MANAGEMENT SCIENCE
Vol. 38, No. 6, June 1992
*Printed in U.S.A.*

# SOME GUIDELINES AND GUARANTEES FOR COMMON RANDOM NUMBERS*

PAUL GLASSERMAN AND DAVID D. YAO
*Graduate School of Business, Columbia University, New York, New York* 10027
*IE/OR Department, Columbia University, New York, New York* 10027

Common random numbers (CRN) is a widely-used technique for reducing variance in comparing stochastic systems through simulation. Its popularity derives from its intuitive appeal and ease of implementation. However, though CRN has been observed to work well with a broad range of models, the class of systems for which it is provably advantageous has remained rather limited.

This paper has two purposes: We first discuss the effectiveness and optimality of CRN in a general setting, stressing the roles played by monotonicity and continuity properties. We then present specific, new classes of systems and comparisons for which CRN is beneficial and even optimal. Our conclusions for these systems are largely consistent with simulation practice and lend further theoretical support to folklore. Our results differ from those of previous analyses primarily because we put conditions on the timing of events, rather than the sequence of states, in a discrete-event simulation.

We formulate our results in three settings corresponding to three applications of CRN: distributional comparisons, structural comparisons, and sensitivity analysis. In each case, we make use of conditions that simultaneously ensure monotonicity and continuity in the timing of events. These properties are established through explicit recursions for event epochs in terms of increasing, continuous functions.

(SIMULATION; VARIANCE REDUCTION TECHNIQUES; STOCHASTIC ORDERING; COMMON RANDOM NUMBERS)

## 1. Introduction

A simulator seldom evaluates just one system. More often, simulation is used to compare alternative models or designs. When comparison is the goal, the cost of a simulation study is measured best not by the work required to evaluate each system separately, but by the efficiency with which valid estimates of *differences* in performance may be obtained.

*Common random numbers* (CRN) is the simplest and probably the most widely used method for increasing the efficiency of comparisons via simulation. It is intuitively appealing and easy to implement in either a custom-made simulation or in a simulation package. In its simplest form, CRN just requires that the systems under study be simulated with the same stream of random numbers. Intuitively, this seems fair since it ensures that the systems are compared under the same conditions.

More generally (and more precisely), CRN is a mechanism for introducing dependence to reduce variance. Suppose that the "systems" under consideration are two random variables, vectors or sequences $X$ and $Y$. The goal is to estimate the expectation $\mathbf{E}[f(X) - g(Y)]$, where $f$ and $g$ are real-valued cost or performance functions associated with the two systems. (Analogous remarks apply if one examines ratios of expectations instead of differences.) The effort required to obtain a valid estimate of this difference depends critically on the variance of $f(X) - g(Y)$, which is given by

$$\mathbf{Var}\,[f(X) - g(Y)] = \mathbf{Var}\,[f(X)] + \mathbf{Var}\,[g(Y)] - 2\,\mathbf{Cov}\,[f(X), g(Y)].$$

The first two terms on the right are determined by the individual distributions of $X$ and $Y$, which are fixed by the systems being modeled. But the last term is under the simulator's

* Accepted by James R. Wilson; received August 7, 1990. This paper has been with the authors 1 month for 2 revisions.

0025-1909/92/3806/0884$01.25

control. Simulating $X$ and $Y$ independently makes the covariance zero; deliberately introducing dependence changes the variance on the left. CRN attempts to reduce this variance by introducing positive dependence between $f(X)$ and $g(Y)$.

This raises two questions regarding the use of CRN: When does it work, and when is it optimal, in the sense that no other mechanism introduces greater positive dependence. For the simulator, these translate to two practical questions: If the same random numbers are used, will variance be reduced? Is this is the best one can do?

The analysis in this paper has two purposes. We first investigate the questions above for CRN in general; we then provide specific conditions under which they can be answered affirmatively. Our *general* analysis points out that the question of optimality is, in a sense, ill-posed. We argue that "common random numbers" is both too broad and too narrow for a meaningful answer, and so split the optimality question into two subquestions. In practice, it seems that one can only hope to establish optimality in a restricted (but useful) sense. Within this restricted notion of CRN, we also consider the role of *inversion* in generating random variables, and consider the problem of optimal *synchronization*—that is, the proper assignment of random "seeds" to random variables.

Our results based on *specific* conditions *guarantee* variance reduction for a class of systems and performance measures. They also *suggest* that CRN is advantageous for a broader class of performance measures and for systems in which our conditions are not grossly violated. In addition, these results provide guidelines for the implementation of CRN, partly validating simulation folklore on synchronization.

The specific conditions that lead to variance reduction are developed in three settings, corresponding to what we view as the most useful applications of CRN. These are as follows:

I. *Distributional comparisons*. Here we have in mind comparisons of essentially the same system driven by different stochastic inputs. As an example, consider the comparison of two single-server queueing systems differing only in their service time distributions.

II. *Structural comparisons*. This refers to comparisons of, for example, queues with the same service and interarrival times, but different buffer sizes, number of servers, etc.

III. *Sensitivity analysis*. By this we mean comparisons of systems that differ only through a small change in a continuous parameter. Think, for example, of a comparison of queues with service rates $\mu$ and $\mu + \epsilon$, with $\epsilon$ small.

The distinctions between these categories cannot be pushed too far; they are to some extent subjective. Nevertheless, the division is useful for the analysis and is meaningful in practice.

The benefit from CRN depends on properties of *monotonicity* (particularly in categories I and II) and *continuity* (particularly in III). While the importance of monotonicity is well known, the role of continuity seems to have been less well appreciated. Both properties are tied to the intuitive justification for CRN: comparing two systems by using the same input makes most sense if the systems respond similarly to changes in inputs (monotonicity), and if their outputs are close when their inputs are close (continuity). The conditions we propose simultaneously ensure monotonicity and continuity.

There is surprisingly little work linking the general observation that monotonicity is important in inducing positive correlation to specific guarantees for variance reduction. Some empirical results are reported in Wright and Ramsay (1979), and specific examples are considered in most simulation texts. Theoretical results for static systems (random vectors) are derived in Rubinstein and Samorodnitsky (1985) and Rubinstein, Samorodnitsky and Shaked (1985), based on notions of positive dependence. Other interesting methods of coupling random samples are developed in Devroye (1990), Schmeiser and Kachitvichyanukul (1986), and Shaked and Shanthikumar (1986).

The only general results for stochastic processes appear to be those of Heidelberger and Iglehart (1979), further discussed in Glynn (1985) and in Glynn and Iglehart (1988), Section 12. These apply when the system simulated is a *stochastically monotone Markov chain* in the sense of Daley (1968). Let us briefly review this setting. A Markov chain $X = \{X_n, n \geq 0\}$ on $\mathbf{R}$ with transition kernel $P$ is stochastically monotone if, for all $y \in \mathbf{R}$, $P(x,[y,\infty))$ is an increasing function of $x$. For any SMMC there are increasing functions $h_n : [0, 1]^{n+1} \mapsto \mathbf{R}$, $n = 0, 1, 2, \ldots$, such that if $(U_1, U_2, \cdots)$ is an i.i.d. sequence, each $U_i$ uniform on $[0, 1]$, then $\{h_n(U_1, \ldots, U_{n+1}), n \geq 0\}$ is equal in law to $\{X_n, n \geq 0\}$. (See the proof of Theorem 3.9 of Heidelberger and Iglehart.)

Suppose now that $X$ and $Y$ are SMMCs, and that $f$ and $g$ are increasing functions of $X$ and $Y$. If $X$ and $Y$ are generated from a single sequence $U$, then $(f(X), g(Y)) =_{\mathrm{st}} (\hat{f}(U), \hat{g}(U))$, for some increasing functions $\hat{f}$ and $\hat{g}$, where $=_{\mathrm{st}}$ denotes equality in distribution. Increasing functions of independent random variables are positively correlated (see, e.g., p. 31 of Barlow and Proschan 1975 for finite-dimensional $U$; the extension to an infinite sequence follows from Lindqvist 1988, p. 121). Thus, $\mathbf{Cov}\,[f(X), g(Y)] = \mathbf{Cov}\,[\hat{f}(U), \hat{g}(U)] \geq 0$. In other words, generating $X$ and $Y$ monotonically from the same $U$ is possible and advantageous.

Compared with the types of processes typically studied through simulation, the SMMCs form a restricted class. The Markov assumption itself is somewhat limiting (though the analysis of Heidelberger and Iglehart implicitly applies to more general monotone processes). While most simulated processes become Markov through an augmentation of the state, the inclusion of supplementary variables may destroy monotonicity. Since the class of processes satisfying the SMMC condition *exactly* is small, the result above is best viewed as a guideline, and suggests checking for rough monotonicity and blatant departures from monotonicity.

In the same spirit, we identify a different class of problems for which CRN is guaranteed to work and propose these, too, as guidelines for thinking about CRN. Our results complement those for SMMCs by taking a different point of view. The principal difference is that we look for monotonicity in the *event epochs* rather than in the sequence of states. This perspective has several important consequences:

- Since the "state" in many simulations is multi-dimensional and does not always have a meaningful ordering, monotonicity in the timing of events is often more natural.
- While conditions for positive dependence are traditionally given in purely probabilistic terms, our conditions are stated directly in terms of the structure of the simulated system, and are therefore based on information readily available to the simulator.
- Because our conditions are easy to understand, they are useful as guidelines even when they are not satisfied exactly.

It seems fair to say that if events never changed order the analysis of common random numbers would be a trivial matter. The question of whether or not CRN works is difficult primarily because, in most meaningful comparisons, the sequence of events may differ across systems on any given run. Not surprisingly, then, our conditions for guaranteed variance reduction restrict the possible effects of order changes. (A precise statement is given in Definition 3.1.) In a sense, the key "guideline" behind all our results is this: to check the benefit of CRN, look at what happens when events change order.

In §2 we discuss common random numbers in general; we consider the optimality of CRN, the role of inversion, and the problem of synchronization. §3 reviews the *generalized semi-Markov process* model of simulation and introduces the key properties on which subsequent results rely. In particular, it provides conditions under which it is possible to give explicit recursions for event epochs purely in terms of increasing, continuous functions. In §4 we use this structure to consider distributional comparisons (as described above) and verify variance reduction. §5 is a similar analysis of structural comparisons. In §6 we consider the application of common random numbers to sensitivity analysis and show that the special structure of §3 leads to an order of magnitude reduction in variance compared to CRN for "arbitrary" systems. The appendix contains all proofs.

Throughout this paper, "increasing," when applied to a function of vectors or sequences means "nondecreasing in the componentwise ordering." Unless otherwise stated, statements about continuity refer to the "product topology:" a mapping of $(x_1, x_2, \cdots)$ to $(f_1, f_2, \cdots)$ is continuous in this sense if $f_i^{(n)} \rightarrow f_i$, for all $i$, whenever $x_i^{(n)} \rightarrow x_i$ for all $i$. A right-continuous function is one which is continuous through every decreasing sequence of arguments.

Finally, a caveat: Our analysis assumes the availability of an ideal random number generator. We make no attempt to model the serial correlation in any real sequence of pseudorandom numbers. Such correlation may influence the implementation and performance of CRN.

## 2. General Considerations Regarding CRN

Before we can show that CRN works or is optimal in any specific settings we must investigate what this means in general. Our discussion points out that, in practice, CRN is at best optimal within a limited class of sampling schemes.

### 2.1. *Optimality of CRN*

Let $X$ and $Y$ be random objects with distributions $P_X$ and $P_Y$, taking values in sets $\mathbf{S}_X$ and $\mathbf{S}_Y$. These sets are essentially arbitrary; in particular, $X$ and $Y$ could be scalars, vectors, or processes. Let $f$ and $g$ be real-valued functions on $\mathbf{S}_X$ and $\mathbf{S}_Y$. In comparing $\mathbf{E}[f(X)]$ and $\mathbf{E}[g(Y)]$ through simulation using CRN, one might ask, Does generating $X$ and $Y$ with the same random numbers make $\mathbf{Cov}\ [f(X), g(Y)]$ positive? Does CRN maximize this covariance?

Without further elaboration, these questions (especially the second one) are meaningless. Of course, part of the problem is that we have not said anything about how $X$ and $Y$ are to be generated from random numbers. But the problem is deeper than that. A result in measure theory (see p. 327 of Royden 1968; see Whitt 1976 and Wilson 1983 for closely related applications) states that any "reasonable" probability space can be represented as the image of a measurable function on the unit interval with Lebesgue measure. This means that, in a precise sense, virtually any random object can be sampled by appropriately transforming a single uniform random variable, $U$. In particular, any joint distribution of $(X, Y)$ on $\mathbf{S}_X \times \mathbf{S}_Y$ (with marginals $P_X$ and $P_Y$) can be realized from a single $U$. Thus, any value of $\mathbf{Cov}\ [f(X), g(Y)]$ that can arise through some joint distribution of $(X, Y)$ can be realized using a common random number; the use of CRN in no way restricts the possible values of this covariance. In this sense, "common random numbers" is simply too general.

There is another sense in which CRN, if taken literally, is too narrow. The value of $\mathbf{Cov}\ [f(X), g(Y)]$ is determined by the joint distribution of $(X, Y)$, but does not otherwise depend on how $X$ and $Y$ are generated. Any sampling scheme that induces the same dependence between $X$ and $Y$ as CRN (with no additional effort) is just as good as CRN, even if it does not literally use common random numbers. We should not, therefore, restrict CRN to mean running different simulations with the same "seeds". We will return to this point in §2.2.

With these remarks in mind, let us focus the problem of CRN, defining what we see as the key practical and theoretical issues. Let $\mathcal{M}(X, Y)$ be the set of probability measures on $\mathbf{S}_X \times \mathbf{S}_Y$ with marginals $P_X$ and $P_Y$. These are the *admissible* joint distributions of $X$ and $Y$. The analysis of CRN is concerned with determining which distributions in $\mathcal{M}(X, Y)$ maximize $\mathbf{Cov}\ [f(X), g(Y)]$, and with finding ways of sampling from these distributions.

We obtain greater generality with little additional complexity if we consider the maximization problem

$$\sup_{\mathcal{M}(X,Y)} \mathbf{E}[\psi(X, Y)], \tag{1}$$

for some function $\psi : \mathbf{S}_X \times \mathbf{S}_Y \mapsto \mathbf{R}$. If we choose $\psi(x, y) = f(x)g(y)$, the solution to (1) maximizes the covariance of $f(X)$ and $g(Y)$.

