Proceedings of the 1992 Winter Simulation Conference
ed. J. J. Swain, D. Goldsman, R. C. Crain, and J. R. Wilson

SENSITIVITY ANALYSIS OF DISCRETE EVENT SYSTEMS WITH AUTOCORRELATED INPUTS

Benjamin Melamed

NEC USA, Inc.
4 Independence Way
Princeton, NJ 08540, U.S.A.

Reuven Y. Rubinstein

Faculty of Industrial Engineering and Management
Technion – Israel Institute of Technology
Haifa 32000, ISRAEL

ABSTRACT

We consider the model $\ell(\boldsymbol{v}) = \mathbb{E}_{\boldsymbol{v}_1}\{L(\underline{\boldsymbol{Y}}_t, \boldsymbol{v}_2)\}$, $\boldsymbol{v} = (\boldsymbol{v}_1, \boldsymbol{v}_2)$, in a Monte Carlo simulation context with particular emphasis on queueing models with TES input sequences. Here L is the sample performance driven by an input sequence $\underline{\boldsymbol{Y}}_t = (\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_t)$, each $\boldsymbol{Y}_j$, $j = 1, \ldots, t$, being a sample from a probability density function (pdf) $f(\boldsymbol{y}, \boldsymbol{v}_1)$, and $\boldsymbol{v} = (\boldsymbol{v}_1, \boldsymbol{v}_2)$ is a vector of parameters (the subscript $\boldsymbol{v}_1$ in $\mathbb{E}_{\boldsymbol{v}_1} L$ indicates that the expectation is taken with respect to the pdf $f(\boldsymbol{y}, \boldsymbol{v}_1)$); TES is a versatile class of Markovian processes which can be used to model empirical time series by simultaneously approximating their empirical marginal (histogram) and leading autocorrelations.

To estimate the performance $\ell(\boldsymbol{v})$ and the associated sensitivities $\nabla^k \ell(\boldsymbol{v}) = \{\nabla^k_{\boldsymbol{v}_1} \ell(\boldsymbol{v}), \nabla^k_{\boldsymbol{v}_2} \ell(\boldsymbol{v})\}$, $k = 0, 1$, we consider the so-called "push out" and "push in" (infinitesimal perturbation analysis) approaches. We show that the "push out" technique merely replaces the original sample function $L(\boldsymbol{v}_2)$ by an auxiliary one $\tilde{L}$ while "pushing out" the parameter vector $\boldsymbol{v}_2$ from $L(\boldsymbol{v}_2)$ to an auxiliary pdf $\tilde{f}(\boldsymbol{y}, \boldsymbol{v})$. We also show that the IPA method, introduced by Ho and his co-workers, corresponds to the "push in" technique; the latter can be viewed as a dual of the "push out" technique. We finally show that the "push in" transformation $\boldsymbol{x} = \boldsymbol{x}(\boldsymbol{y}, \boldsymbol{v}_1)$ typically leads to a non-smooth sample performance function, violating the interchangeability conditions of expectation and differentiation.

1 INTRODUCTION

This paper deals with sensitivity analysis of DES (discrete event systems) for the model

$$\ell(\boldsymbol{v}) = \mathbb{E}_{\boldsymbol{v}_1}\{L(\underline{\boldsymbol{Y}}_t, \boldsymbol{v}_2)\}. \tag{1.1}$$

Here L is the sample performance driven by an input sequence $\underline{\boldsymbol{Y}}_t = (\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_t)$, each $\boldsymbol{Y}_j$, $j = 1, \ldots, t$, being a sample from a probability density function (pdf) $f(\boldsymbol{y}, \boldsymbol{v}_1)$, and $\boldsymbol{v} = (\boldsymbol{v}_1, \boldsymbol{v}_2)$ is a vector of parameters (the subscript $\boldsymbol{v}_1$ in $\mathbb{E}_{\boldsymbol{v}_1} L$ indicates that the expectation is taken with respect to the pdf $f(\boldsymbol{y}, \boldsymbol{v}_1)$). We assume that f depends on $\boldsymbol{v}_1$ and not on $\boldsymbol{v}_2$, whereas L depends on $\boldsymbol{v}_2$ and not on $\boldsymbol{v}_1$. Note also that the standard model $\ell(\boldsymbol{v}) = \mathbb{E}_{\boldsymbol{v}}\{L(\underline{\boldsymbol{Y}}_t)\}$ can be considered as a particular case of the model (1.1) with L independent of $\boldsymbol{v}_2$ and $\boldsymbol{v} = \boldsymbol{v}_1$.

