Some invariance properties of monotone failure rate in the $M/G/1$ queue

Yoav Kerner

EURANDOM, Eindhoven University of Technology, Eindhoven, The Netherlands.
E-mail: kerner@eurandom.tue.nl.

Abstract. We show that in the stationary $M/G/1$ queue, if the service time distribution is with increasing (decreasing) failure rate (IFR (DFR)), then (a) The distribution of the number of customers in the system is also IFR (DFR), (b) The conditional distribution of the remaining service time given the number of customers in the system is also IFR (DFR) and (c) The conditional distribution of the remaining service time given the number of customers in the system, is stochastically decreasing (increasing) with the number of customers in the system.

1 Introduction

The distribution of the residual service time in a single server queue, has many reasons of interest. For example it is a non-trivial component of the sojourn time distribution. Furthermore, the residual service time is the remaining time until the next departure from the queue, and in a network this may contribute to the remaining time until the arrival to another queue. The literature on the departure process from a queue (which is the arrival process to the next queue) is wide and most of it is relatively old. See e.g. a survey by Daley [5]. Most of the literature deals with the marginal distribution of the departure process and or/the inter departure times. In this paper we discuss the *conditional* residual service time, which is the conditional time until the next departure, given the number of customers in the queue. The dependence between the residual service time and the number of customers in the system can be explained as follows. The distribution of the residual service time depends on the past service time. However, if the past itself is not given, the number of customers in the system, which is a function of both the arrival process and the service process, supplies information about the past service time. This information changes our prior belief regarding the distribution of the residual service time.
We consider an $M/G/1$ queue with arrival rate λ and service time CDF $G(\cdot)$, with $G^*(\cdot)$ being its Laplace-Stieltjes Transform (LST) and $\bar{x}$ being its associated mean value. Assume that $\lambda\bar{x} < 1$ and that the system is in steady state. Denote the number of customers in the system by Q and denote the generic random variable R_n having the same distribution as the residual service time given $Q = n$. For any work conserving and non-preemptive service regime, the model's parameters λ and G determine the joint and marginal distributions of

Q and R_n, $n \geq 1$. Here we study the influence of G being with monotone failure rate on the distributions of R_n and Q. The literature regarding the conditional residual service time (or its joint distribution with the queue length) is quite wide, see e.g. [1, 4, 6–8]. A relatively new study with the scope of this paper was done by Ross [10]. Ross showed that if G is DMRL (IMRL) (that is, the expected residual life time is decreasing (increasing) with its past life time) then Q is stochastically larger (smaller) than the queue length in the $M/M/1$ queue with the same utility level. In this work we show that the IFR (DFR) property is transferred from G to R_n and to Q. Also, we show that if G is IFR (DFR), then R_n is stochastically decreasing (increasing) with n.
The paper is organized as follows. In Section 2 we give definitions of stochastic order and monotone failure rate distributions and state two lemmas which connect these two and can be applied to the residual service time. In section 3 we state our main theorems and discuss a few consequences of these theorems. We conclude this paper in section 4 in which two examples are given, both of them are such that the underlying $M/G/1$ queue can be described as two dimensional Markov chain. In these two examples we present the conditional distribution of the residual service time in a way that our results are observed almost immediately.

2 Properties of distributions with monotone failure rate

In this section we give some definitions and notations. Also we state some properties of distributions with monotone failure rate, which come from reliability theory and stochastic order theory.

Definition 1. *A random variable X is said to be stochastically larger than a random variable Y (denote by $X >_{st} Y$), if for any x, $P(X > x) \geq P(Y > x)$.*

Definition 2. *A non-negative random variable (or its associated distribution function F) is said to be with increasing (decreasing) failure rate (IFR (DFR)) if $\bar{F}(t+s)/\bar{F}(t)$ is non-increasing (non-decreasing) with t for any $s \geq 0$, such that $\bar{F}(s+t) > 0$), where $\bar{F} = 1 - F$.*

Now, for two independent random variables X and Y, define the generic random variable $\{X\}_Y$ having the distribution of $X - Y | X > Y$. Note that $\{\{X\}_Y\}_Z \stackrel{d}{=} \{X\}_{Y+Z}$. The following lemmas present relations between X (when it has a monotone failure rate) and $\{X\}_Y$.