The solution to (1) is known in significant generality when $X$ and $Y$ are random variables (i.e., real-valued). Recall that a function $\psi : \mathbf{R}^2 \mapsto \mathbf{R}$ is called *supermodular* if, whenever $x_1 \leq x_2$ and $y_1 \leq y_2$,

$$\psi(x_1, y_1) + \psi(x_2, y_2) \geq \psi(x_1, y_2) + \psi(x_2, y_1).$$

(This definition extends immediately to the case where $\mathbf{S}_X$ and $\mathbf{S}_Y$ are arbitrary partially ordered sets. We use this generalization below.) Suppose $X$ and $Y$ have distribution functions $F_X$ and $F_Y$. For any distribution $F$ on $\mathbf{R}$, define the inverse of $F$ by $F^{-1}(u) = \inf\{x : F(x) > u\}$. From Cambanis, Simons, and Stout (1976) we have

PROPOSITION 2.1. *Suppose that $\psi$ is right-continuous and supermodular, and that $U$ is uniformly distributed on* $[0, 1]$. *Then*

$$\sup_{\mathcal{M}(X,Y)} \mathbf{E}[\psi(X, Y)] = \mathbf{E}[\psi(F_X^{-1}(U), F_Y^{-1}(U))],$$

*assuming all expectations on the left exist and are finite. In other words,* (1) *is solved generating $X$ and $Y$ by inversion using a common random number.*

(Cambanis et al. use weaker conditions. Lorentz 1953 proved a slightly less general version of this result. He also showed that, in a precise sense, the supermodular functions form the largest class for which this result holds. This should not be surprising since the supermodular functions are just those which reward "alignment" of their arguments.)

If $f$ and $g$ are increasing functions from $\mathbf{R}$ to $\mathbf{R}$, then their product is supermodular on $\mathbf{R}^2$; hence, Proposition 2.1 shows how to maximize $\mathbf{Cov}\,[f(X), g(Y)]$ in this case. The element of $\mathcal{M}(X, Y)$ that attains the maximum in Proposition 2.1 is given by the distribution

$$\begin{aligned} H(x, y) = P(F_X^{-1}(U) \leq x, F_Y^{-1}(U) \leq y) &= P(U \leq F_X(x), U \leq F_Y(y)) \\ &= F_X(x) \wedge F_Y(y), \end{aligned} \tag{2}$$

as noted in Hoeffding (1940); see also Lehmann (1966), Cambanis et al., and Whitt (1976).

Thus, the solution to (1) is known when $X$ and $Y$ are random variables. Unfortunately, this is virtually the *only* case in which the solution is known. (See Rachev 1984 for a survey of existing results and related open problems.) Even if (1) could be solved when $X$ and $Y$ are stochastic processes, there would be no guarantee that sampling from the optimal joint distribution would be feasible. Thus, both theoretical and practical considerations lead us to narrow the problem.

In practice, the method of sampling each of $X$ and $Y$ may be determined by considerations (computational efficiency, ease of implementation) other than applicability to CRN. Hence, it is natural to look at the simpler problem in which the sampling algorithms are fixed, and the simulator merely controls the assignment of seeds. Let $U = (U_1, U_2, \cdots)$ and $V = (V_1, V_2, \cdots)$ be sequences of independent random variables, each $U_i$ and $V_i$ uniformly distributed on $[0, 1]$. Fix *sampling functions* $\Phi_X$, $\Phi_Y$ for which $\Phi_X(U) =_{\text{st}} X$ and $\Phi_Y(V) =_{\text{st}} Y$ ($X$ and $Y$ are now general). For example, if $X$ is an i.i.d. sequence, $\Phi_X$ could map $U$ to $X$ by mapping $U_i$ to $X_i$, $i = 1, 2, \ldots$. If $X$ is a dependent sequence, it can be generated by first sampling $X_1$, then $X_2$ given $X_1$, then $X_3$ given $X_1, X_2$, and so on. In this case, it is natural for each $X_i$ to be a function of $U_1, \ldots, U_i$. At this point, the details of $\Phi_X$ and $\Phi_Y$ do not concern us.

With the sampling functions fixed, the problem becomes choosing the joint distribution of $(U, V)$ to maximize $\mathbf{Cov}\,[f \circ \Phi_X(U), g \circ \Phi_Y(V)]$. Even this problem is a bit too general, so we add an additional constraint: we only allow dependence between corresponding elements of the sequences $U$ and $V$. In other words, we require that for all $n$, all $i_1, \ldots, i_n$, and all $u_{i_1}, v_{i_1}, \ldots, u_{i_n}, v_{i_n} \in [0, 1]$,

$$P(U_{i_j} \leq u_{i_j}, V_{i_j} \leq v_{i_j}, j = 1, \ldots, n) = \prod_{j=1}^{n} P(U_{i_j} \leq u_{i_j}, V_{i_j} \leq v_{i_j}). \tag{3}$$

Denote by $\mathcal{M}_0(U, V)$ the set of joint distributions of $(U, V)$ satisfying (3). As shown in the Appendix, Proposition 2.1 proves

PROPOSITION 2.2. *If* $\mathbf{E}[f^2(X)] < \infty$ *and* $\mathbf{E}[g^2(Y)] < \infty$, *and if* $f$, $g$, $\Phi_X$ *and* $\Phi_Y$ *are increasing right-continuous functions, then*

$$\sup_{\mathcal{M}_0(U,V)} \mathbf{Cov}\,[f \circ \Phi_X(U), g \circ \Phi_Y(V)] \tag{4}$$

*is attained by setting* $V = U$. *The maximizing element of* $\mathcal{M}_0$ *is defined by setting the jth factor on the right side of* (3) *equal to* $u_{i_j} \wedge v_{i_j} \equiv P(U_{i_j} \leq u_{i_j}, U_{i_j} \leq v_{i_j})$.

In practice, optimality in the sense of Proposition 2.2 seems to be the most one can hope for. Clearly, a joint distribution which is optimal for (4) may be only suboptimal for (1); the possible distributions $\{(\Phi_X(U), \Phi_Y(V)), (U, V) \in \mathcal{M}_0\}$ form a subset of $\mathcal{M}(X, Y)$. The practical distinction between these problems is the following: In (1), the simulator asks, What is the best way to sample $X$ and $Y$? In (4), the question is, Given algorithms for sampling $X$ and $Y$ from random number streams, what is the best way to allocate random numbers to the two simulations?

In §3 and §4, we define a general class of simulations for which standard choices of $\Phi_X$ and $\Phi_Y$ are, in fact, increasing and right-continuous. Hence, for these systems, comparisons using CRN are optimal, in the sense of Proposition 2.2.

### 2.2. *The Role of Inversion*

It is known that inversion plays a special role in CRN; one occasionally finds the recommendation that *only* inversion be used with CRN. Here, we look more closely at its role in problems (1) and (4).

Inversion is important because monotonicity and continuity are important, and inversion is closely related to these properties. Consider, first, monotonicity. Let $X$ be a random variable with distribution $F_X$. Consider the set $\mathcal{R}_X$ of "rearrangements of $X$"; i.e., the set of functions $\Phi_X : [0, 1] \mapsto \mathbf{R}$ for which $P(\Phi_X(U) \leq x) = F_X(x)$ whenever $U$ is uniform on $[0, 1]$ and $x \in \mathbf{R}$. Clearly, $F_X^{-1}$ is in $\mathcal{R}_X$; in fact, $F_X^{-1}$ is the *unique increasing* element of $\mathcal{R}_X$ (unique up to equality almost everywhere on $[0, 1]$). Since monotonicity of $\Phi_X$ is needed in Proposition 2.2, this alone would distinguish $F_X^{-1}$ among elements of $\mathcal{R}_X$. (The monotonicity of inversion is also the key to the analysis of Heidelberger and Iglehart of CRN with SMMCs.)

This observation allows us to reinterpret Proposition 2.1. Suppose we have decided *a priori* to use CRN and would like to know how best to implement it. We consider the optimization problem

$$\sup_{\Phi_X \in \mathcal{R}_X, \Phi_Y \in \mathcal{R}_Y} \mathbf{E}[\psi(\Phi_X(U), \Phi_Y(U))], \tag{5}$$

in which $U$ is uniform and $\psi$ is continuous and supermodular. From Proposition 2.1, we see that $F_X^{-1}$ and $F_Y^{-1}$ are the optimal choices. Thus, Proposition 2.1 states not only that CRN is optimal if inversion is used, but also that inversion is optimal if CRN is used. (This interpretation is a generalization of a rearrangement inequality of Hardy, Littlewood, and Pólya 1952, due to Lorentz. Whitt 1976 also considers maximal correlation via rearrangements.)

The second property mentioned above is continuity. This property is especially important when the performance functions $f$ and $g$ are, in fact, the same. It may happen—or we may imagine—that the random variables $X$ and $Y$ are connected through a sequence $X^{(0)}, X^{(1)}, \dots$, where $X^{(0)} = X$ and $\{X^{(n)}\}$ converges in distribution to $Y$. In this case, the comparison of $\mathbf{E}[f(X)]$ and $\mathbf{E}[f(Y)]$ becomes the limit of a sequence of comparisons. (Of course, this can always be achieved by letting every $X^{(n)}$, $n > 0$, have the same distribution as $Y$, but it is more meaningful to think of a sequence of "small" steps from $X$ to $Y$.) In making comparisons, it seems reasonable to require that the random variables be generated so that as the distributions of $\{X^{(n)}\}$ draw near to that of $Y$, the *values* draw near as well. (We justify this in §2.3 and §4.4.) Inversion achieves this goal: if $X^{(n)} \Rightarrow Y$ ($\Rightarrow$ denotes convergence in distribution) then $F^{-1}_{X^{(n)}}(U) \rightarrow F_Y^{-1}(U)$ with probability one, which is what we want. However, other rearrangements of $\{X^{(n)}\}$ and $Y$ may also provide continuity, so this property is not unique to inversion.

The above remarks, properly understood, indicate the importance of inversion in specifying joint distributions for simulation comparisons; they should not, however, be taken to imply that inversion itself must be used in implementation. For example, in Proposition 2.1, optimality is achieved by the distribution $H(x, y) = F_X(x) \wedge F_Y(y)$. This distribution is conveniently represented in terms of inversion and CRN, but any other method of sampling from $H$ would achieve the same covariance.

Suppose in comparing two systems we need to generate i.i.d. sequences (e.g., service times) $X = (X_1, X_2, \cdots)$ and $Y = (Y_1, Y_2, \cdots)$. Suppose we would like each pair $(X_i, Y_i)$ to have the distribution $H$. Using inversion and CRN is one option, but not the only one. We might generate each $X_i$ using any sampling scheme (e.g., acceptance-rejection), and then generate $Y_i$ given $X_i$. If $F_Y$ is easily inverted (but $F_X$ is not), we may conditionally sample $Y_i$ by setting $Y_i = F_Y^{-1}(F_X(X_i))$. If $F_X$ and $F_Y$ are continuous, then

$$P(X_i \le x, F_Y^{-1} \circ F_X(X_i) \le y) = H(x, y),$$

as is easily checked.

In some cases, inversion can be used to derive functional relations between random variables; the relation can then be used without inversion. This is the case when $X_i$ and $Y_i$ belong to a *scale* or *location* family, as noted in Glasserman (1988a) and Glynn and Iglehart (1988). Suppose that $F_X(\cdot) = F(\cdot, \theta_X)$ and $F_Y(\cdot) = F(\cdot, \theta_Y)$ for some collection $\{F(\cdot, \theta), \theta \in \Theta\}$ of distributions. This is a scale family if $F(x, \theta_2) = F(\theta_1 x/\theta_2, \theta_1)$, and a location family if $F(x, \theta_2) = F(x + \theta_1 - \theta_2, \theta_1)$, for all $x \in \mathbf{R}$ and all $\theta_1, \theta_2 \in \Theta$. In the first case, we may set $Y_i = \theta_Y X_i/\theta_X$, in the second case $Y_i = X_i - \theta_X + \theta_Y$, to achieve the joint distribution $H$ without necessarily using inversion.

Inversion has one additional property which does make it particularly convenient for the implementation of CRN (and which is often noted in the simulation literature): it requires exactly one uniform variate for each non-uniform variate generated. This makes programming for CRN particularly easy, and also simplifies the analysis of simulations driven by CRN. A method which always requires some fixed number (or at most some fixed number) of variates is almost as convenient. One could group the stream of random numbers into blocks and dedicate each block to a specified variate. Some methods, however, require a potentially unbounded number of uniform variates for each transformation; this is the case with acceptance-rejection, for example. CRN may be difficult to implement in comparing two simulations using such a method. However, if such a method is used to sample from $H(x, y)$ *across* simulations (e.g., as described above), then it is just as good as inversion and the fact that it uses a random number of variates is irrelevant.

In subsequent sections, we often specialize our results to the setting in which sequences of independent random variables are generated by inversion. This sometimes clarifies the results and may make them more immediately applicable. However, references to CRN with inversion should be understood as shorthand for sampling from a specified joint distribution. The results do not depend on literal use of inversion.

### 2.3. *Synchronization*

In the simulation literature one finds the recommendation that when CRN is used it should be implemented so that "corresponding" random variables across simulations are generated from the same random numbers. This is the issue of *synchronization*. Posing the problem this way presupposes that closely matching random numbers is advantageous, so let us formulate the question more generally.