Our goal is to perform sensitivity analysis, namely, to estimate the expected steady-state performance $\ell(\boldsymbol{v})$, $\boldsymbol{v} = (\boldsymbol{v}_1, \boldsymbol{v}_2)$ and the associated sensitivities $\nabla^k \ell(\boldsymbol{v})$, $k \geq 1$. We focus on queueing models, with particular emphasis on autocorrelated TES arrival processes; TES is a versatile class of Markovian processes which can be used to model empirical time series by simultaneously approximating their empirical marginal (histogram) and leading autocorrelations (see below).

Assume that the output process $\{L_t : t > 0\}$ is driven by an autocorrelated input sequence $\{\boldsymbol{X}_t : t > 0\}$; that is, $L_t(\cdot) = L_t(\underline{\boldsymbol{X}}_t)$, where $\underline{\boldsymbol{X}}_t = (\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_t)$ and $L_t(\cdot)$ is a sequence of real-valued functions. As an example one may think of autocorrelated interarrival times and service time random variables in a $G/G/1$ queue with $\{L_t : t > 0\}$ being the waiting time process. We assume, for simplicity, that $\{X_t : t > 0\}$ is a 1-dimensional process and that it can be written recursively as

$$X_t(\boldsymbol{v}_2) = \boldsymbol{x}(X_{t-1}, Y_t, \boldsymbol{v}_2) = \boldsymbol{x}(\underline{Y}_t, \boldsymbol{v}_2,), \ t > 0, \tag{1.2}$$

where $\boldsymbol{x}$ is a real-valued function, and $\{Y_t : t > 0\}$ is an iid sequence as before. It follows that the sequence $\{X_t : t > 0\}$ itself is a function of an iid sequence $\{Y_t : t > 0\}$; i.e., $X_t(\cdot) = X_t(\underline{Y}_t, \boldsymbol{v}_2)$, where the $X_t(\cdot)$ are real-valued functions, $\underline{Y}_t = (Y_1, Y_2, \ldots, Y_t)$, where $Y_j \sim f(y, \boldsymbol{v}_1)$, $j = 1, 2, \ldots$, and $\boldsymbol{v}_2$ is the

parameter vector associated with the autocorrelation function of the sequence $X_1, X_2, \ldots, X_t$. The parameter vector v_2 will be called the *autocorrelation parameter vector* (see examples below). We also assume that as $t \to \infty$, both processes $\{X_t : t > 0\}$ and $\{L_t : t > 0\}$ become stationary and ergodic. As examples of $\{X_t : t > 0\}$ consider:

(1) **AR(1): First order autoregressive.**

$$X_t(v_2) = \begin{cases} X_1, & \text{if } t = 1, \\ v_2 X_{t-1} + Y_t, & \text{if } t > 1, \end{cases} \tag{1.3}$$

where $-1 < v_2 < 1$, and similarly for higher order autoregressive processes and ARMA(p,q) processes.

(2) **TES: Transform-Expand-Sample.** Consider random variables of the form

$$X_t(\boldsymbol{v}_2) = F^{-1}[\eta_t(\boldsymbol{v}_2)], \tag{1.4}$$

where F^{-1} is the inverse function of a cdf $F(y, \boldsymbol{v})$. Then $X_t(\boldsymbol{v}_2) \sim F$; for example, if $Y \sim \exp(v)$, then $X_t = (-1/v)\ln(1 - \eta_t)$, where η_t is uniform on $[0, 1)$. There are two basic classes of TES processes: TES$^+$ and TES$^-$ (see Melamed 1991, Jagerman and Melamed 1992ab). The class TES$^+$ includes the simple processes $\{\eta_t^+(\boldsymbol{v}_2)\}$, given recursively by

$$\eta_t^+(\boldsymbol{v}_2) = \begin{cases} U_0, & \text{if } t = 0, \\ \langle \eta_{t-1}^+ + L + (R - L)U_t \rangle, & \text{if } t > 0. \end{cases} \tag{1.5}$$