Lemma 1. *Let X and Y be two non-negative independent random variables. If X is IFR (DFR) then $\{X\}_Y$ is IFR (DFR) as well.*

Proof. We prove the result for the IFR case, while the proof for the DFR case is equivalent. We want to prove that for any positive t, s, u,

$$\frac{P(\{X\}_Y > t+u)}{P(\{X\}_Y > t)} > \frac{P(\{X\}_Y > t+u+s)}{P(\{X\}_Y > t+s)}$$

which is equivalent to

$$P(X > Y + t + u)P(X > Y + t + s) > P(X > Y + t + u + s)P(X > Y + t)$$

or,

$$\log P(X > Y + t + s) + \log P(X > Y + t + s) > \tag{1}$$
$$> \log P(X > Y + t + u + s) + \log P(X > Y + t).$$

Note that X is IFR if and only if $\bar{F}$ is log-concave. Thus, for any positive y, we have

$$\log(\bar{F}(y+t+u)) + \log(\bar{F}(y+t+s)) - \log(\bar{F}(y+t+u+s)) - \log(\bar{F}(y+t)) \geq 0.$$

Multiplying the latter by the density of Y and integrating implies (1). In the DFR case all the inequality signs are reversed.

Lemma 2. *Let X, Y, Z be non-negative random variables such that X is independent on (Y, Z) and $P(Y \leq Z) = 1$. If X is IFR, then $\{X\}_Y \geq_{st} \{X\}_Z$.If X is DFR, the stochastic inequality is reversed.*

Proof. We want to show that

$$\frac{P(X > Y + t)}{P(X > Y)} > \frac{P(X > Z + t)}{P(X > Z)}$$

which is equivalent to

$$\log P(X > Y + t) - \log P(X > Y) - \log P(X > Z + t) + \log P(X > Z) > 0$$

Since X IFR, for any realization y, z of (Y, Z), we have

$$\log P(X > y + t) - \log P(X > y) - \log P(X > z + t) + \log P(X > z) > 0.$$

Thus, multiplying the latter by the joint density (assume for simplicity that it exists) of (Y, Z) keeps the inequality. In the DFR case, all the inequalities are reversed.

3 Main results

In this section we apply the results from section 2 to the conditional residual service time in the $M/G/1$ queue. It is shown in [7] that the CDF of R_n follows the recursion

$$F_n(t) = P(R_n \leq t) = (1 - G^*(\lambda))P(\{X\}_Y \leq t) + G^*(\lambda)P(\{R_{n-1}\}_Y \leq t) \tag{2}$$

where $X \sim G$ and $Y \sim Exp(\lambda)$. Denote by $G_k(t)$ the CDF of the random variable $\{X\}_{Y_k}$ where X follows the service time distribution and $Y_k \sim Erlang(k, \lambda)$.

Applying the recursive formula of the distribution R_n $n-1$ times implies that the CDF of R_n can be presented as

$$F_n(t) = (1 - G^*(\lambda)) \sum_{i=1}^{n-1} (G^*(\lambda))^{i-1} G_i(t) + (G^*(\lambda))^{n-1} G_n(t). \tag{3}$$

This representation allows us to prove the following theorem.

Theorem 1. *In the stationary $M/G/1$ queue, if the service time distribution is IFR (DFR), then for any $n \geq 1$, R_n is IFR (DFR).*

Proof. Let N be a geometric random variable with probability of success $1 - G^*(\lambda)$ and let $Y_1, Y_2, \ldots$ be i.i.d. random variables with the distribution $Exp(\lambda)$. Also, let S_n be a generic random variable with the same distribution as $\sum_{i=1}^{N \wedge n} Y_i$. Following (3), we observe that $R_n \stackrel{d}{=} \{X\}_{S_n}$. The rest is immediate from lemma 1.

Remark 1. The above proof holds for both IFR and DFR cases. An alternative proof for the DFR case is as follows. First, by Lemma 1, R_1 is DFR. Second, as a mixture of DFR is also DFR (see [3]), and from (2) we see that R_n is a mixture of $\{X\}_Y$ and $\{R_{n-1}\}_Y$. Hence, by induction, R_n is DFR as well.

In the next theorem we observe a stochastic order in the random sequence R_n.

Theorem 2. *In the stationary $M/G/1$ queue, if the service time distribution is IFR (DFR), then R_n is stochastically decreasing (increasing) in n.*

Proof. We prove the theorem first for the IFR case. We consider the sequence S_n which was defined in the proof for Theorem 1. We write

$$S_{n+1} = S_n + Y I_{\{N>n\}}.$$

Hence, by Lemma 2, since $P(S_n \leq S_{n+1}) = 1$, we have $\{X\}_{S_n} >_{st} \{X\}_{S_{n+1}}$ which is equivalent to $R_n >_{st} R_{n+1}$.

Next we learn from the monotonicity of the failure rate function of the service time distribution, about the monotonicity of the (discrete) failure function of the underlying distribution of the number of customers in the system.