Let $X$ and $Y$ be vectors or sequences. Once we have fixed sampling functions $\Phi_X$ and $\Phi_Y$, and have made the decision to use CRN, we must still decide how to assign random numbers across simulations—i.e., how to synchronize. Let $\Pi$ be the set of one-to-one functions $\pi$ mapping $\{1, 2, \cdots\}$ into itself. Think of elements of $\Pi$ as "permutations" of the positive integers. Given a sequence $U = (U_1, U_2, \cdots)$, denote by $U^\pi$ the sequence with $i$th element $U_{\pi(i)}$. If $\Phi_X(U) =_{\text{st}} X$, then $\Phi_X(U^\pi) =_{\text{st}} X$ as well, because $U^\pi =_{\text{st}} U$. Thus, we are free to compare $g \circ \Phi_Y(U)$ with $f \circ \Phi_X(U^\pi)$ for any $\pi \in \Pi$. The analysis of §2.1 provides no guidance in choosing $\pi$; if $\Phi_X$ is increasing then, as a function of $U$, $\Phi_X(U^\pi)$ is increasing, too. Any $\pi$ will make the covariance of $f(X)$ and $g(Y)$ positive, but which will maximize it? We are faced with the problem

$$\sup_{\pi \in \Pi} \mathbf{E}[\psi(\Phi_X(U^\pi), \Phi_Y(U))]. \tag{6}$$

As an example, consider a comparison of two single-server queueing systems. Suppose that $X$ is the sequence of interarrival and service times for one queue and $Y$ is the same sequence for the other queue. Suppose that all components of $X$ (and all components of $Y$) are independent, and let $X_i$ and $Y_i$ have marginal distributions $F_i$ and $G_i$. Suppose that the $i$th component of $\Phi_X(u)$ is given by $F_i^{-1}(u_i)$ and that of $\Phi_Y(u)$ by $G_i^{-1}(u_i)$. Let $f$ be an increasing, continuous function of $X$ or $Y$. From Proposition 2.2, we know that in estimating $\mathbf{E}[f(X) - f(Y)]$ we obtain variance reduction (compared with independent samples) by setting $X = \Phi_X(U)$ and $Y = \Phi_Y(U)$. However, for any $\pi \in \Pi$ we also obtain variance reduction by setting $X = \Phi_X(U^\pi)$ and $Y = \Phi_Y(U)$. Intuitively, it seems best to match seeds so that corresponding service and interarrival times in $X$ and $Y$ are generated from the same component of $U$; however, the results and discussion above do not by themselves justify this intuition.

Some justification is provided by the following result, which uses continuity. Suppose that $X$ and $Y$ take values in a common set $\mathbf{S}$, a complete, separable metric space. Let $X^{(0)} = X$ and let $\{X^{(n)}, n \geq 0\}$ be a sequence for which $X^{(n)} \Rightarrow Y$. Let $\Phi_{X^{(n)}}(U) =_{\text{st}} X^{(n)}$, $n = 0, 1, \ldots,$ and $\Phi_Y(U) =_{\text{st}} Y$, where $U$ is a sequence of independent, uniformly distributed random variables.

PROPOSITION 2.3. *Suppose that, for almost every* $u \in [0, 1]^\infty$, $\Phi_{X^{(n)}}(u) \to \Phi_Y(u)$ *as* $n \to \infty$. *Let* $f: \mathbf{S} \mapsto \mathbf{R}$ *be continuous. Suppose there exists an* $\epsilon > 0$ *such that* $\sup_{n \geq 0} \mathbf{E}[|f(X^{(n)})|^{2+\epsilon}] < \infty$. *Then* $\mathbf{Var}\,[f \circ \Phi_{X^{(n)}}(U) - f \circ \Phi_Y(U)] \to 0$ *as* $n \to \infty$. *Moreover, if* $\pi \in \Pi$ *is nontrivial, in the sense that* $\mathbf{Var}\,[f \circ \Phi_Y(U^\pi) - f \circ \Phi_Y(U)] \neq 0$, *then for all sufficiently large* $n$,

$$\mathbf{Var}\,[f \circ \Phi_{X^{(n)}}(U) - f \circ \Phi_Y(U)] < \mathbf{Var}\,[f \circ \Phi_{X^{(n)}}(U^\pi) - f \circ \Phi_Y(U)].$$

PROOF. The conclusion is an immediate consequence of the assumptions. Continuity of $f$ implies $f \circ \Phi_{X^{(n)}}(u) \to f \circ \Phi_Y(u)$ for almost every $u$, and the moment condition on $f(X^{(n)})$ implies that $\{[f \circ \Phi_{X^{(n)}}(U) - f \circ \Phi_Y(U)]^2, n \ge 0\}$ is uniformly integrable. Hence,

$$\lim_{n\to\infty} \mathbf{Var}\,[f \circ \Phi_{X^{(n)}}(U) - f \circ \Phi_Y(U)] = \mathbf{Var}\,[\lim_{n\to\infty} f \circ \Phi_{X^{(n)}}(U) - f \circ \Phi_Y(U)] = 0.$$

Similarly, for nontrivial $\pi$ we have $\lim_{n\to\infty} \mathbf{Var}\,[f \circ \Phi_{X^{(n)}}(U^\pi) - f \circ \Phi_Y(U)] = c$, for some $c > 0$. For all sufficiently large $n$,

$$\mathbf{Var}\,[f \circ \Phi_{X^{(n)}}(U) - f \circ \Phi_Y(U)] < c/2 < \mathbf{Var}\,[f \circ \Phi_{X^{(n)}}(U^\pi) - f \circ \Phi_Y(U)]. \qquad \square$$

Let us interpret this result in the comparison of two single-server queues discussed above. Consider a sequence of single-server queueing systems through which the service and interarrival times of one of the two systems under comparison converge in distribution to those of the other: for all $i = 1, 2, \ldots$, the sequence $\{F_i^{(n)}, n \ge 0\}$, $F_i^{(0)} = F_i$, converges to $G_i$ at all continuity points of $G_i$. If $\Phi_{X^{(n)}}$ and $\Phi_Y$ generate components of $X^{(n)}$ and $Y$ by inversion, then $X_i^{(n)} \to Y_i$, for all $i$, for almost every $u$. Hence, when the other hypotheses of the proposition are in force, we may conclude that assigning components of $U$ to corresponding components of $X$ and $Y$ beats any other (nontrivial) assignment if the distributions for the two queues are sufficiently close. The restriction to nontrivial permutations is needed to exclude the possibility that $\pi$ merely permutes elements of $U$ that do not affect $f \circ \Phi_Y(U)$.

Suppose, now, that we have two streams of random numbers $U^{(1)}$ and $U^{(2)}$. We use one for interarrival times and one for service times. In comparing two systems, we must decide whether to use the streams in the same way or, perhaps, to swap them. Proposition 2.3 suggests that the *standard* synchronization (using the streams the same way for both systems) is best if the service and interarrival times in the two systems are closer than the service times in either and the interarrival times in the other. Further support for the standard synchronization is given in subsequent sections.

For finite-horizon simulations—simulations of random vectors—a different approach to (6) and synchronization can be developed, based on *arrangement* orderings; see §6.F of Marshall and Olkin (1979). A function $f: [0, 1]^n \mapsto \mathbf{R}$ is called *arrangement increasing* if, for any $u_1, \ldots, u_n$ and $1 \le i < j \le n$, $f(u_1, \ldots, u_i, \ldots, u_j, \ldots, u_n) \ge f(u_1, \ldots, u_j, \ldots, u_i, \ldots, u_n)$ whenever $u_i \ge u_j$. For the following, let $\pi$ be any permutation of $\{1, \ldots, n\}$.

PROPOSITION 2.4. *Suppose $f$ and $g$ are arrangement increasing functions on* $[0, 1]^n$. *If the components of $U = (U_1, \ldots, U_n)$ are independent and uniform on* $[0, 1]$, *then* $\mathbf{Cov}\,[f(U), g(U)] \ge \mathbf{Cov}\,[f(U^\pi), g(U)]$ *for all $\pi$. In other words,* (6) *is attained by the identity permutation.*

This result shows that a sufficient condition for closely "matching" random numbers to be optimal is that the functions applied to them be arrangement increasing. The relevant functions are typically compositions of sampling and performance functions, so the arrangement increasing property may be difficult to satisfy or verify. It is unclear whether Proposition 2.4 can be applied to queueing systems with any generality. A simple example—the single-server queue—is given in §4.4.

## 3. Specially Structured Systems

We now turn to the investigation of a specific class of systems for which CRN is guaranteed to reduce variance and for which CRN is optimal in the sense of Proposition 2.2. To define this class of systems, we need to specify a precise model for discrete-event simulation.

### 3.1. *A Model of Simulation*

The most appropriate setting for our analysis is that of *generalized semi-Markov processes* or GSMPs. These processes are sufficiently general to model most systems studied through simulation, and their dynamics closely mimic those of event-driven simulations. (See Glynn and Iglehart for an overview of the role of GSMPs in analyzing simulation.)

A GSMP is defined in terms of a generalized semi-Markov *scheme*, which may be thought of as the structure of a simulation algorithm. A scheme is a 4-tuple $\mathcal{G} = (\mathbf{S}, \mathbf{A}, \mathcal{E}, p)$ where $\mathbf{S}$ is a set of *states* or system configurations; $\mathbf{A}$ is a finite set of *events*; $\mathcal{E}$ is a mapping from $\mathbf{S}$ to subsets of $\mathbf{A}$ with the interpretation that $\mathcal{E}(s)$ is the set of active events—the *event list*—in state $s$; and $p$ is a transition probability function: if $\alpha \in \mathcal{E}(s)$, then $p(s'; s, \alpha)$ is the probability that the process moves to state $s'$ from state $s$ upon the occurrence of event $\alpha$. Once the scheme is given, the stochastic description of a GSMP is completed by specifying an input process of *clock times*. This is a doubly-indexed sequence $\xi = \{\xi_\alpha(n), \alpha \in \mathbf{A}, n = 1, 2, \cdots\}$ with the interpretation that $\xi_\alpha(n)$ is the length of the $n$th "clock" or lifetime for event $\alpha$. For example, if $\alpha$ is an arrival or a service completion event, then $\xi_\alpha(n)$ is the $n$th interarrival time or service time. Naturally, $P(\xi_\alpha(n) \geq 0) = 1$ for all $\alpha$ and $n$.

We now describe how $\xi$ drives the evolution of the system. We construct a sequence $\{(Y_n, C_n), n \geq 0\}$ in which $Y_n$ is the $n$th state visited by the system and $C_n$ is the vector of residual clock times just after the $n$th transition. The sequence $\{(Y_n, C_n), n \geq 0\}$ is a general state-space Markov chain when the clock times are independent. Fix an initial state $s_0$ and let $Y_0 = s_0$. Initially, clocks are set for active events: if $\alpha \in \mathcal{E}(s_0)$ then $C_0(\alpha) = \xi_\alpha(1)$, and if $\alpha \notin \mathcal{E}(s_0)$ then $C_0(\alpha) = 0$. Among the elements of $\mathcal{E}(s_0)$, the event with the smallest clock time—call it $a_1$—is the first to occur, and it occurs at $C_0(a_1)$. (Use an arbitrary rule to break ties.) Upon the occurrence of $a_1$, the process moves to state $Y_1$ which is sampled from the probability mass function $p(\cdot; s_0, a_1)$. In the new state, the clock readings are adjusted: If $\alpha \in \mathcal{E}(Y_1) \cap [\mathcal{E}(Y_0) - \{a_1\}]$, then the clock for $\alpha$ continues to run in the new state, and its clock reading is given by $C_1(\alpha) = C_0(\alpha) - C_0(a_1)$. If $\alpha \in \mathcal{E}(Y_1) \setminus [\mathcal{E}(Y_0) - \{a_1\}]$, then a new clock must be set for $\alpha$. Thus, $C_1(\alpha) = \xi_\alpha(k + 1)$, where $k$ is the number of times a clock has previously been set for $\alpha$ (which, so far, is just $\mathbf{1}\{\alpha = a_1\}$). Finally, if $\alpha \in [\mathcal{E}(Y_0) - \{a_1\}] \setminus \mathcal{E}(Y_1)$, then $\alpha$ becomes inactive in the new state, $C_1(\alpha) = 0$, and the clock for $\alpha$ is said to be *interrupted*. By repeating this procedure we obtain $(Y_2, C_2)$ from $(Y_1, C_1)$, and so on.

If we let $\tau_n = \sum_{i=0}^{n-1} \min\{C_i(\alpha) : \alpha \in \mathcal{E}(Y_i)\}$, then $\tau_n$ is the epoch of the $n$th transition. We always assume that the system is *nonexplosive*, in the sense that, for all $s_0$, $P(\sup_{n>0} \tau_n = \infty) = 1$. This is a condition on $\xi$. With this assumption, the state of a GSMP $\{X_t, t \geq 0\}$ is defined by setting $X_t = Y_n$ for $\tau_n \leq t < \tau_{n+1}$.

From the evolution described above we also obtain a sequence $T = \{T_\alpha(n), \alpha \in \mathbf{A}, n = 1, 2, \cdots\}$, where $T_\alpha(n)$ is the epoch of the $n$th occurrence of event $\alpha$. Define $T_\alpha(n)$ to be infinity if $\alpha$ does not occur $n$ times. By convention, for every $\alpha \in \mathbf{A}$, $T_\alpha(0) = 0$ and $T_\alpha(\infty) = \infty$. If we define $D = \{D_\alpha(t), \alpha \in \mathbf{A}, t \geq 0\}$ by $D_\alpha(t) = \sup\{n \geq 0 : T_\alpha(n) \leq t\}$, then $D_\alpha(t)$ is the number of occurrences of $\alpha$ in $(0, t]$.

It is important to note that we have not placed any restrictions on the dependence among components of $\xi$, except (implicitly) that the dependence is determined *a priori* and is not affected by the evolution of the process. In practice, one often takes the sequences $\xi_\alpha$ and $\xi_{\alpha'}$ to be independent for $\alpha \neq \alpha'$, and the elements of each sequence $\{\xi_\alpha(n), n = 1, 2, \cdots\}$ to be i.i.d. with some distribution $F_\alpha$. Let us refer to this case as *the standard independent input*.

### 3.2. *Key Properties*

We begin by considering GSMPs based on *deterministic* schemes—i.e., schemes in which $p(s'; s, \alpha)$ takes only the values 0 and 1. For such schemes, we may define $\phi(s,$

$\alpha$) to be the unique state reached from $s$ upon the occurrence of $\alpha$, if $\alpha \in \mathcal{E}(s)$. Call a finite sequence of events $\beta_1 \cdots \beta_n$ *feasible* starting in a state $s_0$ if $\beta_{i+1} \in \mathcal{E}(s_i)$ for $i = 0, \ldots, n-1$, with $s_i = \phi(s_{i-1}, \beta_i)$. The feasible sequences are those that can arise as the sequence of actual events through some choice of $\xi$. A finite sequence of events is also called a *string* and denoted $\sigma$. If $\sigma$ is feasible starting in $s$, we write $\phi(s, \sigma)$ for the state reached from $s$ through the occurrence of the sequence $\sigma$. For simplicity, in the sequel we take the initial state $s_0$ to be fixed. Finally, for any string $\sigma$, let $N_\alpha(\sigma)$ be the number of occurrences of $\alpha$ in $\sigma$ and let $N(\sigma) = (N_\alpha(\sigma), \alpha \in \mathbf{A})$.

