Double Random Element: Definition, Property, and Example

Guang-Liang Li Victor O. K. Li

Department of Electrical and Electronic Engineering
The University of Hong Kong
{glli,vli}@eee.hku.hk

May 21, 2003

(draft for comments)

Abstract

A double random element describes a phenomenon that cannot be described by a single random variable. We define the double random element, and discuss its basic properties. Taking Jackson's theorem as an example, we show that the distinction between a double random element and a random variable can help identify and prevent mistakes in stochastic modeling.

1 Introduction

A double random element is a collection of random variables, indexed by possible values of another random variable. Corresponding to a double random element, there exists a random variable, which we call "the marginal version of the double random element." Though double random elements are widely used in stochastic modeling, they are confused with their marginal versions. Such confusion can even be found in standard textbooks. In fact, double random elements are not included in the existing literature.

The purpose of this paper is to define the double random element, discuss its basic properties (Section 2), and clarify the confusion between double random elements and their marginal versions in stochastic modeling of networking studies (Sections 3 and 4).

We have recently shown that Jackson's theorem is incorrect [3, 4]. By distinguishing between double random elements and their marginal versions, we try to explain why Jackson's proof, and all other proofs of Jackson's theorem, are flawed, and why Jackson's theorem has been considered correct for almost half a century.

Four queues are used in the exposition. In Section 3, two queues are used to explain why Jackson's theorem is incorrect in the presence of feedback. In Section 4, two queues are used to explain why Jackson's theorem is incorrect even if feedback is prohibited. We conclude in Section 5.

2 Double Random Element: Definition and Basic Properties

A double random element is defined based on a random variable X, a collection E of random variables, and a map from A, the set of all possible values of X, to E. The random variable X and the set A are respectively the *index* and the *index set* of the double random element. Denote such an object by $X|E$. We may read $X|E$ as "double random element X, E," or "double random element indexed by X," if omission of E will not cause confusion. The members of E are *components* of $X|E$.

Definition 1 *A double random element $X|E$ is a bijection from A to E, where A is the set of all possible values of a random variable X, and E is a collection of random variables. Any realization of $X|E$ begins with a realization of X.*

Corresponding to $x \in A$, the random variable in E is denoted by $x|E$. We call $x|E$ "the component of $X|E$ indexed by x," or "the component indexed by x," if X and E can be omitted without causing confusion. The theorem below is an immediate result of Definition 1, and can be used as an alternative definition of double random elements.

Theorem 1 *(a) Any realization of $X|E$ follows two steps: X must be realized first, and a realization of $x|E$ follows, where x is the realized value of X. (b) z is a realized value of $X|E$, if and only if z is a realized value of some $x|E$, where $x \in A$.*

In the rest of this paper, we shall focus on a special type of double random elements, such that X is a discrete random variable, and $x|E$ is a continuous random variable for each $x \in A$.

Definition 2 *The marginal version of a double random element $X|E$, denoted by $[X|E]$, is a random variable with a distribution*

$$P\{[X|E] \leq z\} = \sum_{x \in A} P\{x|E \leq z\} P\{X = x\}.$$

Definition 3 *Two double random elements $X_1|E_1$ and $X_2|E_2$ are statistically identical, if X_1 and X_2 are identically distributed, and if $x|E_1$ and $x|E_2$ are identically distributed for any possible value x of X_1.*

If A is a set of nonnegative integers, we denote X and $X|E$ by N and $N|E$, respectively. The distribution of the marginal version of $N|E$ is

$$P\{[N|E] \leq z\} = \sum_{n \in A} P\{n|E \leq z\} P\{N = n\}. \tag{1}$$

For a double random element $N|E$, let F_n and f_n be, respectively, the distribution and density functions of $n|E$, where $n \in A$, and denote by F and f, respectively, the distribution and density functions of $[N|E]$.

Definition 4 *A double random element $N|E$ is linearly independent, if $\{F_n : n \in A\}$ is linearly independent, or equivalently, if $\{f_n : n \in A\}$ is linearly independent.*

Two functions F_1 and F_2 are *essentially identical*, if values of F_1 and F_2 may differ only on some sets of measure zero in their domains; otherwise the two functions are *essentially different*. Without considering any set of measure zero, we write $F_1 \neq F_2$ if F_1 and F_2 are essentially different, and $F_1 = F_2$ otherwise.

