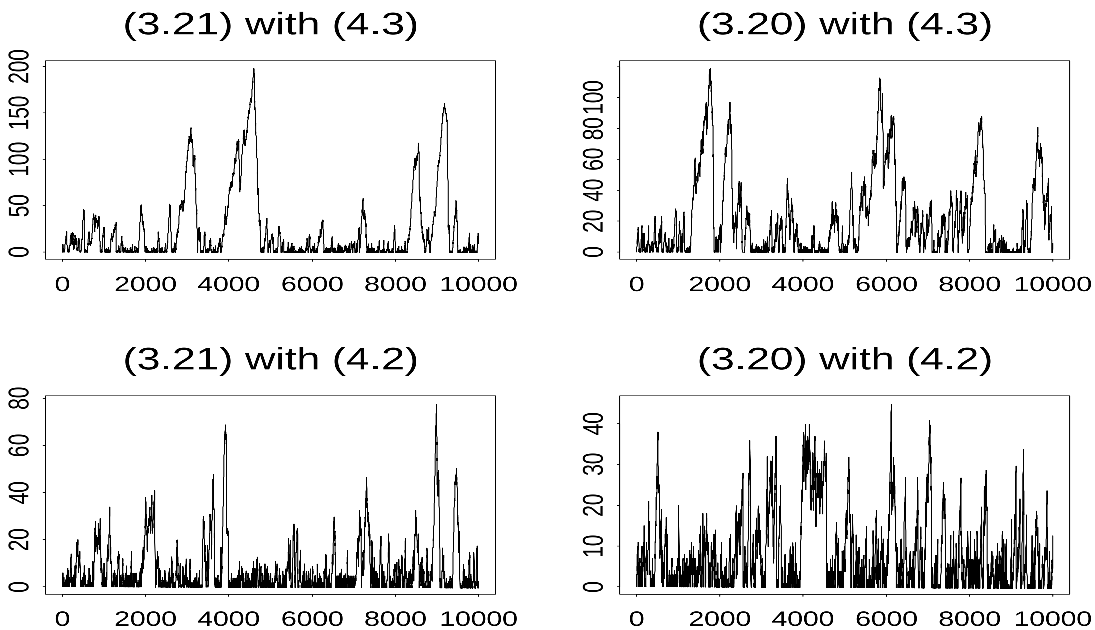

Figure 4.1: What 10,000 arriving customers see.
Dependent interarrival times

For comparison we simulated the classical $G/M/1$ (actually, $M/M/1$) queue, with the same arrival and service rates.

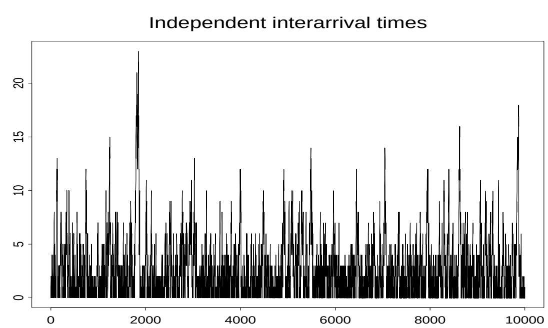

Figure 4.2: What 10,000 arriving customers see, $G/M/1$ case

Clearly, the customers in the queues with interarrival times generated by our procedure tend to see much longer lines, and the lines get longer as θ decreases (as expected from our discussion in the previous section).

Further the long range dependent processes given by (3.10) tend to create the longest lines.

These conclusions are further strengthened by looking at the histograms of the marginal distributions of the numbers of customers seen by new arrivals in the five cases considered above. The distribution tails seem to be much fatter for the long range dependent input (Figure 4.1) than for the renewal input (Figure 4.3); in the latter case the actual marginal distribution is well known to be geometric, with parameter $1 - \rho = .3$.