Here $\boldsymbol{v}_2 = (L, R)$ where $-1/2 \le L < R \le 1/2$, $\langle x \rangle = x - \lfloor x \rfloor$ is the fractional part of x, $\lfloor x \rfloor = \max\{n \text{ integer} : n \le x\}$ is the integral part of x, and $\{U_t\}$ is a sequence of iid random variables, uniform on [0,1). It can be shown that the sequences $\{\eta_t^+\}$ cover all positive lag-1 autocorrelations in the range $[0, 1)$ by varying the parameters L and R. To similarly cover all negative lag-1 autocorrelations in the range $[-1, 0)$ one may use the class TES$^-$ which includes processes $\{\eta_t^-(\boldsymbol{v}_2)\}$, defined via the (antithetic) formula

$$\eta_t^- = \begin{cases} \eta_t^+, & \text{if } t \text{ even} \\ 1 - \eta_t^+, & \text{if } t \text{ odd.} \end{cases} \tag{1.6}$$

It can be shown that the marginals of both sequences $\{\eta_t^+\}$ and $\{\eta_t^-\}$ are uniformly distributed on the interval $[0, 1)$, and therefore, can be readily transformed to general marginal distributions F via (1.4). Furthermore, TES processes possess a variety of autocorrelation functions, which make them particularly suitable for fitting empirical data. In particular, if $R - L \to 0$, then $\{X_t : t \ge 0\}$ approaches the (conditionally) deterministic process $\{F^{-1}(U_0), F^{-1}(U_0), \ldots,\}$; if $R - L = 1$, $\{X_t : t \ge 0\}$ reduces to an iid process with marginals F; and if $0 < R - L < 1$, then $\{X_t : t \ge 0\}$ exhibits a variety of autocorrelation structures, including monotone, oscillating and alternating. For more details on general TES processes see Melamed (1991), and Jagerman and Melamed (1992a,b).

The rest of the paper is organized as follows. Section 2 deals with sensitivity analysis of the model (1.1). Special emphasis is placed on the smoothness of the sample function $L(\boldsymbol{y}, \boldsymbol{v}_2)$ and on variance reduction. Here we present two techniques, based on transforming random variables; these are called the "push out" and "push in" techniques, respectively. The terms "push out" and "push in" derive from the fact (see below) that in the first case we "push out" the parameter vector $\boldsymbol{v}_2$ from the original sample performance $L(\boldsymbol{Y}, \boldsymbol{v}_2)$ into an auxiliary pdf by a suitable transformation, and then apply the standard *Score Function* (SF) method; in the second case we operate the other way around, namely, we first "push in" (by a suitable transformation) the parameter vector $\boldsymbol{v}_1$ into the sample performance $L(\boldsymbol{Y}, \boldsymbol{v}_2)$, and then differentiate the resulting (auxiliary) sample performance with respect to $\boldsymbol{v} = (\boldsymbol{v}_1, \boldsymbol{v}_2)$. We discuss conditions under which such transformations are useful in the sense that they either generate *smooth sample performances* or lead to *variance reduction*. We also show how the *infinitesimal perturbation analysis* (IPA) method, introduced by Ho and his co-workers, corresponds to the "push in" technique; the latter can be viewed as a dual of the "push out" technique and fits into the paradigm described below. Section 3 presents numerical results.

2 "PUSH OUT" AND "PUSH IN" ESTIMATORS

2.1 Smoothness and Variance Reduction

This subsection presents some background material on two techniques, called "push out" and "push in". We shall show below that the first technique typically *smooths out* the sample performance function

$L(\boldsymbol{y}, \boldsymbol{v}_2)$ with respect to $\boldsymbol{v}_2$ by making it *independent* of $\boldsymbol{v}_2$, while the second technique can lead to *variance reduction*. Both techniques are based on the standard change of variables method.