Corollary 1. *In the stationary $M/G/1$ queue, if the service time distribution is IFR (DFR), then the number of customers in the system is IFR (DFR) as well.*

Proof. Let $\pi_n = P(Q = n)$ and let $h_n = P(Q = n | Q \geq n)$ be the failure function of Q. In [8], the following formula which connects the distribution of Q and the expected value of R_n appears:

$$E(R_n) = \frac{1-\rho}{\lambda P(Q=n)} P(Q > n) = \frac{1-\rho}{\lambda} \left(\frac{1}{h_n} - 1 \right) \tag{4}$$

Thus, $E(R_n)$ is decreasing (increasing) with n if and only if h_n is increasing (decreasing) with n. From Theorem 2 we have that if the service time distribution is IFR (DFR) then R_n is stochastically decreasing (increasing) which implies that $E(R_n)$ is decreasing (increasing), which in turn is equivalent to Q being IFR (DFR).

Remark 2. As the set of IFR distributions is closed under convolutions, Theorem 1 implies immediately that if the service time distribution is IFR, then the conditional sojourn time, given the number of customers in the system upon arrival is IFR as well. This of course does not imply that the marginal distribution of the sojourn time is IFR, as the set of IFR distributions is not closed under mixtures.

4 Examples

In this section we give two examples, one is IFR and the other is DFR, in which the results in Theorems 1 and 2 can be derived directly. The $M/G/1$ queue when the service time distribution is one of these two examples was studied intensively in the literature, mostly in the two dimensional Markov process setting. See e.g. [9]. For our purposes, the simplest way to present the residual service time is using (3).

Example 1: E_k distribution
Assume that the service time distribution is the Erlang distribution with k phases and a rate μ of each phase. In this case, as we show next, the distribution of R_n can be obtained explicitly. Given $n \wedge N = m$, i.e. $S_m \sim E(m, \lambda)$, we have

$$P(X > S_n + u) = \int_{y=0}^{\infty} \int_{x=y+u}^{\infty} \frac{\mu^k \lambda^m x^{k-1} y^{m-1} e^{-\lambda y - \mu x}}{(m-1)!(k-1)!} dx dy =$$

$$= \sum_{i=0}^{k-1} \sum_{j=0}^{i} e^{-\mu u} \frac{(\mu u)^{i-j}}{(i-j)!} \binom{m-1+j}{j} \left(\frac{\mu}{\mu+\lambda}\right)^j \left(\frac{\lambda}{\mu+\lambda}\right)^m.$$

From the latter we learn that R_n can be written as a mixture of independent random variables. In particular, let W_n be an integer valued random variable which gets values between 1 and k. Thus, $R_n|W_n = i \sim Erlang(i, \mu)$. Moreover, $W_n \stackrel{d}{=} k - B_n | B_n < k$ where B_n is a negative binomial random variable with number of successes $n \wedge N$ and probability of success $\frac{\lambda}{\lambda+\mu}$. According to Theorem 7.1 in [11], a random sum of i.i.d. exponential random variables, in which the number of addend is IFR, is IFR as well. Hence, to show the result of Theorem 1 we need to show that W_n is IFR. As B_n is a sum of geometric random variables, it is IFR. It can be shown (see e.g. [2], p. 37) that it is equivalent to $P(B_n = j | B_n \leq j)$ is decreasing, which is in turn equivalent to W_n being IFR. Obviously, $B_n <_{st} B_{n+1}$ and hence $W_n >_{st} W_{n+1}$. As R_n and R_{n+1} are random

sums of random variables from the same distribution (see, e.g., Theorem 1.A.4 in [12]), $R_n >_{st} R_{n+1}$.

Example 2: H_2 distribution
Assume that the service time follows the Hyper exponential distribution. That is $G(x) = 1 - \alpha e^{-\mu_1 x} - (1-\alpha)e^{-\mu_2 x}$ for some α, μ_1, μ_2, such that $0 < \alpha < 1$. Assume w.l.o.g. that $\mu_1 < \mu_2$. It is clear that for any n, the residual service time is hyper exponential as well and hence DFR. Let α_n be the probability that the server is working at a rate of μ_1, given that there are n customers in the system. In our case, the result of Theorem 2 is equivalent to α_n being increasing in n. We show that next. Given $n \wedge N = m$ we have, using (3) and conditioning on the service rate selected in the beginning of the service,

$$P(R_n > r) = P(X > S_m + r | X > S_m) = \frac{\alpha e^{-\mu_1 r}\left(\frac{\lambda}{\lambda+\mu_1}\right)^m + (1-\alpha)e^{-\mu_2 r}\left(\frac{\lambda}{\lambda+\mu_2}\right)^m}{\alpha\left(\frac{\lambda}{\lambda+\mu_1}\right)^m + (1-\alpha)\left(\frac{\lambda}{\lambda+\mu_2}\right)^m}.$$