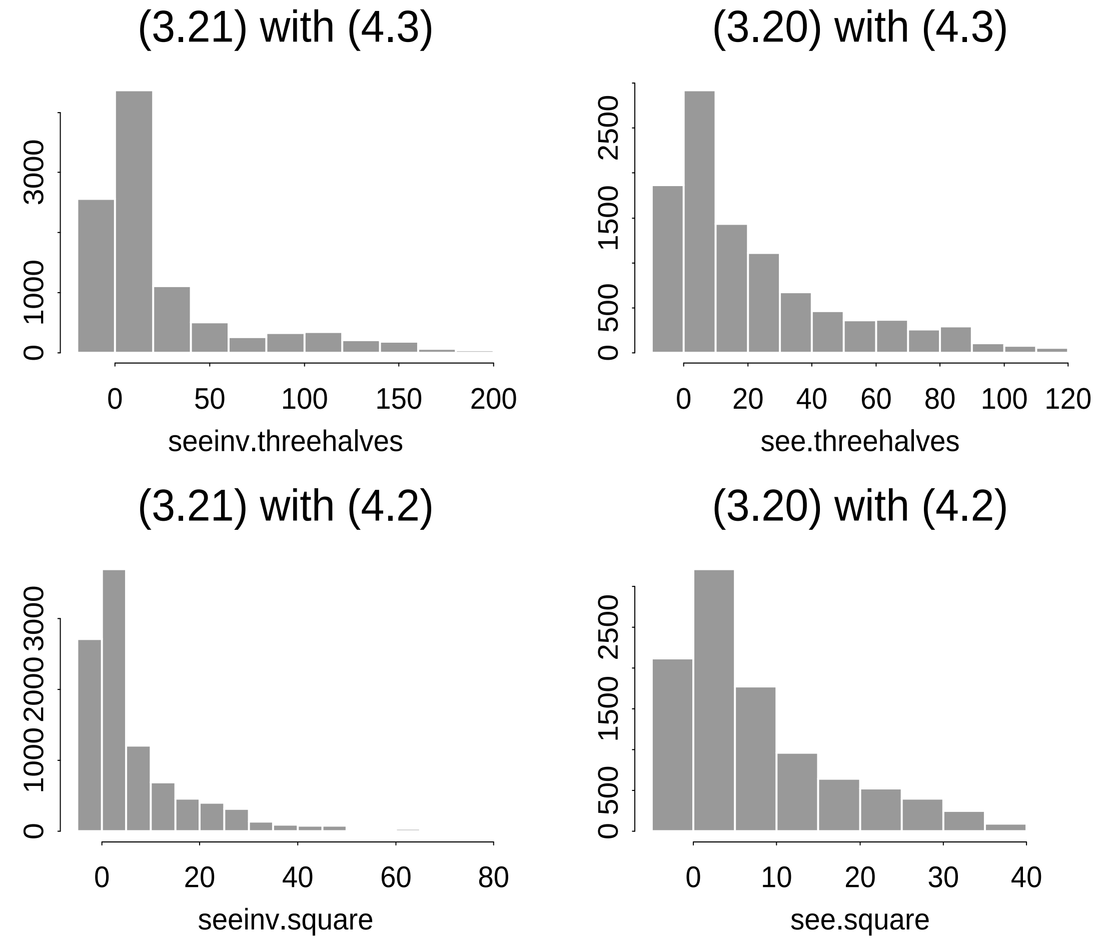

Figure 4.3: Histograms for dependent input

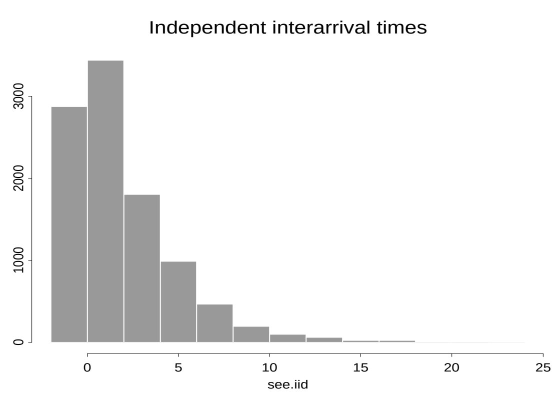

Figure 4.4: Histogram for independent input

Acknowledgement: Ranjni Natarajan generously shared with us her knowledge of computing and statistical simulation.

References

Asmussen, S., *Applied Probability and Queues*, J. Wiley and Sons, New York, 1987.

Beran, J., *Statistical methods for data with long-range dependence*, Statistical Science **7** (1992), 404–427.

Beran, J., Sherman, R., Taqqu, M., and Willinger, W., *Variable–bit–rate video traffic and long–range dependence*, Preprint (1992).

Brockwell, P. and Davis, R., *Time Series: Theory and Methods, 2nd edition*, Springer–Verlag, New York, 1991.

Cambanis, S., Lawniczak, A., Podgórski, K. and Weron, A., *Ergodicity and mixing of symmetric infinitely divisible processes.*, Center for Stochastic Processes, UNC Chapel Hill, NC **Technical Report 346** (1991).

Davis, R. and Resnick, S., *Basic properties and prediction of max–ARMA processes*, Adv. Applied Probability **21** (1989), 781–803.

Davis, R. and Resnick, S., *Prediction of stationary max–stable processes*, Annals of Applied Probability **3** (1993), 497–525.

Feller, W., *An Introduction to Probability Theory and its Applications*, Vol. II, second edition, Wiley, New York, 1971.

Heyman, D., Lakshman, T., and Tabatabai, A., *Statistical analysis and simulation study of video teleconference traffic in ATM networks*, Preprint, Bellcore (1992).