Our results are based on structural properties of schemes specified in

DEFINITION 3.1. Define the following three properties for deterministic schemes:

(i) *noninterruption*: for all $s \in \mathbf{S}$, all distinct $\alpha, \beta \in \mathbf{A}$, if $\alpha, \beta \in \mathcal{E}(s)$ then $\beta \in \mathcal{E}(\phi(s, \alpha))$;

(ii) *permutability*: for all feasible $\sigma$ and $\sigma'$, $N(\sigma) = N(\sigma')$ implies $\mathcal{E}(\phi(s, \sigma)) = \mathcal{E}(\phi(s, \sigma'))$;

(iii) *strong permutability*: for all feasible $\sigma$ and $\sigma'$, $N(\sigma) = N(\sigma')$ implies $\phi(s, \sigma) = \phi(s, \sigma')$.

The first property states that the occurrence of one event never interrupts the clock of another; a clock, once set, runs out at its scheduled time regardless of the occurrence of other events. The property of permutability states that permuting the order of events (while maintaining feasibility) does not change the event list of the state reached. Strong permutability is indeed stronger because it requires that permuting events not change the state reached.

In a queueing context, most nonpreemptive disciplines satisfy noninterruption, and most permutable schemes are in fact strongly permutable. For instance, the first-come-first-served, single-server queue is noninterruptive and strongly permutable. But permutability is incompatible with, for example, multiple job classes arriving in separate streams to a single queue: a change in the order of arrivals of different classes can change the event list reached. Examples are detailed in §4.2.

The conditions in Definition 3.1 have fundamental implications for the method of CRN; their significance derives from the following:

THEOREM 3.2. *Suppose the deterministic scheme $\mathcal{G}$ is noninterruptive and permutable. Then for every $\alpha \in \mathbf{A}$ and every $n = 1, 2, \cdots$ there exists a set of (nonrandom) indices $\{x^j_\beta(\alpha, n), \beta \in \mathbf{A}, j = 1, \ldots, J\}$, for some finite $J$ (depending on $\alpha$ and $n$), such that for all $\xi$*

$$T_\alpha(n) = \xi_\alpha(n) + \min_{1\le j\le J} \max_{\beta\in\mathbf{A}} \{T_\beta(x^j_\beta(\alpha, n))\}. \tag{7}$$

*It follows that $T$ is an increasing, continuous function of $\xi$.*

Instances of (7) are given for specific examples in §4.2. For the single-server queue, (7) is especially simple so we display it here. Let $\alpha$ denote arrival, let $\beta$ denote service completion and let the system be empty initially. The corresponding recursions are thus:

$$T_\alpha(n) = \xi_\alpha(n) + T_\alpha(n-1); \tag{8}$$

$$T_\beta(n) = \xi_\beta(n) + \max\{T_\alpha(n), T_\beta(n-1)\}. \tag{9}$$

To see this, observe that the $n$th interarrival time starts at the $(n-1)$st arrival, the $n$th service time starts at either the $n$th arrival or the $(n-1)$st departure, whichever is later.

That noninterruption and permutability imply (7) is established in Glasserman and Yao (1992b), so here we only sketch the argument. In a noninterruptive scheme, $T_\alpha(n) - \xi_\alpha(n)$ is the epoch of the $n$th setting of a clock for $\alpha$. Denote this epoch by $S_\alpha(n)$; we argue that $S_\alpha(n)$ is given by the second term on the right side of (7). Suppose that $T_\alpha(n) < \infty$ and consider the set of strings $\Sigma$ leading to the $n$th activation of $\alpha$; i.e., $\sigma \in \Sigma$ if and only if $N_\alpha(\sigma) = n - 1$ and $\alpha \in \mathcal{E}(\phi(s_0, \sigma))$. Let $N(\Sigma) = \{N(\sigma), \sigma \in \Sigma\}$. Call an element $x$ of $N(\Sigma)$ *minimal* if for any other element $y \in N(\Sigma)$ we have $y_\beta > x_\beta$ for at least one $\beta \in \mathbf{A}$. Let each $x^j(\alpha, n) = \{x^j_\beta(\alpha, n), \beta \in \mathbf{A}\}$ be a minimal element of $N(\Sigma)$. Notice that each $x^j(\alpha, n)$ is a deterministic vector not depending on the outcome of $\xi$. Permutability implies that $\alpha$ is activated for the $n$th time by time $t$ if and only if $D(t)$ dominates some element of $N(\Sigma)$. Hence, $S_\alpha(n)$ is the smallest $t$ for which $D(t)$ dominates some element of $N(\Sigma)$, and therefore the smallest $t$ for which $D(t)$ dominates some minimal element of $N(\Sigma)$. This explains the "min" in (7). For $D(t)$ to dominate some $N(\sigma)$, *every* $D_\beta(t)$, $\beta \in \mathbf{A}$, must be greater than or equal to $N_\beta(\sigma)$; i.e., $D(t)$ dominates $x^j(\alpha, n)$ if and only if $t \geq \max_\beta \{T_\beta(x^j_\beta(\alpha, n))\}$. This explains the "max" in (7).

The second part of the theorem follows from (7). Let $\xi$ and $\xi'$ be two realizations of the clock time process. Write $\xi \leq \xi'$ if, for every $\alpha$ and $n$, $\xi_\alpha(n) \leq \xi'_\alpha(n)$. Then the mapping from $\xi$ to $T$ given by (7) is increasing because $\xi \leq \xi'$ implies $T \leq T'$, in the same componentwise ordering. (Min, max and addition are increasing functions.) The mapping is also continuous: if $\xi^{(k)}_\alpha(n) \to \xi_\alpha(n)$ as $k \to \infty$ for every $\alpha$ and $n$, then, for the resulting epochs $T^{(k)}$ and $T$, $T^{(k)}_\alpha(n) \to T_\alpha(n)$ as $k \to \infty$, for every $\alpha$ and $n$. This holds because min, max and addition are continuous.

REMARKS. (i) If some event $\alpha$ can never occur $n$ or more times, we can drop the "min" in (7) and set $x_\beta(\alpha, n) = \infty$ for all $\beta \neq \alpha$. (Recall our convention that $T_\beta(\infty) = \infty$ for all $\beta$.) In an *irreducible* scheme (i.e., one in which every state can be reached from any other state through some sequence of events) all events can occur infinitely many times, so every index appearing on the right side of (7) is finite. Many commonly simulated systems are irreducible.

(ii) Most of our results follow from (7), so we could simply take that representation as our starting point. Recursions like (7) are known for specific systems; see, e.g., Baccelli, Massey, and Towsley (1989), Tsoucas and Walrand (1989), and see Greenberg, Lubachevsky, and Mitrani (1990) for an application to parallel simulation. Our results do not depend on Theorem 3.2 except through (7), so they hold whenever such a recursion is available. But Definition 3.1 does provide a convenient set of "primitive" conditions that ensure the structure of (7). It also underscores the connection between monotonicity, continuity and changes in the order of events.

## 4. Distributional Comparisons

We now apply the structure of Theorem 3.2 to CRN. We consider two GSMPs based on the same scheme but driven by different clock processes. We show that if the scheme is permutable and non-interruptive, variance reduction is achieved using CRN.

### 4.1. *Guaranteed Variance Reduction*

Let $\xi^{(1)}$ and $\xi^{(2)}$ be alternative inputs to the same scheme, and let $T^{(1)}$ and $T^{(2)}$ be the resulting event epochs. Let $\Phi^{(1)}$ and $\Phi^{(2)}$ generate $\xi^{(1)}$ and $\xi^{(2)}$ from a sequence of i.i.d. uniform random variables. Our first result is relevant to finite-horizon comparisons. It is an immediate consequence of Proposition 2.2.

THEOREM 4.1. *Suppose the deterministic scheme $\mathcal{G}$ is noninterruptive and permutable. Suppose that $\Phi^{(1)}$ and $\Phi^{(2)}$ are increasing, right-continuous functions of sequences $U$ and $V$, and that $f$ and $g$ are increasing, right-continuous functions of $T^{(1)}$ and $T^{(2)}$ for which $f(T^{(1)})$ and $g(T^{(2)})$ have finite second moments. Then generating $\xi^{(1)}$ and $\xi^{(2)}$ with $U = V$ minimizes* $\mathbf{Var}\,[f(T^{(1)}) - g(T^{(2)})]$ *among all joint distributions of $(U, V)$ in $\mathcal{M}_0(U, V)$. In particular, it minimizes* $\mathbf{Var}\,[T^{(1)}_\alpha(n) - T^{(2)}_\alpha(n)]$ *for all $\alpha$ and $n$.*

An important special case is that of the standard independent input described at the end of §3.1. By simply relabeling, we may take $U$ to be a doubly-indexed sequence $\{U_\alpha(n), \alpha \in \mathbf{A}, n = 1, 2, \cdots\}$. Suppose $\Phi$ transforms $U$ to $\xi$ by setting $\xi_\alpha(n)$

$= F_\alpha^{-1}(U_\alpha(n))$ and that $\Phi^{(2)}$ works analogously. These maps are monotone increasing and right-continuous. Call this case the standard independent input with inversion.

COROLLARY 4.2. *In the case of standard independent input with inversion, $\Phi^{(1)}$ and $\Phi^{(2)}$ are automatically increasing and right-continuous. Hence, CRN is optimal, in the sense of Theorem* 4.1, *if the other conditions of the theorem are in effect.*

We can weaken the assumption of independent inputs if we assume that $\xi_\alpha$ and $\xi_\beta$ are independent, for $\alpha \neq \beta$, and that, for each $\alpha \in \mathbf{A}$, $\{\xi_\alpha(n), n = 1, 2, \cdots\}$ is a *conditionally increasing sequence*; i.e., for all $n$, and all $x_n$,

$$P(\xi_\alpha(n) \geq x_n \mid \xi_\alpha(1) = x_1, \ldots, \xi_\alpha(n-1) = x_{n-1})$$

is increasing in $(x_1, \ldots, x_{n-1})$. (If $\{\xi_\alpha(n), n = 1, 2, \cdots\}$ is a stochastically monotone Markov chain, it is a conditionally increasing sequence.) Using the construction described by Rubinstein et al. (1985), for such a sequence it is possible to represent each $\xi_\alpha(n)$ as an increasing, right-continuous function of $U_\alpha(1), \ldots, U_\alpha(n)$. Thus, we have

COROLLARY 4.3. *The conclusion of Corollary* 4.2 *holds if $\{\xi_\alpha^{(1)}, \alpha \in \mathbf{A}\}$ and $\{\xi_\alpha^{(2)}, \alpha \in \mathbf{A}\}$ are sets of independent conditionally increasing sequences.*

Finally, since $D = \{D_\alpha(t)\}$ is monotone in $\xi$ whenever $T$ is, we also have

COROLLARY 4.4. *Results* 4.1–4.3 *hold with $T^{(1)}$ and $T^{(2)}$ replaced by $D^{(1)}$ and $D^{(2)}$.*

In order to establish analogous results for steady-state simulations, we need to make some assumptions about the behavior of $T_\alpha(n)$ as $n$ grows. We assume that, for all $\alpha$, there are finite constants $m_\alpha^{(1)}$ and $m_\alpha^{(2)}$ such that

$$\frac{1}{n} T_\alpha^{(i)}(n) \to m_\alpha^{(i)}, \qquad \text{in probability as } n \to \infty \qquad \text{for} \qquad i = 1, 2. \tag{10}$$

Each $m_\alpha^{(i)}$ is the asymptotic *cycle time* of $\alpha$, the steady-state mean time between occurrences of $\alpha$. If $m_\alpha^{(i)} > 0$, then $1/m_\alpha^{(i)} = \lim_{t\to\infty} D_\alpha(t)/t$ is the asymptotic *throughput* of $\alpha$.

Denote by $\mathcal{M}_1(U, V)$ the set of elements of $\mathcal{M}_0(U, V)$ for which there exist finite constants $\sigma_\alpha$, $\alpha \in \mathbf{A}$, such that

$$n^{-1/2}[T_\alpha^{(1)}(n) - T_\alpha^{(2)}(n) - n(m_\alpha^{(1)} - m_\alpha^{(2)})] \Rightarrow \sigma_\alpha \mathcal{N}(0, 1), \tag{11}$$

when $\xi^{(1)}$ is generated from $U$ and $\xi^{(2)}$ from $V$. (In (11), $\mathcal{N}(0, 1)$ denotes a standard normal random variable.) This assumption of asymptotic normality is broadly applicable in practice. Finally, we also need to assume that

$$\left\{\frac{1}{n}[T_\alpha^{(i)}(n) - nm_\alpha^{(i)}]^2, n > 0\right\} \quad \text{is uniformly integrable,} \qquad i = 1, 2. \tag{12}$$

See, e.g., p. 32 of Billingsley (1968) for the definition and implications of uniform integrability. We now have

THEOREM 4.5. *Consider a GSMP based on a noninterruptive, permutable, deterministic scheme. Suppose that $\xi^{(1)}$ and $\xi^{(2)}$ are generated monotonically and right-continuously from $U$ and $V$. Suppose that* (10)–(12) *hold and that the joint distribution obtained by setting $U = V$ is in $\mathcal{M}_1$. Then CRN achieves maximal variance reduction* (*the smallest $\sigma_\alpha^2$*) *among elements of $\mathcal{M}_1$.*

It is a straightforward matter to rephrase this result in terms of *functions* of the $T_\alpha^{(i)}(n)$'s. Conclusions along the lines of Corollaries 4.2–4.4 are also obtained from Theorem 4.5 with obvious modifications. In particular, a similar result applies to throughputs.

Let us now briefly indicate how the conditions of noninterruption and permutability extend from deterministic to *probabilistic* schemes; i.e., schemes in which $p(s'; s, \alpha)$ need not be zero or one. Noninterruption is easy: we simply require that if $\{\alpha, \beta\} \subseteq \mathscr{E}(s)$, then $\beta \in \mathscr{E}(s')$ for all $s'$ such that $p(s'; s, \alpha) > 0$. The generalization of permutability is more complicated and depends on a GSMP definition of *state-independent routing* formulated in Glasserman and Yao (1992a). This definition is somewhat tedious but coincides with the usual sense of state-independent routing in queueing systems. Rather than review the concept in detail, we provide a brief description.