Theorem 2 *(a) If all $n|E, n \in A$ follow the same distribution, $[N|E]$ is distributed as the components of $N|E$. (b) If $N|E$ is linearly independent, $F_n \neq F$ and $f_n \neq f$ for any $n \in A$.*

Proof: (a) This follows immediately from (1). (b) If there exists some $k \in A$, such that $F_k = F$, from (1), F_k could be expressed as a linear combination of the other members of $\{F_n, n \in A\}$. But this contradicts the linear independence of $\{F_n, n \in A\}$. So $F_n \neq F$ for all $n \in A$. Similarly, $f_n \neq f$ for all $n \in A$.

□

3 Jackson's Feedback Queue and Its Modified Counterpart

The simplest Jackson network is a single-server queue with instantaneous Bernoulli feedback. The waiting capacity is infinite. Customers arrive from the outside in accordance with a Poisson process at rate λ. Once served, a customer may either depart, with probability $1-\theta$, or return to the system immediately, with probability θ. The first-come-first-served discipline is applied to both external and returning customers. Service times of both types of customers are independent and identically distributed (i.i.d.) exponential random variables, with parameter μ. At time t, the state of the feedback queue is modeled by the number of customers waiting and being served in the queue, denoted by $Q(t)$.

If the state of a queue can be modeled as a stationary process, the queue has a steady state. For the feedback queue, Jackson's theorem says, in essence, that "so far as steady states are concerned," this queue would behave as if it was an ordinary M/M/1 queue without feedback. The basic *assumption* is that $Q(t)$ could be modeled as a Markov process [2].

This simplest Jackson network has been used in the literature to explain Jackson's theorem, and is directly obtained from Jackson's model [2]. It is, however, a counterexample of Jackson's theorem [3, 4]. The flaw in Jackson's proof has been identified [3].

A remedy for Jackson's theorem in the case of the feedback queue argues that, by allowing any returning customer to enter service again immediately rather than waiting in line, Jackson's feedback queue can be modified without affecting the number of customers in the queue. This remedy claims that the evolution of the state probability of the feedback queue and the modified queue is governed by the same equation

$$\begin{aligned}&P\{U(t+\Delta t)=k\}=[1-\lambda\Delta t-\mu(1-\theta)\Delta t]P\{U(t)=k\}+\mu(1-\theta)\Delta tP\{U(t)=k+1\}\\&+\lambda\Delta tP\{U(t)=k-1\}+o(\Delta t)\end{aligned} \tag{2}$$

where $U(t)$ is the state of the queue at time t, $k>1$, and $\Delta t>0$ is an infinitesimally small time increment. But (2) actually corresponds to an ordinary M/M/1 queue without feedback. In other words, the remedy tries to salvage Jackson's theorem, by considering the state of the feedback queue identical to the state of the M/M/1 queue without feedback, *at all times*. This is not even claimed in Jackson's theorem. Such argument is an example of the confusion between double random elements and their marginal versions.

Consider first the feedback queue. Two time epochs, r and s, where $r<s$, define a time interval $[r,s)$, such that $Q(r-)\neq k$, where $k>0$ is given, $Q(t)=k$ for $t\in[r,s)$, and $Q(s)=k-1$. Clearly, $s-r$ is a sum of a random number of exponential random variables of rate μ. Denote by Y_j the jth random variable in $[r,s)$.

$$s-r=\sum_{j=1}^{N}Y_j \tag{3}$$

where N is a geometric random variable.

As indicated clearly by (3), any realization of $s-r$ follows two steps: N must be realized first, and a realization of $\sum_{j=1}^{n}Y_j$ follows, where n is a realized value of N. Moreover, any realized value of $s-r$ must be a realized value of $\sum_{j=1}^{n}Y_j$ for some realized value n of N.

From Theorem 1, we recognize that $s-r$ is a double random element. We use $N|E$ to represent this double random element, where E is a collection of Erlang random variables with parameters μ and n, denoted by $n|E, n=1,2,\cdots$.

However, in the literature, $s-r$ is confused with its marginal version $[N|E]$, which is an exponential random variable of rate $\mu(1-\theta)$. Since $N|E$ is linearly independent, from Theorem 2, $[N|E]$ is not equal in distribution to any component of $N|E$.

Thus, there are two interpretations of Jackson's feedback queue. In one interpretation, $s-r$ is $N|E$, a double random element. In the other interpretation, $s-r$ is $[N|E]$, the marginal version of $N|E$. But only one of the two interpretations is correct.