Jagerman, D.L. and Melamed, B., *The transition and autocorrelation structure of TES processes Part I: General theory*, Stochastic Models **8** (1992a), 193-219.

Jagerman, D.L. and Melamed, B., *The transition and autocorrelation structure of TES processes Part II: Special cases*, Stochastic Models, **8, 499-527** (1992b).

Hurst, H.E., *Long-term storage capacity of reservoirs*, Transactions of the American Society of Civil Engineers **116** (1951), 770-808.

Hurst, H.E., *Methods of using long-term storage in reservoirs*, Proceedings of the Institution of Civil Engineers, Part I (1955), 519-577.

Lewis, P.A.W. and McKenzie, E., *Minification processes and their transformations*, Journal of Applied Probability 28 (1991), 45-57.

Livny, M., Melamed, B. and Tsiolis, A.K., *The impact of autocorrelations on queueing systems*, Management Sciences **39** (1993), 322-339.

Melamed, B., *TES: a class of methods for generating autocorrelated uniform random variates*, ORSA Journal of Computing **3** (1991), 317-329.

Rajput, B. and Rosinski, J., *Spectral representation of infinitely divisible processes*, Probability Theory and Related Fields **82** (1989), 451-458.

Resnick, Sidney, *Extreme Values, Regular Variation, and Point Processes*, Springer-Verlag, New York, 1987.

Resnick, S., *Adventures in Stochastic Processes*, Birkhauser, Boston, 1992.

Samorodnitsky, G. and Taqqu, M., *Stable Non–Gaussian Random Processes: Stochastic Models with Infinite Variance*, Chapman and Hall, New York, 1994.

Schruben, Lee, *Graphical Simulation Modeling and Analysis using Sigma for Windows*, Scientific Press, San Francisco, CA, 1994.

Appendix.

In this Appendix we present a broader point of view on the input process which aids in understanding how the long range dependence arises in our models and further shows that our model is a particular case of a family of stationary ergodic inputs. This allows one to use different simulation designs, depending on the need.

To describe the general model we introduce some terminology and background information. Let $\mathcal{B}$ be the Borel σ-field on $\mathbb{R}$, and let M be an independently scattered infinitely divisible random measure on $\mathcal{B}$ with control measure

$$\lambda(ds) = ds, \quad s \in \mathbb{R}, \tag{A.1}$$

and instantaneous Lévy measures

$$q(dx, s) = \frac{e^{-x}}{x} 1(x > 0)dx,\ x \in \mathbb{R},\ s \in \mathbb{R}, \tag{A.2}$$

(see, for example, Samorodnitsky and Taqqu (1994)). That is, M is just a stochastic process indexed by Borel sets of finite Lebesgue measure such that if $A_1, A_2, \dots$ are pairwise disjoint, then $M(A_1), M(A_2), \dots$ are independent (that is what *independently scattered* means), and if furthermore the union of the A_n's above has a finite Lebesgue measure as well, then $M(\cup_{n=1}^{\infty} A_n) = \sum_{n=1}^{\infty} M(A_n)$ with probability 1. The fact that the control measure λ is Lebesgue measure and that the instantaneous Lévy measure q does not depend on s implies that the distribution of the random variable $M(A)$ depends only on the Lebesgue measure of the set A. This is easily seen from the fact that each $M(A)$ is a non-negative infinitely divisible random variable with Laplace transform

$$Ee^{-\alpha M(A)} = \exp\{-\int_0^{\infty} (1 - e^{-\alpha x})\nu_{M(A)}(dx)\}, \ \alpha \geq 0 \tag{A.3}$$

and Lévy measure $\nu_{M(A)}$ given by

$$\nu_{M(A)}(dx) = \int_A q(dx, s)\lambda(ds) = \lambda(A)\frac{e^{-x}}{x}1(x > 0)dx. \tag{A.4}$$

In particular, $M(A)$ is a Gamma random variable with scale parameter 1, and shape parameter equal to $\lambda(A)$ (cf. Feller, 1971, page 451 for example). If $\lambda(A) = 1$, then $M(A)$ is the standard exponential random variable. Note that $\{M((0,t], t \geq 0\}$ is a Gamma process and has stationary independent increments. We refer the reader to Rajput and Rosinski (1989) for more information on infinitely divisible random measures and stochastic integrals with respect to these measures.