A GSMP has state-independent routing if for each $\alpha \in \mathbf{A}$ and for every $s_1, s_2 \in \mathbf{S}$ such that $\alpha \in \mathscr{E}(s_1)$ and $\alpha \in \mathscr{E}(s_2)$, the possible transitions out of $s_1$ and $s_2$ due to $\alpha$ are in one-to-one correspondence, with corresponding transitions having equal probabilities. For such a GSMP, it is possible to define a random sequence $\nu = \{\nu_\alpha(n), \alpha \in \mathbf{A}, n = 1, 2, \cdots\}$ such that $\nu_\alpha(n)$ determines the state transition deterministically upon the $n$th occurrence of $\alpha$. In this case, we may write $\phi(s, \alpha, \nu_\alpha(n))$ for the state reached from $s$ if the $n$th occurrence of $\alpha$ occurs in $s$. We now require that $\phi$ satisfy the condition of permutability for all outcomes of $\nu$. When this holds, we obtain a representation like (7) in which the indices $x^j(\alpha, n)$ depend on $\nu$ but not $\xi$. A queueing network with probabilistic routing is considered in Example 4.9 below.

## 4.2. *Examples*

We now discuss some simple examples that do and do not satisfy noninterruption and permutability. In verifying the second of these, it is useful to note that if the scheme is noninterruptive and if, for all $s$, and all distinct $\alpha$, $\beta$

$$\alpha, \beta \in \mathscr{E}(s) \Rightarrow \phi(s, \alpha\beta) = \phi(s, \beta\alpha), \tag{13}$$

then the scheme is strongly permutable, hence permutable. Property (13) is the *commuting condition* of Glasserman (1988b).

EXAMPLE 4.6. Consider the single-server queue. The state space is $\mathbf{S} = \{0, 1, 2, \cdots\}$, the set of possible queue lengths. Let $\mathbf{A} = \{\alpha, \beta\}$ where $\alpha$ denotes arrival and $\beta$ denotes service completion. If $s = 0$ then $\mathscr{E}(s) = \{\alpha\}$ and (13) is satisfied vacuously. If $s > 0$, then $\mathscr{E}(s) = \{\alpha, \beta\}$ and $\phi(s, \alpha\beta) = s = \phi(s, \beta\alpha)$. From Theorem 3.2, it follows that all arrival and service completion epochs are increasing, continuous functions of the interarrival and service times. Explicit recursions were given in (8–9).

EXAMPLE 4.7. Consider $k$ single-server queues in tandem. The first has an infinite buffer, all the others may be finite or infinite. The events are arrival to the first queue, $\alpha$, and service completion at the $i$th queue, $\beta_i$, $i = 1, \ldots, k$. Various forms of blocking are possible. In *manufacturing* blocking, if the buffer at $i + 1$ is full upon the completion of service at $i$, the completed job waits at $i$ (preventing the next initiation of service) until room becomes available at $i + 1$. This system is noninterruptive and permutable, as shown in Glasserman and Yao (1992a). One may verify that, in fact, (13) holds. The arrival epochs follow (8). Suppose there is room for $b_i$ jobs at queue $i$ (including the server); then the service-completion epochs follow these recursions:

$$T_{\beta_1}(n) = \xi_{\beta_1}(n) + \max\{T_\alpha(n), T_{\beta_1}(n-1), T_{\beta_2}(n - b_2 - 1)\};$$

$$T_{\beta_i}(n) = \xi_{\beta_i}(n) + \max\{T_{\beta_{i-1}}(n), T_{\beta_i}(n-1), T_{\beta_{i+1}}(n - b_{i+1} - 1)\},$$

$$T_{\beta_k}(n) = \xi_{\beta_k}(n) + \max\{T_{\beta_{k-1}}(n), T_{\beta_k}(n-1)\}.$$

$$i = 2, \ldots, k-1;$$

Several modifications of this system, including queues with *communication* or *kanban* blocking, also satisfy (13) and lead to slightly modified recursions.

EXAMPLE 4.8. In this example permutability is violated. Consider a queue fed by two classes of arrivals $\alpha_1$ and $\alpha_2$. Let $\beta_1$ and $\beta_2$ be the corresponding service completion events. Changing the order of events in this system changes the event list reached. For example, suppose the queue is initially empty. The sequence $\alpha_1\alpha_2$ makes the event list $\{\alpha_1, \alpha_2, \beta_1\}$ (the class 1 job arrives first and goes into service), but the sequence $\alpha_2\alpha_1$ makes the event list $\{\alpha_1, \alpha_2, \beta_2\}$. Intuitively, we would expect that making the interarrival times for class 1 jobs shorter would *delay* the departures, $\beta_2$,. of class 2 jobs. Thus, in comparing systems with different interarrival processes $\xi_{\alpha_1}$ and $\xi_{\alpha_2}$, variance reduction is not guaranteed if CRN is used. This is especially true if the departure times of class 1 jobs in one system are compared with those of class 2 jobs in another. Since there is some negative correlation between these epochs—speeding up one class of jobs will slow down the other—CRN may actually increase variance. If we give one class of jobs preemptive priority over the other, we violate noninterruption and compound the departure from monotonicity.

If, in the original system, the two classes of jobs have the same service time distribution and if we do not distinguish between service completions of the two classes, then the system becomes permutable. For example, if $\beta$ denotes a service completion of either class then the event list is $\{\alpha_1, \alpha_2, \beta\}$ following the arrival of a class 1 and class 2 job, regardless of their order. For this modification, the recursions become

$$T_{\alpha_i}(n) = \xi_{\alpha_i}(n) + T_{\alpha_i}(n-1), \qquad i = 1, 2;$$

$$T_\beta(n) = \xi_\beta(n) + \min_{j=0,\ldots,n} \max\{T_{\alpha_1}(j), T_{\alpha_2}(n-j), T_\beta(n-1)\},$$

with the convention that $T_{\alpha_i}(0) = 0$, $i = 1, 2$.

EXAMPLE 4.9. To illustrate how our results can be applied to probabilistic schemes, we consider a closed network of single-server, infinite-buffer queues with Markovian routing. Denote a typical state by $\mathbf{n} = (n_1, \ldots, n_M)$, where $n_i$ is the number of jobs at node $i$ and $M$ is the number of nodes. Let $\beta_i$ denote service completion at the $i$th node. If routing is governed by a routing matrix $(P_{ij})$, then $p(\mathbf{n} - e_i + e_j; \mathbf{n}, \beta_i) = P_{ij}$, where $e_i$ and $e_j$ are the $i$th and $j$th unit vectors. Let $\nu_{\beta_i}(n)$ take the value $j$ with probability $P_{ij}$, $j = 1, \ldots, M$. Then we may define $\phi(\mathbf{n}, \beta_i, \nu) = \mathbf{n} - e_i + e_\nu$. It is easy to see that this system is noninterruptive. It is also permutable because for all $\nu_i$ and $\nu_j$, if $P_{i\nu_i} > 0$ and $P_{j\nu_j} > 0$ then $\phi(\phi(\mathbf{n}, \beta_i, \nu_i), \beta_j, \nu_j) = \phi(\phi(\mathbf{n}, \beta_j, \nu_j), \beta_i, \nu_i)$. In other words, changing the order in which jobs move from nodes $i$ and $j$ does not change the resulting state, provided the same routing decisions are made in both cases.

The examples above carry over to multiple-server queues, provided all servers at a particular queue are identical and we do not distinguish among departures from different servers at the same queue. Verifying this requires a minor modification of permutability tailored to *clock multiplicity*—an extension of the usual GSMP framework—developed in Glasserman and Yao (1992a). In effect, clock multiplicity forces a GSMP analog of the following simulation rule: let the multiple servers at each queue draw service times from a common stream. The details are in Glasserman and Yao (1992a).

### 4.3. *Other Performance Measures*

Thus far, we have only considered throughput-like performance measures—i.e., quantities defined purely in terms of event epochs. But event epochs are also building blocks for more general performance measures, and it is often possible to extend monotonicity results by examining how more general measures change with the event epochs. Our goal here is not to give the most general results possible, but rather to illustrate this idea with two examples.

*Sojourn Times*. Consider the time spent by jobs in a single-server queue. (Queues in tandem work similarly; a single queue makes the discussion simpler.) Suppose the queue starts empty: $s_0 = 0$. The sojourn time $W_n$ of the $n$th job, including its time in service, is $T_\beta(n) - T_\alpha(n)$ (in the notation of Example 4.6). Therefore, $W_n$ is increasing in $\xi_\beta$ (the service times), decreasing in $\xi_\alpha$ (the interarrival times), and continuous in both. The same is true of $\bar{W}_N = W_1 + \cdots + W_N$, for any $N = 1, 2, \ldots$. For simplicity, consider the case of standard independent input. Let $U^{(i)} = \{(U_\alpha^{(i)}(n), U_\beta^{(i)}(n)), n = 1, 2, \cdots\}$, $i = 1, 2$, and suppose $\Phi^{(i)}$ generates $\xi^{(i)}$, $i = 1, 2$, by setting $\xi_\beta^{(i)}(n) = F_\beta^{(i)^{-1}}(U_\beta^{(i)}(n))$ and $\xi_\alpha^{(i)}(n) = F_\alpha^{(i)-1}(1 - U_\alpha^{(i)}(n))$. Suppose $F_\alpha^{(i)}, F_\beta^{(i)}$, $i = 1, 2$, are all strictly increasing. This makes $\xi_\beta^{(i)}$ an increasing, continuous function of $U_\beta^{(i)}$, and $\xi_\alpha^{(i)}$ a decreasing, continuous function of $U_\alpha^{(i)}$. Hence, it makes $\bar{W}_N^{(i)}$ an increasing, continuous function of $U^{(i)}$. Proposition 2.2 is therefore immediately applicable in comparing $\bar{W}_N^{(1)}$ and $\bar{W}_N^{(2)}$.

To guarantee variance reduction for steady-state comparisons, we need analogs of (10)–(12). Fix $\xi^{(i)}$, $i = 1, 2$, and let $\{W_n^{(i)}, n \geq 0\}$ be the associated sojourn time sequences. Suppose there are constants $w^{(i)}$, $i = 1, 2$, such that

$$\frac{1}{n}\bar{W}_n \twoheadrightarrow w^{(i)} \quad \text{in probability}, \qquad i = 1, 2. \tag{14}$$

Let $\mathcal{M}_1$ be the set of joint distributions of $(U^{(1)}, U^{(2)})$ for which (11) holds with $T_\alpha^{(i)}(n)$ replaced by $W_n^{(i)}$, $m_\alpha^{(i)}$ replaced by $w^{(i)}$, and $\sigma_\alpha$ replaced by some $\sigma_w$. The following is a consequence of Theorem 4.5.

COROLLARY 4.10. *Suppose that* $\xi^{(i)}$, $i = 1, 2$, *are generated as described above, that the joint distribution obtained by setting* $U^{(1)} = U^{(2)}$ *is in* $\mathcal{M}_1$ *and that* $\{n^{-1}[\bar{W}_n^{(i)} - nw^{(i)}]^2, n > 0\}$ *is uniformly integrable for* $i = 1, 2$. *Then CRN achieves maximal variance reduction* (*minimizes* $\sigma_w^2$) *among elements of* $\mathcal{M}_1$.

Heidelberger and Iglehart establish variance reduction (though not optimality) for this example using the fact that $\{W_n, n \geq 0\}$ is an SMMC. Our analysis holds in the case of dependent $\xi$ and sojourn times through, e.g., a serial subnetwork, where the Markov assumption is violated.

*Queue Lengths.* Recall that $\{X_t, t \geq 0\}$ is the state of the GSMP. In the case of a single-server queue, $X_t$ is just the queue length at time $t$. In a network, by suitable choice of $f$ we can make $\{f(X_t), t \geq 0\}$ the queue-length process at a particular node. Suppose, more generally, that $\mathbf{S}$ is partially ordered and that $f: \mathbf{S} \mapsto \mathbf{R}$ is increasing with respect to this partial order. Often, $f(X_t)$ is an increasing function of some subset of sequences $\xi_\alpha$, $\alpha \in \mathbf{A}$, and a decreasing function of others. For example, in tandem queues (using the notation of Example 4.5) the queue length at the $i$th node is an increasing function of $\xi_{\beta_j}$, $j \geq i$ and a decreasing function of $\xi_\alpha$ and $\xi_{\beta_j}$, $j < i$. Through appropriate transformations we can therefore make $f(X_t)$ an increasing function of $U = \{U_\alpha(n), \alpha \in \mathbf{A}, n = 1, 2, \cdots\}$ for all $f$.

Now consider two systems with inputs $\xi^{(1)}$ and $\xi^{(2)}$. Let $\{X_t^{(i)}, t \geq 0\}$ be the corresponding state processes. Suppose we want to compare

$$\frac{1}{t}\int_0^t f(X_s^{(i)})ds, \qquad i = 1, 2, \tag{15}$$

as $t \twoheadrightarrow \infty$. If we assume the existence of these limits, as in (10), asymptotic normality, as in (11), and uniform integrability, as in (12), then common random numbers is optimal:

COROLLARY 4.11. *Suppose that, for all* $t \geq 0$, $f(X_t)$ *is increasing and right-continuous in* $U$. *Suppose that* (11) *and* (12) *hold when applied to* (15) *and that* $U^{(1)} = U^{(2)}$ *is in the corresponding class* $\mathcal{M}_1$. *Then CRN achieves maximal variance reduction among elements of* $\mathcal{M}_1$ *for comparison of* (15).

Corollary 4.11 applies, for example, to acyclic open queueing networks with a single class of jobs and infinite buffers by applying the device described above for the single-server queue. When service rates are exponential, Shanthikumar and Yao (1986) show that the queue lengths in a *closed* network are monotone in each service rate. Using inversion to generate service times, this makes queue lengths monotone in the $U_\alpha(n)$'s. For more general networks, monotonicity of $f(X_t)$ may fail. Note, however, that the corollary remains true if monotonicity of $f(X_t)$ is replaced by the weaker assumption that the time-average of $f(X_s)$ over $[0, t]$ (as in (15)) is increasing in $U$.