Is $s-r$ a double random element or its marginal version? Since $s-r$ is a directly observable quantity, if $s-r$ is $[N|E]$, in a sample path of $Q(t)$, realized values of the exponential random variable of rate $\mu(1-\theta)$ should be directly observed. But such realized values do not exist in any sample path of $Q(t)$. On the contrary, realized values of $s-r$ observed in any realization of $Q(t)$ are realized values of $n|E, n=1,2,\cdots$, the components of $N|E$. Therefore, $s-r$ is a double random element, not its marginal version.

Theorem 3 *The state of Jackson's feedback queue, i.e., $Q(t)$, is not a Markov process.*

Proof: Since $s-r$ is a double random element $N|E$, whose components $n|E, n=1,2,\cdots$ follow different distributions, $s-r$ is not a random variable with an unvarying distribution. The result to be proved then follows.

□

Now consider the modified queue. Denote by $R(t)$ the state of the modified queue at time t. Two time epochs, u and v, where $u<v$, define a time interval $[u,v)$, such that $R(u-)\neq k$ for a given $k>0$, $R(t)=k$ for $t\in[u,v)$, and $R(v)=k-1$. The customer in service at time u will be served during the whole interval $[u,v)$.

Theorem 4 *The state of the modified queue, i.e., $R(t)$, is not a Markov process.*

Proof: Evidently, $[u,v)$ consists of M i.i.d. exponential random variables of rate μ, where M is a geometric random variable, distributed as N in (3). Let Y_j be the jth random variable in $[u,v)$. We have

$$v-u=\sum_{j=1}^{M}Y_j. \tag{4}$$

From (4) and (3), $v-u$ and $s-r$ are statistically identical double random elements. Like $Q(t)$, $R(t)$ cannot be a Markov process.

□

Theorem 5 *Neither $Q(t)$ nor $R(t)$ is stationary.*

Proof: Since $v-u$, the service time (or residual service time) of any customer at the modified queue, is a double random element statistically identical to $s-r$, the service time sequence of the modified queue is not stationary. Consequently, the modified queue is not stable, and hence $R(t)$ is not stationary. As a result, $Q(t)$ cannot be stationary, since $Q(t)$ is identical to $R(t)$.

□

The instability of a queue means that the behavior of the queue cannot be described by a time-independent probability law, though the state of an unstable queue may always be finite.

4 Downstream Queue in Tandem Network

Consider a simple Jackson network of two single-server queues in tandem, such that all customers arrive at the first queue according to a Poisson process, go to the second queue after service, and leave the network from there. At each queue, the waiting capacity is infinite, service times are i.i.d.

exponential random variables, and customers are served in the order of arrivals. Service times required by the same customer at different queues are independent. The first queue is a stable M/M/1 queue in steady state.

What is the second queue? The answer depends on what is the input of the second queue. But the input of the second queue seems simply to be the output of the first queue. However, the problem is not that simple. What is "the output" of the first queue?

Let us simulate the first queue, or do the simulation as a thought experiment, and observe the inter-departure time sequence of this stable M/M/1 queue in steady state [3, 4].

Let s and t, where $s < t$, be two arbitrary, consecutive departure epochs at the first queue. There are two possibilities: The server of the first queue is exclusively either busy or idle at time s. From a straightforward observation, if the server is busy at s, the inter-departure time $t - s$ is distributed as a service time Y; otherwise, $t - s$ is distributed as the sum of an idle time Z of the server and a service time Y. Define

$$I = \begin{cases} 1, & \text{the server is busy at } s \\ 0, & \text{otherwise.} \end{cases}$$

The observation on the inter-departure time can be described by the following expression.

$$t - s = IY + (1 - I)(Z + Y).$$

Thus, any realization of $t - s$ follows two steps: I must be realized first, and depending on the realized value of I, a realization of Y or $Z + Y$ follows. As a result, any realized value of $t - s$ must be a realized value of either Y or $Z + Y$ exclusively.

We recognize now that $t - s$ is a double random element (see Theorem 1). Denote this double random element by $I|E$, which has two components: $0|E = Z + Y$, and $1|E = Y$.

Denote by F_1 the distribution of Y, and F_2 the distribution of $Z + Y$. According to the observation, in the *physical* departure process of the first queue, the distribution of any inter-departure time is exclusively either F_1 or F_2, depending on the state of the queue, characterized by I. Since $F_1 \neq F_2$, the physical departure process is not stationary.

If the two queues in tandem are considered jointly as a network, the physical departure process of the first queue is the arrival process of the second queue. As a result, the second queue, and hence the network as a whole, are not stable. In other words, denote the state of the first queue by $Q_1(t)$, and the state of the second queue by $Q_2(t)$, the joint process $(Q_1(t), Q_2(t))$, i.e., the state of the network, is not stationary, since $Q_2(t)$ is not stationary. Like the feedback queue, the tandem network of two queues is a counterexample to Jackson's theorem.