Another description of the process M is as follows: Let

$$\sum_k \epsilon_{(t_k, j_k)}$$

be a Poisson random measure (PRM) on $[0,\infty) \times (0,\infty]$ with mean measure $\lambda \times x^{-1}e^{-x}dx$. (Cf. Resnick, 1987.) Here $\epsilon_{(x,y)}$ is the probability measure concentrated at the point (x, y). Then we may define for any $A \in \mathcal{B}$ such that $\lambda(A) < \infty$

$$M(A) = \sum_{t_k \in A} j_k.$$

Our basic input process is a *moving average* with respect to the random measure M, which formally is a stochastic integral of the form

$$\xi_n = \int_{-\infty}^{\infty} f(n + s)M(ds), \ n = 0, 1, 2, \dots. \tag{A.5}$$

Here $f \geq 0$ is a kernel function, which is our degree of freedom in specifying the dependence between the interarrival times. The moving average (A.5) is well defined if and only if

$$\int_{[0,\infty)\times(0,\infty]} (f(t)x \wedge 1)dt\frac{e^{-x}}{x}dx < \infty \tag{A.6}$$

(cf. Rajput and Rosinski (1989)). In this case $\{\xi_n, \, n \geq 0\}$ is a (strictly) stationary process (infinitely divisible, in fact), such that each ξ_n is a non-negative infinitely divisible random variable with Laplace transform

$$Ee^{-\alpha \xi_n} = \exp\{-\int_0^{\infty} (1 - e^{-\alpha x})\nu_{\xi}(dx)\}, \ \alpha \geq 0 \tag{A.7}$$

and Lévy measure ν_ξ

$$\nu_\xi(B) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} 1(xf(s) \in B) q(dx, s)\lambda(ds). \tag{A.8}$$

This process is also ergodic (cf. Cambanis et. al. (1991)).

The input processes used in this paper correspond to a specific choice of kernel f, which we now give.

Let $\{a_i,\, i \geq 1\}$ be a nonincreasing sequence of nonnegative numbers satisfying (3.1). Let I_j denote the interval $(j-1, j-1+a_j)$, $j = 1, 2, \ldots$ and define

$$f(u) = 1(u \in \cup_{j=1}^{\infty} I_j). \tag{A.9}$$

One readily checks that (A.6) holds, and therefore our input process (A.5) is a stationary process, and we see from (A.8) that each ξ_n has the standard exponential distribution. Note also that we may write

$$\xi_n = \sum_{j=1}^{\infty} M((j-1-n, j-1-n+a_j]), \tag{A.10}$$

so that ξ_n is a sum of Gamma distributed random variables; recall that

$$M((j-1-n, j-1-n+a_j])$$

is Gamma distributed with shape parameter a_j.

The dependence structure of the input process is obviously determined by the numbers $\{a_n,\, n \geq 1\}$. Since the kernel $f(n + \cdot)$ in (A.5) is defined by shifting the set $\cup_{j=1}^{\infty} I_j$, and since the random measure M assigns independent random weights to disjoint parts of the real line, it follows that the larger the intersection of the set $\cup_{j=1}^{\infty} I_j$ with the shifted copies of itself, the stronger the dependence the interarrival time process $\{T_n,\, n \geq 0\}$ exhibits. In particular, in the extreme case $a_1 = 1$, $a_n = 0$ for every $n \geq 2$ this intersection is empty, and the interarrival times are i.i.d. On the other hand, we can make the dependence stronger by making the sequence $\{a_n,\, n \geq 1\}$ decrease in a regular and *slow* way.

Consider the the distributional structure of $\xi_0, \xi_1, \ldots, \xi_{n-1}$. For a subset $A \subset \{0, 1, \ldots, n-1\}$ let

$$I_A = \{x \in \mathbb{R} :\ f(i+x) = 1 \text{ for all } i \in A,\ f(i+x) = 0 \text{ for all } i \notin A\}.$$

Then, obviously,

$$\xi_i = \sum_{A \ni i} M(I_A),\ i = 0, 1, \ldots, n-1.$$

Since for different A's the sets I_A's are pairwise disjoint, the corresponding terms $M(I_A)$'s are independent. Moreover, each $M(I_A)$ has Gamma distribution with scale parameter 1 and shape parameter $|I_A|$.

It is easy to understand the structure of the numbers $|I_A|$'s. First of all,

$$|I_A| = 0 \text{ if for some } i < j < k,\ i \in A,\ k \in A,\ j \notin A,$$

and so we consider now sets A of the form

$$A(i,j) = \{i, i+1, \ldots, j\},\ i \leq j.$$

Let $\gamma(i,j) = |I_{A(i,j)}|$. It is simple to see that these numbers are given by (3.2).

Therefore, denoting $\Gamma(i,j) = M(I_{A(i,j)})$, $0 \leq i \leq j \leq n-1$, we conclude that the vector $(\xi_0, \xi_1, \ldots, \xi_{n-1})$ can be represented as in (3.3).

Sidney I. Resnick, Cornell University, School of Operations Research and Industrial Engineering, ETC Building, Ithaca, NY 14853 USA

E-mail: sid@orie.cornell.edu

Gennady Samorodnitsky, Cornell University, School of Operations Research and Industrial Engineering, ETC Building, Ithaca, NY 14853 USA

E-mail: gennady@orie.cornell.edu