## 4.4. *Synchronization*

As mentioned in §2.3, CRN folklore recommends closely matching random numbers to corresponding events across simulations; see especially the discussion in Bratley, Fox, and Schrage (1987). The continuity of the event epochs implied by (7) allows us to apply Proposition 2.3 in support of the conventional wisdom. Suppose that $\{\xi^{(k)}, k \geq 0\}$ converges in law to $\xi$ as $k \to \infty$, and suppose that $\Phi^{(k)}$ generates $\xi^{(k)}$ and $\Phi$ generates $\xi$ so that $\xi^{(k)} \to \xi$ with probability one.

THEOREM 4.12. *Let $\{\xi^{(k)}\}$ and $\xi$ be as above. Suppose that, for some $\epsilon > 0$, $\sup_{k\geq 0} \mathbf{E}[(\xi_\alpha^{(k)}(n))^{2+\epsilon}] < \infty$, for all $\alpha$ and $n$. If the scheme is noninterruptive and permutable, then* $\mathbf{Var}\,[T_\alpha^{(k)}(n) - T_\alpha(n)] \to 0$ *as $k \to \infty$ for all $\alpha$ and $n$ for which every $x_\beta^i(\alpha, n)$ in (7) is finite.*

Let us interpret this result in the setting of the standard independent input with inversion; i.e., $\xi_\alpha^{(k)}(n)$ and $\xi_\alpha(n)$ are generated by inversion from $F_\alpha^{(k)}$ and $F_\alpha$ using $U_\alpha(n)$. Almost sure convergence of the clock times follows from convergence of the corresponding distributions. Under the moment conditions in Theorem 4.12, the variance of $T_\alpha^{(k)}(n) - T_\alpha(n)$ vanishes as $k \to \infty$. Moreover, Proposition 2.3 shows that this synchronization is asymptotically optimal; it eventually beats any assignment of seeds that permutes some of the $U_\alpha(n)$'s. In practice, this means keeping a separate stream $\{U_\alpha(n), n > 0\}$ of random numbers for each event—a recommendation often encountered in the simulation literature.

REMARK. A referee points out that, for fixed $k$, $T_\alpha^{(k)}(n)$ typically satisfies a central limit theorem, as $n \to \infty$, with variance proportional to $n$. This suggests that the rate at which the variances in Theorem 4.12 go to zero decreases with $n$. However, for the mean time between occurrences of an event, $T_\alpha^{(k)}(n)/n$, the effectiveness of common random numbers does not appear to diminish as $n \to \infty$.

The monotonicity implied by (7) provides a different justification, in a special class of comparisons, for assigning a separate stream $\{U_\alpha(n), n > 0\}$ to each event. Recall that one distribution $F$ is stochastically smaller than another distribution $G$, if $F(x) \geq G(x)$ for all $x$. Denote this relation by $F \leq_{\text{st}} G$.

THEOREM 4.13. *Consider a noninterruptive, permutable scheme driven by standard independent input with inversion. Suppose that, for every $\alpha$, $F_\alpha^{(2)} \leq_{\text{st}} F_\alpha^{(1)}$. If, for every $\alpha$ and $n$, $\xi_\alpha^{(1)}(n)$ and $\xi_\alpha^{(2)}(n)$ are generated from the same $U_\alpha(n)$, then $T_\beta^{(2)}(m) \leq T_\beta^{(1)}(m)$ for every $U = \{U_\alpha(n), \alpha \in \mathbf{A}, n > 0\}$ and every $\beta$ and $m$.*

This result shows that if the clock distributions for the systems under comparison are stochastically ordered, then the standard synchronization (assigning a random number stream to each event type) makes $T^{(2)} \leq T^{(1)}$ on *every* simulation run. It also yields a zero-variance estimate of $\mathbf{1}\{T_\alpha^{(2)}(m) \leq T_\alpha^{(1)}(m)\}$. While these properties do not by themselves guarantee variance reduction for all comparisons of the two systems, they are intuitively appealing and they suggest that each pair of runs across systems is indicative of their relative performance.

Theorem 4.13 admits extensions to more general performance measures. Consider the sojourn time in a queue (the sojourn time in tandem queues works similarly). Use the notation of §4.3. If $\xi_\alpha^{(1)} \leq_{\text{st}} \xi_\alpha^{(2)}$ and $\xi_\beta^{(2)} \leq_{\text{st}} \xi_\beta^{(1)}$, we may simulate the two systems so that on every run $W_n^{(2)} \leq W_n^{(1)}$ for every $n$. It follows that we can order the average sojourn times on each run. Using the argument preceding Corollary 4.11, we may simulate the two systems ensuring that certain queue lengths are always ordered.

As discussed in §2.3, another justification for synchronization is available for functions that are arrangement increasing, but this condition is hard to satisfy. Here, we give one very simple example. Consider a single-server queue with standard independent input. For fixed service times $\{\xi_\beta(i), i = 1, 2, \cdots\}$, it can be verified that every departure epoch $T_\beta(n)$ and arrival epoch $T_\alpha(n)$ is an arrangement increasing function of the interarrival times $\{\xi_\alpha(i), 1 \leq i \leq n\}$. If inversion is used, then they are also arrangement increasing functions of $\{U_\alpha(i), 1 \leq i \leq n\}$. Hence, there is no advantage to permuting the random numbers used to generate interarrival times in comparing two queues. This holds, more generally, for any event $\alpha$ which is in the event list of every state.

## 5. Structural Comparisons

The previous section considered comparisons of essentially the same system driven by different clock processes. In this section, we show that the properties of noninterruption and permutability also guarantee benefits from CRN in comparing systems with different structure. Different "structure" means different schemes in the GSMP setting, and this translates to, e.g., different buffer sizes, different job populations, and different numbers of nodes in queueing systems.

Let $\mathcal{G}^{(1)}$ and $\mathcal{G}^{(2)}$ be two noninterruptive, permutable schemes driven by inputs $\xi^{(1)}$ and $\xi^{(2)}$. Let $\Phi^{(i)}$, $i = 1, 2$, be increasing, right-continuous functions generating $\xi^{(i)}$ from i.i.d. uniform sequences $U^{(i)}$. Let $T^{(i)}$, $i = 1, 2$, be the event epoch sequence associated with $\mathcal{G}^{(i)}$ and $\xi^{(i)}$, and let $f$ and $g$ be increasing, right-continuous functions of $T^{(1)}$ and $T^{(2)}$ for which $f(T^{(1)})$ and $g(T^{(2)})$ have finite second moments. The following is an immediate extension of Theorem 4.1:

THEOREM 5.1. *With the notation and conditions above, CRN minimizes* $\mathbf{Var}\,[f(T^{(1)}) - g(T^{(2)})]$ *among all joint distributions in* $\mathcal{M}_0(U^{(1)}, U^{(2)})$.

Results along the lines of 4.2–4.5 follow immediately.

Comparison of two schemes $\mathcal{G}^{(1)}$ and $\mathcal{G}^{(2)}$ is most meaningful when the respective event sets $\mathbf{A}^{(i)}$, $i = 1, 2$, are the same set $\mathbf{A}$. (There is no loss of generality in assuming this, in any case, because we can always set $\mathbf{A} = \mathbf{A}^{(1)} \cup \mathbf{A}^{(2)}$ and then take $\xi_\alpha^{(i)}(n) \equiv \infty$ if $\alpha \notin \mathbf{A}^{(i)}$.) When this holds, we obtain analogs of Corollary 4.10. Consider queues in tandem, as in Example 4.7. Recall that we require that the first queue have infinite capacity, and let $b^{(i)} = (b_2^{(i)}, \ldots, b_k^{(i)})$, $i = 1, 2$, be the vectors of buffer capacities at the other queues for the two systems. Regardless of the values of the $b_j^{(i)}$'s, in comparing sojourn times in the two systems we get (maximal) variance reduction using common random numbers, provided only that service and interarrival times are appropriately generated monotonically. It is not even necessary that the number of queues in the two systems be the same.

In general, we may also assume that $\mathcal{G}^{(1)}$ and $\mathcal{G}^{(2)}$ have the same state space by, if necessary, taking both to have state space $\mathbf{S} = \mathbf{S}^{(1)} \cup \mathbf{S}^{(2)}$ (though this may destroy irreducibility). We then obtain results for comparisons of queue-length-like quantities through the argument of Corollary 4.11.

There is a special class of structural comparisons for which we can also make a statement about synchronization. These are comparisons of *subschemes*: $\hat{\mathcal{G}}$ is a *subscheme* of $\mathcal{G}$ (denoted $\hat{\mathcal{G}} \subseteq \mathcal{G}$) if $\hat{\mathbf{S}} \subseteq \mathbf{S}$, $\hat{\mathbf{A}} \subseteq \mathbf{A}$, $\hat{\mathcal{E}}(s) \subseteq \mathcal{E}(s)$ for all $s \in \hat{\mathbf{S}}$, and $\hat{p}(s'; s, \alpha) = p(s'; s, \alpha)$ for all $s, s' \in \hat{\mathbf{S}}$ and all $\alpha \in \hat{\mathcal{E}}(s)$. As explained above, we may take $\hat{\mathbf{A}} = \mathbf{A}$ and $\hat{\mathbf{S}} = \mathbf{S}$. For deterministic schemes, $\hat{p}(s'; s, \alpha) = p(s'; s, \alpha)$ reduces to $\hat{\phi}(s, \alpha) = \phi(s, \alpha)$.

In the example of tandem queues discussed above, a system with buffer vector $\hat{b}$

$= (\hat{b}_2, \ldots, \hat{b}_k)$ is a subscheme of one with vector $b = (b_2, \ldots, b_k)$ if and only if $\hat{b} \le b$. In comparing two closed queueing networks as in Example 4.9 but with different job populations, the one with the smaller population is a subscheme of the other. Eliminating a queue from a tandem network or reducing the number of servers at a multiple-server queue also produce subschemes; see Glasserman and Yao (1992a).

It follows from the definition of subscheme that if $\hat{\mathcal{G}} \subseteq \mathcal{G}$ then any sequence of events $\sigma$ feasible for $\hat{\mathcal{G}}$ is also feasible for $\mathcal{G}$. From this observation we can prove

LEMMA 5.2. *Suppose that $\hat{\mathcal{G}}$ and $\mathcal{G}$ are deterministic, noninterruptive, permutable schemes and that $\hat{\mathcal{G}} \subseteq \mathcal{G}$. Let $\hat{\xi}$ and $\xi$ be the corresponding clock processes and $\hat{T}$ and $T$ the corresponding event epoch sequences when both start from the same initial state $s_0$. $\hat{T}$ and $T$ have representations* (7) *with indices $\{\hat{x}^j(\alpha, n)\}$ and $\{x^i(\alpha, n)\}$. Every such vector $\hat{x}^j(\alpha, n)$ dominates some $x^i(\alpha, n)$.*

From this it follows that if $\hat{\xi} = \xi$ with probability one, then $T \le \hat{T}$ with probability one. In fact, we have

THEOREM 5.3. *Let $\hat{\mathcal{G}}$ and $\mathcal{G}$ be as in Lemma 5.2. Consider the case of standard independent input with inversion. Let $\{\hat{F}_\alpha, \alpha \in \mathbf{A}\}$ and $\{F_\alpha, \alpha \in \mathbf{A}\}$ be the clock distributions for the two systems, $\hat{\xi}_\alpha(n) = \hat{F}_\alpha^{-1}(U_\alpha(n))$ and $\xi_\alpha(n) = F_\alpha^{-1}(U_\alpha(n))$. If, for every $\alpha \in \mathbf{A}$, $F_\alpha \le_{\mathrm{st}} \hat{F}_\alpha$, then for every $U$, $T_\alpha(n) \le \hat{T}_\alpha(n)$, for every $\alpha$ and $n$.*

Theorem 5.3 shows that in comparing one system $\mathcal{G}$ with a subsystem $\hat{\mathcal{G}}$, if both are noninterruptive and permutable and if their clock distributions are stochastically ordered, then the standard synchronization makes $T \le \hat{T}$ on every simulation run. In particular, it yields a zero-variance estimate of every $\mathbf{1}\{T_\alpha(n) \le \hat{T}_\alpha(n)\}$. For example, in comparing two tandem networks with buffer vectors $\hat{b} \le b$, we can simulate the two systems so that every service completion in the smaller system occurs after the corresponding event in the larger system. Using the arguments of §4.3 we can also guarantee that the sojourn times for the two systems and the queue lengths at a particular node are ordered on every simulation run. Similar conclusions apply whenever the subscheme relation holds.

## 6. Sensitivity Analysis

We now return to the setting of §4—a single scheme driven by different clock processes—but with a different emphasis. We consider comparisons based on small changes in a continuous parameter. In this context, we are primarily interested in the variance of a difference estimate as the magnitude of the parameter change goes to zero. For "reasonable" comparisons, the use of CRN guarantees that this variance goes to zero. But we show that for noninterruptive, permutable schemes the convergence to zero can be an order of magnitude faster.

We begin with a simple, general result. Let the parameter set be $\Theta$, an interval of the real line. Let $L(\theta)$ be a statistic computed from a simulation at parameter value $\theta$. Fix a nominal $\theta_0 \in \Theta$; performance at other $\theta$ values is to be compared with performance at $\theta_0$. Recall that a function $\psi$ on $\Theta$ is *Lipschitz continuous* if there is a $K > 0$ (the modulus) such that $|\psi(\theta_2) - \psi(\theta_1)| \le K|\theta_2 - \theta_1|$ for all $\theta_1, \theta_2 \in \Theta$.

LEMMA 6.1. *For any $\theta_0 \in \Theta$, $\mathbf{Var}\,[L(\theta_0 + h) - L(\theta_0)]$ is*

(i) $O(1)$, *if $\{L(\theta), \theta \in \Theta\}$ are independent and* $\sup_\theta \mathbf{E}[L^2(\theta)] < \infty$;

(ii) $o(1)$, *if $L$ is continuous at $\theta_0$ with probability one and* $\sup_\theta \mathbf{E}[|L(\theta)|^{2+\epsilon}] < \infty$ *for some $\epsilon > 0$;*

(iii) $O(h^2)$, *if $L$ is Lipschitz continuous throughout $\Theta$ with probability one and its* (*random*) *modulus $K_L$ satisfies* $\mathbf{E}[K_L^2] < \infty$.