For this network, the basic assumption of Jackson's theorem is that $(Q_1(t), Q_2(t))$ could be modeled as a stationary Markov process. Based on this assumption, Jackson's theorem claims that both queues could be stable and hence the network could be stable, and the two queues in "steady state" would behave as if they were independent M/M/1 queues, i.e.,

$$P\{Q_1(t) = k_1, Q_2(t) = k_2\} = P\{Q_1(t) = k_1\}P\{Q_2(t) = k_2\}$$

where both $P\{Q_1(t) = k_1\}$ *and* $P\{Q_2(t) = k_2\}$ are considered the state distributions of M/M/1 queues in statistical equilibrium.

However, as we have shown, the assumption is invalid. So Jackson's theorem does not hold. A confusion between the double random element $I|E$ and its marginal version $[I|E]$ might be the cause of the invalid assumption.

The marginal version of $I|E$ is an exponential random variable, with a parameter identical to the arrival rate of the first queue. Denote by F the distribution of $[I|E]$. Clearly, $F \neq F_1$, and $F \neq F_2$, since $I|E$ is linearly independent (see Theorem 2).

By sampling random variables from F independently, *regardless of the state of the first queue*, and using so sampled random variables as "inter-departure times," a "departure" process of the first queue is *constructed*. This "departure" process is the Poisson "output" referred to in Burke's theorem [1]. If the constructed Poisson process *is used as* the arrival process of the second queue, the second queue *is altered* and *becomes* an M/M/1 queue, and can now be stable.

However, since $I|E$ represents the observation on the physical departure process, and since $I|E$ differs from $[I|E]$, the "departure" process constructed based on $[I|E]$ does not possess the physical meaning of a departure process, i.e., any "departure" epoch in this "departure" process is not a physical departure epoch of the first queue. Indeed, the "inter-departure times" do not exist in the physical departure process [3, 4].

If such "departure" process is used as the arrival process of the second queue, the two queues, which used to be connected in a network, are now separated. Such separation of queues (referred to by Burke also as "stages") in a tandem system is actually the motivation of Burke's seminal work, as Burke himself said in his original paper [1]:

"It is intuitively clear that, in tandem queuing processes of the type mentioned above, if the output distribution of each stage was of such character that the queuing system formed by the second stage was amenable to analysis, then the tandem queue could be analyzed stage-by-stage insofar as the separate delay and queue-length distributions are concerned. Such a stage-by-stage analysis can be expected to be considerably simpler than the simultaneous analysis heretofore necessary. Fortunately, under the conditions stated below, it is true that the output has the required simplicity for treating each stage individually."

Furthermore, such separation is also the *only* way to make a downstream queue in a tandem network "amenable to analysis," if "analysis" is to be performed in steady state. If we are interested in "the separate delay and queue-length distributions," e.g., if we want to *construct* a *marginal* state distribution to describe the steady-state behavior of the second queue, the queue must be separated first, based on the marginal version $[I|E]$ of the double random element $I|E$.

As we have shown, the separation alters the second queue in the tandem network. The second queue *in isolation, after separation* is not the second queue *in the network, before separation* anymore. The separated queue becomes M/M/1, and can be stable. The original queue in the network is neither M/M/1, nor stable.

5 Concluding Remarks

A double random element describes a phenomenon that cannot be described by a single random variable. The existing theory of stochastic modeling fails to capture such phenomena, and even contradicts observed facts. Since double random elements are not included in the current literature, phenomena that should be modeled as double random elements are incorrectly modeled by their marginal versions. The limitation of the existing theory and the contradiction between the theory and the facts so caused must be resolved.

As a result of our effort to this end, we have defined the double random element, and discussed its basic properties. Four queues exemplify the difference between double random elements and their marginal versions. The confusion between double random elements and their marginal versions may explain why all proofs of Jackson's theorem are flawed, and why Jackson's theorem has been considered correct for so long.

References

[1] P. J. Burke, "The output of a queueing system," *Operations Research*, vol. 4, pp. 699-714, 1956.

[2] J. R. Jackson, "Networks of waiting lines," *Operations Research*, vol. 5, no. 4, 518-521, 1957.

[3] Guang-Liang Li and Victor O. K. Li, "Networks of queues: myth and reality," 2003.

[4] Guang-Liang Li and Victor O. K. Li, "Frequently asked questions about why Jackson' theorem is incorrect," 2003.