Thus,

$$\alpha_n = \frac{\alpha \mathrm{P}_n\left(\frac{\lambda}{\mu_1+\lambda}\right)}{\alpha \mathrm{P}_n\left(\frac{\lambda}{\mu_1+\lambda}\right) + (1-\alpha)\mathrm{P}_n\left(\frac{\lambda}{\mu_2+\lambda}\right)} = \frac{\alpha \mathrm{P}_n\left(\frac{\lambda}{\mu_1+\lambda}\right)/\mathrm{P}_n\left(\frac{\lambda}{\mu_2+\lambda}\right)}{\alpha \mathrm{P}_n\left(\frac{\lambda}{\mu_1+\lambda}\right)/\mathrm{P}_n\left(\frac{\lambda}{\mu_2+\lambda}\right) + 1 - \alpha} \tag{5}$$

where

$$\mathrm{P}_n(z) = \mathrm{E}\left(z^{n \wedge N}\right) = \frac{z(1-G^*(\lambda))}{1 - zG^*(\lambda)} + \frac{1-z}{1-zG^*(\lambda)}(zG^*(\lambda))^n.$$

Note that since $\mu_1 < \mu_2$, $\frac{\lambda}{\lambda+\mu_1} < \frac{\lambda}{\lambda+\mu_2}$. Thus, what we need to show is that the ratio $\mathrm{P}_n(z_1)/\mathrm{P}_n(z_2)$ is increasing in n, for $z_1 > z_2$. We write

$$\frac{\mathrm{P}_n(z_1)}{\mathrm{P}_n(z_2)} = C\left(\frac{a_1 + b_1 e^{-\theta_1 n}}{a_2 + b_2 e^{-\theta_2 n}}\right)$$

where

$$C = \frac{1 - z_2 G^*(\lambda)}{1 - z_1 G^*(\lambda)}, \quad a_i = z_i(1 - G^*(\lambda)), \quad b_i = 1 - z_i \quad \textit{and} \quad \theta_i = -\log(z_i G^*(\lambda)).$$

A simple calculus shows that if $C > 0, a_1 > a_2, b_1 < b_2$ and $\theta_1 < \theta_2$ (as in our case), this ratio is increasing.

Acknowledgments

Thanks are due to Ivo Adan, Onno Boxma, and Moshe Haviv for discussion and remarks.

References

1. I. Adan and M. Haviv (2009), "Conditional ages and residual service times in the M/G/1 queue," *Stochastic Models*, Vol. 25, pp 110-128.
2. R.E. Barlow and F. Proschan (1965), *Mathematical theory of reliability,* Wiley, New York.
3. R.E. Barlow, A.W. Marshall and F. Proschan (1963), Properties of probability distributions with monotone failure rate, *Annals of Mathematical Statistics,* vol 34, pp 375-389.
4. O.J. Boxma, (1984), "Joint distribution of sojourn time and queue length in the M/G/1 queue with (in)finite capacity, " *European Journal of Operational Research,* Vol. 16, pp. 246-256.
5. D.J. Daley, (1975), "Queueing output processes," *Advances in Applied Probability,* Vol. 8, pp. 395-415
6. D. Fakinos, (1990), "Equilibrium queue size distributions for semi-reversible queues," *Stochastic Processes and Their Applications,* Vol. 36, pp. 331–337.
7. Y. Kerner, (2008), "The conditional distribution of the residual service time in the $M_n/G/1$ queue," *Stochastic Models*, Vol. 24, pp. 364-375.
8. A. Mandelbaum and U. Yechiali (1979), "The conditional residual service time in M/G/1 queue. Unpublished manuscript. Also at http://www.math.tau.ac.il/uriy/Publications.html.
9. M.F. Neuts, (1982), "Explicit steady-state solutions to some elementary queueing models," *Operations Research* Vol. 30, pp. 480-489.
10. S.M. Ross, (2006), "Bounding the stationary distribution of the M/G/1 queue size," *Probability in the Engineering and Informational Sciences,* Vol. 20, pp. 571-574.
11. S.M. Ross, J.G. Shanthikumar and Z. Zhu (2005), "On increasing-failure-rate random variables," *Journal of Applied Probability,* Vol. 42, pp. 797-809.
12. M. Shaked and J.G. Shanthikumar (1993), *Stochastic orders and their applications*, Academic Press.