In the setting we have in mind, $L(\theta)$ is computed from a simulation of a process $X(\theta) = \{X_t(\theta), t \ge 0\}$, $X(\theta)$ taking values in a complete, separable metric space. If the family of processes $\{X(\theta),\ \theta \in \Theta\}$ is *weakly continuous* in $\theta$ (meaning that $X(\theta_n) \Rightarrow X(\theta)$ whenever $\theta_n \rightarrow \theta$), and if $L$ is a continuous functional of $X$, then $\mathbf{E}[L(\cdot)]$ is continuous in $\theta$, under the additional technical requirement that $\{L(\theta),\ \theta \in \Theta\}$ be uniformly integrable. Thus, in comparing $\mathbf{E}[L(\theta_0 + h)]$ and $\mathbf{E}[L(\theta_0)]$, we should expect to be able to find a low-variance estimator when $h$ is small.

Part (i) of the lemma points out that sampling $L(\theta_0 + h)$ and $L(\theta_0)$ (or $X(\theta_0 + h)$ and $X(\theta_0)$) independently completely fails to take advantage of the potential continuity of $\mathbf{E}[L(\cdot)]$. Independent sampling will not force $L(\theta_0 + h)$ and $L(\theta_0)$ to be close. Part (ii) describes the typical performance of common random numbers. For most models, there is, with probability one, *some* neighborhood of each $\theta$ throughout which $L$ is continuous if the $X(\theta)$'s are sampled using common random numbers. The size of such a neighborhood may vary over different simulation runs. This "local" continuity is enough to ensure that the variance of $L(\theta_0 + h) - L(\theta_0)$ indeed becomes small as $h$ goes to zero, under the additional assumption in the lemma.

Part (iii) strengthens continuity to Lipschitz continuity, but a more important strengthening is the replacement of continuity at $\theta_0$ with continuity *throughout* $\Theta$. In other words, in (iii) the neighborhood of $\theta_0$ throughout which continuity holds is not allowed to depend on the simulation run. (The interval $\Theta$ could be an arbitrarily small neighborhood of $\theta_0$, but it must be fixed.) From this we get the faster convergence to zero indicated in the lemma.

We now show that part (iii) of Lemma 6.1 often holds in comparisons of noninterruptive, permutable schemes. Let $\Phi$ sample $\xi$ from a law that depends on $\theta$; we write $\xi(\theta) = \Phi(U, \theta)$. From Theorem 3.2 we conclude that if every $\xi_\beta(j, \theta)$ is, with probability one, continuous in $\theta$, then so is every $T_\alpha(n, \theta)$. Thus, continuity of the clock times gives us most of what we need to apply Lemma 6.1 (iii). In practice, the clock times are made continuous in $\theta$ by using common random numbers across different parameter values. We now have

THEOREM 6.2. *Suppose that $\mathcal{G}$ is noninterruptive and permutable, and that every $\xi_\alpha(n, \theta)$ is a Lipschitz continuous function of $\theta$ with modulus $K_{\alpha,n}$ satisfying $\mathbf{E}[K^2_{\alpha,n}] < \infty$. Then* $\mathbf{Var}\,[T_\alpha(n, \theta_0 + h) - T_\alpha(n, \theta_0)] = O(h^2)$, *for all $\theta_0 \in \Theta$ and all $\alpha$ and $n$ for which the indices on the right side of* (7) *are finite.*

This result is best viewed as a building block from which results for more interesting comparisons can be derived. For example, consider sojourn times as in Corollary 4.10: $W_n = T_\beta(n) - T_\alpha(n)$ and $\bar{W}_N = W_1 + \cdots + W_N$. From Theorem 6.2, each $T_\beta(n, \theta_0 + h) - T_\beta(n, \theta)$ and $T_\alpha(n, \theta_0 + h) - T_\alpha(n, \theta)$, $n = 1, \ldots, N$ has variance $O(h^2)$. By the Cauchy-Schwarz inequality, covariances among these terms are $O(h^2)$. Thus, sums of such terms have variance $O(h^2)$; in particular,

$$\mathbf{Var}\left[\frac{1}{N}[\bar{W}_N(\theta_0 + h) - \bar{W}_N(\theta_0)]\right] = O(h^2).$$

We can also derive a result for queue-length-like quantities, based on a stronger condition. If $f\colon \mathbf{S} \mapsto \mathbf{R}$ and $t > 0$ is fixed, define

$$L_f(\theta) = \frac{1}{t}\int_0^t f(X_s(\theta))ds. \tag{16}$$

Glasserman (1988b) shows that for noninterruptive, *strongly* permutable schemes, $L_f$ is continuous in $\theta$ with probability one, if every clock time is; however, here we will not apply that result directly. We show that $L_f$ is in fact Lipschitz under some assumptions on $\xi$. We need a generalization of the standard independent input: $\{\xi_\alpha,\ \alpha \in \mathbf{A}\}$ are independent sequences, and, for each $\alpha \in \mathbf{A}$, $\{\xi_\alpha(n, \cdot),\ n \geq 1\}$ are i.i.d. *functions of* $\theta$.

We suppose that every $\xi_\alpha(n, \cdot)$ is, with probability one, Lipschitz continuous with random modulus $K_{\alpha,n}$, and that $\{K_{\alpha,n}, \alpha \in \mathbf{A}, n = 1, 2, \cdots\}$ are independent. (As a practical matter, this is a condition on the dependence of the $\xi(\theta)$'s when sampled using common random numbers.) For each $\alpha$ and $n$ let $\bar{x}(\alpha, n)$ be the maximum of the indices appearing on the right side of (7). Notice that $\bar{x}(\alpha, n) \geq n - 1$.

THEOREM 6.3. *Consider a noninterruptive, strongly permutable scheme. Let $\xi$ be as above and suppose that, for all $\alpha$ and $n$, $\mathbf{E}[K_{\alpha,n}^4] < \infty$ and $P(\inf_\theta \xi_\alpha(n, \theta) = 0) < 1$. Suppose that, for all $\alpha$, $\bar{x}(\alpha, n) = O(n)$. Let $f$ be bounded. Then* $\mathbf{Var}$ $[L_f(\theta_0 + h) - L_f(\theta_0)] = O(h^2)$ *for all* $\theta_0 \in \Theta$.

The condition that $\bar{x}(\alpha, n) = O(n)$ is by no means restrictive. In fact, $\bar{x}(\alpha, n) \in \{n - 1, n\}$ is typical for queueing systems; this is the case, e.g., in Examples 4.6 and 4.7. The condition $\bar{x}(\alpha, n) = O(n)$ excludes the possibility that the $n$th occurrence of $\alpha$ must be preceded by an order of magnitude more occurrences of some other event $\beta$.

The next result points out that a conclusion similar to Theorem 6.3 holds for steady-state comparisons provided variances converge uniformly. We need to assume that there is a deterministic function $l(\cdot)$ such that for all $\theta_0$, $\theta_0 + h \in \Theta$

$$t^{-1/2}\left[\int_0^t \{f(X_s(\theta_0 + h)) - f(X_s(\theta_0))\}ds - t[l(\theta_0 + h) - l(\theta_0)]\right] \Rightarrow \sigma_h \mathcal{N}(0, 1) \quad (17)$$

for some finite $\sigma_h$. We also need a uniform integrability condition: as $t \to \infty$,

$$t^{-1}\,\mathbf{Var}\left[\int_0^t \{f(X_s(\theta_0 + h)) - f(X_s(\theta_0))\}ds - t[l(\theta_0 + h) - l(\theta_0)]\right] \to \sigma_h^2. \quad (18)$$

THEOREM 6.4. *In addition to the conditions of Theorem* 6.3, *assume* (17) *and suppose that* (18) *holds uniformly in $h$ in a neighborhood of $h = 0$. Then $\sigma_h^2$ is $o(h)$.*

Verifying uniform convergence in (18) is difficult and, in practice, generally not possible. This theorem primarily serves to spell out what conditions lead to the conclusion, in principle. However, uniform convergence would seem to be a plausible additional assumption whenever the variances in (18) are continuous in $h$ in some neighborhood of 0 not depending on $t$. Continuity can be expected to hold whenever the input distributions are reasonably smooth in $\theta$.

## 7. Concluding Remarks

We have presented a variety of settings in which the use of common random numbers is effective and even optimal. To a large extent, our results support standard simulation practice; indeed, a principal contribution of this paper is the identification of a class of systems and criteria for which folklore is provably correct. The key guideline that may be extracted from these results is the importance of examining what happens when events change order. Definition 3.1 formalizes this idea and thereby helps formalize intuition about when CRN is effective.

Variance reduction is guaranteed (in comparing throughputs and in some cases sojourn times and queue lengths) whenever changing the order of some events does not radically change the evolution of the system. This is the case for most standard queueing systems with a single class of jobs and a first-come-first-served discipline, but not for most multiclass networks or queues with, e.g., pre-emptive disciplines.

Two further points deserve comment:

(i) If one is willing to restrict the allowable comparisons, then something weaker than permutability would suffice. Permutability need only hold for certain changes in the order of events. This is best illustrated through an example. Consider a queue fed by two classes of jobs, as in Example 4.8, but with an infinite buffer. This system violates our conditions. However, suppose we only consider changes in the service time distributions—we do not compare systems with different interarrival time distributions. It is not hard to see that the departure epochs of all jobs are, in fact, increasing and continuous in service times, because the order of arrivals cannot change. Hence, our results can be applied to this restricted class of comparisons. More generally, one needs to check only those changes in the order of events that can happen through the changes to be compared. This is made precise via a definition of relevance in Glasserman and Yao (1992a).

(ii) The structural properties considered here have implications for variance reduction beyond CRN. Indeed, all variance reduction techniques that depend on inducing positive (or negative) correlation ultimately rely on (or at least benefit from) some type of monotonicity. The techniques of *antithetic variates*, *control variates* and many instances of *indirect estimation* belong to this class. Consider the second of these. Suppose, for some system, $m_\alpha = \lim_{n\to\infty} T_\alpha(n)/n$ is known and $\lim_{n\to\infty} T_\beta(n)/n$ is to be estimated. Strong correlation among the event epochs would make $\{T_\alpha(n),\ n > 0\}$ a good control for estimation of $m_\beta$.[1]

[1] This work was carried out while Paul Glasserman was with the Operations Research Department of AT&T Bell Laboratories. David Yao was supported, in part, by NSF grant ECS-89-96230. The authors thank the Area Editor, the Associate Editor and a referee for their constructive comments.

## Appendix: Proofs

PROOF OF PROPOSITION 2.2. All covariances in (4) exist because $f(X)$ and $g(Y)$ are square-integrable, so we may apply Proposition 2.1. The distributions in $\mathcal{M}_0(U, V)$ are characterized by (3), so we need to maximize over the joint distribution of $(U_i, V_i)$, $i = 1, 2, \ldots$. Suppose all such distributions except that of some $(U_j, V_j)$ have been fixed. For any fixed value of $\{(u_i, v_i),\ i \neq j\}$, $[f \circ \Phi_X(u)] \cdot [g \circ \Phi_Y(v)]$ is supermodular and right-continuous as a function of $(u_j, v_j)$, because each factor is increasing and right-continuous. Hence, its expectation is maximized by setting $P(U_j \le u_j, V_j \le v_j) = H(u_j, v_j) = u_j \wedge v_j$. Since this holds regardless of the distributions of $(U_i, V_i)$, $i \neq j$, this joint distribution must be optimal for all $j$. □

PROOF OF PROPOSITION 2.4. Let $\pi_{ij}$ permute $u_i$ and $u_j$, $i < j$. It is enough to show that $\mathbf{E}[f(U)g(U)] \ge \mathbf{E}[f(U^{\pi_{ij}})g(U)]$, because every permutation is a product of transpositions. Let $\Delta_{ij}f(u) = f(u) - f(u^{\pi_{ij}})$, define $\Delta_{ij}g$ the same way, and note that $\Delta_{ij}f(u^{\pi_{ij}}) = -\Delta_{ij}f(u)$. Using this definition, we get $\mathbf{E}[f(U)g(U)] - \mathbf{E}[f(U^{\pi_{ij}})g(U)] = \mathbf{E}[\Delta_{ij}f(U)g(U)]$, which we may rewrite as $\mathbf{E}[\Delta_{ij}f(U)g(U)\mathbf{1}\{U_i > U_j\} + \Delta_{ij}f(U)g(U)\mathbf{1}\{U_j > U_i\}]$. However, because the $U_k$'s are i.i.d., this expectation is unchanged if we reverse $U_i$ and $U_j$ in the second term. Thus, this expectation equals $\mathbf{E}[\{\Delta_{ij}f(U)g(U) + \Delta_{ji}f(U^{\pi_{ij}})g(U^{\pi_{ij}})\}\mathbf{1}\{U_i > U_j\}]$. This can be further rewritten as $\mathbf{E}[\{\Delta_{ij}f(U)\Delta_{ij}g(U)\}\mathbf{1}\{U_i > U_j\}]$, which is nonnegative because $\Delta_{ij}f(u)$ and $\Delta_{ij}g(u)$ are nonnegative on $\{u : u_i > u_j\}$ if $f$ and $g$ are arrangement increasing. Thus, $\mathbf{E}[f(U)g(U)] - \mathbf{E}[f(U^{\pi_{ij}})g(U)] \ge 0$, which is what we needed to show. □

PROOF OF THEOREM 4.1. The composition of increasing, right-continuous functions is increasing and right-continuous, so $f \circ T^{(1)} \circ \Phi^{(1)}$ and $g \circ T^{(2)} \circ \Phi^{(2)}$ are increasing, right-continuous functions of $U$ and $V$. (Monotonicity and continuity of $T^{(i)}$, $i = 1, 2$ are consequences of Theorem 3.2.) Proposition 2.2 now applies. □

PROOF OF THEOREM 4.5. Uniform integrability implies that

$$\sigma_\alpha^2 = \lim_{n\to\infty} n^{-1} \mathbf{Var}\,\{[T_\alpha^{(1)}(n) - nm_\alpha^{(1)}(n)] - [T_\alpha^{(2)}(n) - nm_\alpha^{(2)}(n)]\};$$

hence, a joint distribution which minimizes the variance of $[T_\alpha^{(1)}(n) - T_\alpha^{(2)}(n)]$ for all $n$ also minimizes the asymptotic variance $\sigma_\alpha^2$. In light of Theorem 4.1, such a distribution is obtained by setting $U = V$. □

PROOF OF THEOREM 4.12. If $\xi^{(k)} \to \xi$, then $T^{(k)} \to T$. From (7) we obtain, for all $k$,

$$T_\alpha^{(k)}(n) \le \xi_\alpha^{(k)}(n) + \sum_{\beta\in\mathbf{A}} \sum_{j\le\bar{x}} \xi_\beta^{(k)}(j),$$

where $\bar{x} = \max_{\beta,j}\{x_\beta^j(\alpha, n)\} < \infty$. Thus, for any $r > 0$, $T_\alpha^{(k)}(n)$ has finite $r$th moment if every $\xi_\beta^{(k)}(j)$ does. In particular, under the hypotheses of the theorem, $\sup_{k\ge0} \mathbf{E}[T_\alpha^{(k)}(n)^{2+\epsilon'}] < \infty$ for some $\epsilon \ge \epsilon' > 0$. Consequently,

$$\lim_{k\to\infty} \mathbf{Var}\,[T_\alpha^{(k)}(n) - T_\alpha(n)] = \mathbf{Var}\,[\lim_{k\to\infty} T_\alpha^{(k)}(n) - T_\alpha(n)] = 0. \qquad \square$$

PROOF OF THEOREM 4.13. If $F \leq_{\text{st}} G$, then $F^{-1}(u) \leq G^{-1}(u)$ for all $u \in [0, 1]$. Thus, $\xi_\alpha^{(2)}(n) \leq \xi_\alpha^{(1)}(n)$ with probability one, for every $\alpha$ and $n$. $T^{(2)}$ and $T^{(1)}$ inherit this ordering via (7). □

PROOF OF LEMMA 5.2. Let $\hat{x}^j(\alpha, n)$ be a vector in the representation (7) of $\hat{T}_\alpha(n)$. Then if $\sigma$ is a sequence of events, feasible for $\hat{\mathcal{G}}$, with $N(\sigma) \geq \hat{x}^j(\alpha, n)$, either $N_\alpha(\sigma) \geq n$, or $N_\alpha(\sigma) = n - 1$ and $\alpha \in \hat{\mathcal{E}}(\hat{\phi}(s_0, \sigma))$. Since $\hat{\mathcal{G}} \subseteq \mathcal{G}$, $\sigma$ is also feasible for $\mathcal{G}$, and $\hat{\mathcal{E}}(\hat{\phi}(s_0, \sigma)) \subseteq \mathcal{E}(\phi(s_0, \sigma))$. Thus, either $N_\alpha(\sigma) = n$ or $N_\alpha(\sigma) = n - 1$ and $\alpha \in \mathcal{E}(\phi(s_0, \sigma))$. But then $N(\sigma)$ must dominate some $x^i(\alpha, n)$. Since $\sigma$ is an arbitrary string dominating $\hat{x}^j(\alpha, n)$, we conclude that $\hat{x}^j(\alpha, n)$ must dominate some $x^i(\alpha, n)$. □

PROOF OF THEOREM 5.3. The ordering of the clock distributions implies that $\xi_\alpha(n) \leq \hat{\xi}_\alpha(n)$ with probability one, for every $\alpha$ and $n$. $T$ and $\hat{T}$ inherit this ordering because they are monotone increasing and because of the ordering of the indices established in Lemma 5.2. □

PROOF OF LEMMA 6.1. (i) In the independent case, $\mathbf{Var}\,[L(\theta_0 + h) - L(\theta_0)] = \mathbf{Var}\,[L(\theta_0 + h)] + \mathbf{Var}\,[L(\theta_0)]$, which is bounded as $h \to 0$ because $\mathbf{Var}\,[L(\theta)]$ is bounded on $\Theta$. (ii) $\mathbf{E}[|L(\theta)|^{2+\epsilon}]$ bounded on $\Theta$ implies

$$\lim_{h\to 0} \mathbf{Var}\,[L(\theta_0 + h) - L(\theta_0)] = \mathbf{Var}\,[\lim_{h\to 0} L(\theta_0 + h) - L(\theta_0)]$$

which is zero if $L$ is almost surely continuous at $\theta_0$. (iii)

$$\begin{aligned}\mathbf{Var}\,[L(\theta_0 + h) - L(\theta_0)] &\leq \mathbf{E}[\{L(\theta_0 + h) - L(\theta_0)\}^2] + \{\mathbf{E}[|L(\theta_0 + h) - L(\theta_0)|]\}^2 \\ &\leq (\mathbf{E}[K_L^2] + \mathbf{E}^2[K_L])h^2 = O(h^2). \qquad \square\end{aligned}$$

PROOF OF THEOREM 6.2. Let $\bar{x}$ be as in the proof of Theorem 4.12. Then

$$|T_\alpha(n, \theta + h) - T_\alpha(n, \theta)| \leq \sum_{\beta\in\mathbf{A}} \sum_{j\leq\bar{x}+1} |\xi_\beta(j, \theta + h) - \xi_\beta(j, \theta)|;$$

hence, $T_\alpha(n, \cdot)$ is Lipschitz with square-integrable modulus

$$\hat{K}_{\alpha,n} = \sum_{\beta\in\mathbf{A}} \sum_{j\leq\bar{x}+1} K_{\beta,j}. \tag{19}$$

The result now follows from Lemma 6.1. □

PROOF OF THEOREM 6.3. Let $\|f\|$ be the supremum of $|f|$. Then

$$\begin{aligned}\left|\int_0^t \{f(X_s(\theta + h)) - f(X_s(\theta))\}ds\right| &\leq 2\|f\| \int_0^t \mathbf{1}\{X_s(\theta + h) \neq X_s(\theta)\}ds \\ &\leq 2\|f\| \int_0^t \mathbf{1}\{D_s(\theta + h) \neq D_s(\theta)\}ds \qquad (20)\end{aligned}$$

because $X_s$ is determined by $D_s$ in a strongly permutable scheme. The total time in $[0, t]$ over which $D_s(\theta + h)$ and $D_s(\theta)$ differ is bounded by the total change in the epochs of events in $[0, t]$; i.e.,

$$\int_0^t \mathbf{1}\{D_s(\theta + h) \neq D_s(\theta)\}ds \leq \sum_{\alpha,n} \mathbf{1}\{T_\alpha(n, \theta_0 + h) \wedge T_\alpha(n, \theta_0) \leq t\}\,|T_\alpha(n, \theta_0 + h) - T_\alpha(n, \theta_0)|, \tag{21}$$

the sum running over all $\alpha \in \mathbf{A}$ and all $n = 1, 2, \ldots$. Let

$$A_\alpha(n) = \inf_\theta \xi_\alpha(1, \theta) + \cdots + \inf_\theta \xi_\alpha(n, \theta);$$

clearly, $A_\alpha(n) \leq T_\alpha(n, \theta)$ for all $\theta$. By hypothesis, the infimum of the clock times is not identically zero, so, for all $\alpha$ and $t$, the indicator $\mathbf{1}\{A_\alpha(n) \leq t\}$ is zero for all but finitely many $n$. (Use renewal theory; e.g., Theorem 5.2.1 of Prabhu 1965). It follows that all but finitely many terms in (21) are zero.

Let $\hat{K}_{\alpha,n}$ be the Lipschitz modulus for $T_\alpha(n, \cdot)$ derived in (19). Combining (20) and (21), we see that $L_f(\cdot)$ is Lipschitz with modulus

$$2\|f\|t^{-1} \sum_{\alpha,n} \mathbf{1}\{A_\alpha(n) \leq t\}\hat{K}_{\alpha,n}.$$

We need to show that the sum has finite second moment. Squaring the sum, taking the expectation and applying the Cauchy-Schwarz inequality twice, we find that it is enough to verify that

$$\sum_{\alpha_1,n_1} \sum_{\alpha_2,n_2} P(A_{\alpha_1}(n_1) \leq t)P(A_{\alpha_2}(n_2) \leq t)\mathbf{E}[\hat{K}^4_{\alpha_1,n_1}]\mathbf{E}[\hat{K}^4_{\alpha_2,n_2}]$$

is finite. For each $\alpha$, since $\bar{x}(\alpha, n) = O(n)$, it follows from (19) and the fact that $\{K_{\alpha,n}, \alpha \in \mathbf{A}, n = 1, 2\}$ are independent and have fourth moments that $\mathrm{E}[\hat{K}^4_{\alpha,n}] = O(n^4)$. On the other hand, since $A_\alpha(n)$ is a sum of i.i.d. random variables which are not strictly zero, each $P(A_\alpha(n) \le t)$ is $O(\rho^n)$ for some $\rho < 1$. Hence, the sum is finite. □

PROOF OF THEOREM 6.4. Let $\sigma_h^2(t)$ be the variance on the left side of (18). Then

$$\begin{aligned}
\lim_{h\to 0} h^{-1}\sigma_h^2 &= \lim_{h\to 0}\lim_{t\to\infty} h^{-1}t^{-1}\sigma_h^2(t) && \text{by (18)},\\
&= \lim_{t\to\infty}\lim_{h\to 0} h^{-1}t^{-1}\sigma_h^2(t) && \text{by uniform convergence over } h,\\
&= \lim_{t\to\infty} 0 && \text{by Theorem 6.3},\\
&= 0. \quad \square
\end{aligned}$$

## References

BACCELLI, F., W. A. MASSEY AND D. TOWSLEY, "Acyclic Fork-Join Queueing Networks," *J. Assoc. Comput. Mach.*, 36 (1989), 615-642.

BARLOW, R. E. AND F. PROSCHAN, *Statistical Theory of Reliability and Life Testing*, Holt, Rinehart and Winston, New York, 1975.

BILLINGSLEY, P., *Convergence of Probability Measures*, John Wiley and Sons, New York, 1968.

BRATLEY, P., B. L. FOX AND L. E. SCHRAGE, *A Guide to Simulation*, (Second Ed.), Springer-Verlag, New York, 1987.

CAMBANIS, S., G. SIMONS AND W. STOUT, "Inequalities for $Ek(X, Y)$ when the Marginals are Fixed," *Z. Wahrsch. Gebiete*, 36 (1976), 285-294.

DALEY, D., "Stochastically Monotone Markov Chains," *Z. Wahrsch. Gebiete*, 10 (1968), 305-317.

DEVROYE, L., "Coupled Samples in Simulation," *Oper. Res.*, 38 (1990), 115-126.

GLASSERMAN, P., "Sensitivity of Sample Values not Generated by Inversion," *J. Option Theory Appl.*, 52 (1988a), 487-493.

———, "Equivalence Methods in the Perturbation Analysis of Queueing Networks," Ph.D. Thesis, Division of Applied Sciences, Harvard University, 1988b.

——— AND D. D. YAO, "Monotonicity in Generalized Semi-Markov Processes," *Math. Oper. Res.*, 17 (1992a) 1-21.

——— AND ———, "Generalized Semi-Markov Processes: Antimatroid Structure and Second-Order Properties," *Math. Oper. Res.*, 17 (1992b) 444-469.

GLYNN, P. W., "Regenerative Structure of Markov Chains Simulated via Common Random Numbers," *Oper. Res. Lett.*, 4 (1985), 49-53.

——— AND D. L. IGLEHART, "Simulation Methods for Queues: An Overview," *Queueing Systems*, 3 (1988), 221-255.

GREENBERG, A. G., B. D. LUBACHEVSKY AND I. MITRANI, "Unboundedly Parallel Simulations via Recurrence Relations," ACM Sigmetrics '90, Boulder, CO, also *ACM Trans. Computer Systems*, (1990) (to appear).

HARDY, G. H., J. E. LITTLEWOOD AND G. POLYA, *Inequalities*, (2nd. ed.), Cambridge University Press, Cambridge, 1952.

HEIDELBERGER, P. AND D. L. IGLEHART, "Comparing Stochastic Systems Using Regenerative Simulation and Common Random Numbers," *Adv. in Appl. Probab.*, 11 (1979), 804-819.

HOEFFDING, W., "Masstabinvariante Korrelationstheorie," *Schrift. Math. Inst. und des Instituts Angewandte Math. der Univ. Berline*, 5 (1940), 179-233.

LEHMANN, E. L., "Some Concepts of Dependence," *Ann. Math. Statist.*, 37 (1966), 1137-1153.

LINDQVIST, B. H., "Association of Probability Measures on Partially Ordered State Spaces," *J. Multivariate Anal.*, 26 (1988), 111-132.

LORENTZ, G. G., "An Inequality for Rearrangements," *Amer. Math. Monthly*, 60 (1953), 176-179.

MARSHALL, A. W. AND I. OLKIN, *Inequalities: Theory of Majorization and Its Applications*, Academic Press, New York, 1979.

PRABHU, N. U., *Stochastic Processes*, Macmillan, New York, 1965.

RACHEV, S. T., "The Monge-Kantorovich Mass Transference Problem and Its Stochastic Applications," *Theor. Probab. Appl.*, 29 (1984), 647-676.

ROYDEN, H. L., *Real Analysis*, (2nd. ed.), Macmillan, New York, 1968.

RUBINSTEIN, R. Y. AND G. SAMORODNITSKY, "Variance Reduction by the Use of Common and Antithetic Random Variables," *J. Statist. Comput. Simul.*, 22 (1985), 161-180.

———, ——— AND M. SHAKED, "Antithetic Variates, Multivariate Dependence and Simulation of Stochastic Systems," *Management Sci.*, 31 (1985), 66-77.

SCHMEISER, B. W. AND V. KACHITVICHYANUKUL, "Correlation Induction Without the Inverse Transform," *Proc. Winter Simulation Conf.*, J. R. Wilson, J. O. Henriksen, and S. D. Roberts (Eds.), 1986, 266–274.

SHAKED, M. AND J. G. SHANTHIKUMAR, "The Total Hazard Rate Construction, Antithetic Variates and Simulation of Stochastic Systems," *Comm. Statist. Stochastic Models*, 2 (1986), 237–249.

SHANTHIKUMAR, J. G. AND D. D. YAO, "The Effect of Increasing Service Rates in a Closed Queueing Network," *J. Appl. Probab.*, 23 (1986), 474–483.

TSOUCAS, P. AND J. WALRAND, "Monotonicity of Throughput in Non-Markovian Networks," *J. Appl. Probab.*, 26 (1989), 134–141.

WHITT, W., "Bivariate Distributions with Given Marginals," *Ann. Math. Statist.*, 4 (1976), 1280–1289.

WILSON, J. R., "Antithetic Sampling with Multivariate Inputs," *Amer. J. Math. Management Sci.*, 3 (1983), 121–144.

WRIGHT, R. D. AND T. E. RAMSAY, JR., "On the Effectiveness of Common Random Numbers," *Management Sci.*, 25 (1979), 649–656.