(a) **The "Push Out" Technique**. To demonstrate the idea of the "push out" technique, first described in Rubinstein (1992a), consider a simple DES and suppose that there exists a vector-valued function $\boldsymbol{x} = \boldsymbol{x}(\boldsymbol{y}, \boldsymbol{v}_2)$ and a real-valued function $\tilde{L}(\boldsymbol{x})$ independent of $\boldsymbol{v}_2$, such that $L(\boldsymbol{y}, \boldsymbol{v}_2)$ can be represented as

$$L(\boldsymbol{y}, \boldsymbol{v}_2) = \tilde{L}[\boldsymbol{x}(\boldsymbol{y}, \boldsymbol{v}_2)]. \tag{2.1}$$

Furthemore, suppose that for $\boldsymbol{Y} \sim f(\boldsymbol{y}, \boldsymbol{v}_1)$, the corresponding random vector $\boldsymbol{X} = \boldsymbol{x}(\boldsymbol{Y}, \boldsymbol{v}_2)$ has a known pdf $\tilde{f}(\boldsymbol{x}, \boldsymbol{v}_1, \boldsymbol{v}_2)$. We can then write $\ell(\boldsymbol{v})$ as

$$\begin{aligned} \ell(\boldsymbol{v}) &= \int L(\boldsymbol{y}, \boldsymbol{v}_2)\, f(\boldsymbol{y}, \boldsymbol{v}_1)\, d\boldsymbol{y} \\ &= \int \tilde{L}(\boldsymbol{x})\, \tilde{f}(\boldsymbol{x}, \boldsymbol{v})\, d\boldsymbol{x} \;=\; \mathbb{E}_{\tilde{f}}\{\tilde{L}(\boldsymbol{X})\}, \end{aligned} \tag{2.2}$$

where the expectation is now taken with respect to the pdf $\tilde{f}(\boldsymbol{x}, \boldsymbol{v}_1, \boldsymbol{v}_2)$. Thus, the parameter vector $\boldsymbol{v}_2$ is effectively "pushed out" from $L(\boldsymbol{y}, \boldsymbol{v}_2)$ to an auxiliary pdf $\tilde{f}(\boldsymbol{x}, \boldsymbol{v}_1, \boldsymbol{v}_2)$.

It is important to understand that the representation of $L(\boldsymbol{y}, \boldsymbol{v}_2)$ in (2.1) and the subsequent transformation (2.2) are not always available, and even when available, the corresponding random vector $\boldsymbol{X}$ may not have a density function (with respect to the Lebesgue measure). Also, even if $\tilde{f}(\boldsymbol{x}, \boldsymbol{v}_1, \boldsymbol{v}_2)$ exists, it may be difficult to calculate. (For more details see Rubinstein and Shapiro 1992).

Suppose now that for every $\boldsymbol{v}_2$ the function $\boldsymbol{x} = \boldsymbol{x}(\boldsymbol{y}, \boldsymbol{v}_2)$ is one-to-one and thus has an inverse $\boldsymbol{y} = \boldsymbol{y}(\boldsymbol{x}, \boldsymbol{v}_2)$, which is assumed to be continuously differentiable in $\boldsymbol{x}$. In this case we have

$$\tilde{f}(\boldsymbol{x}, \boldsymbol{v}_1, \boldsymbol{v}_2) = f[\boldsymbol{y}(\boldsymbol{x}, \boldsymbol{v}_2), \boldsymbol{v}_1] \left| \frac{\partial \boldsymbol{y}(\boldsymbol{x}, \boldsymbol{v}_2)}{\partial \boldsymbol{x}} \right|, \tag{2.3}$$

where $|\partial \boldsymbol{y}/\partial \boldsymbol{x}|$ denotes the absolute value of the determinant of the Jacobian matrix of $\boldsymbol{y}(\boldsymbol{x}, \boldsymbol{v}_2)$ with respect to $\boldsymbol{x}$. For example, suppose that $\boldsymbol{v}_2$ has the same dimensionality as $\boldsymbol{y}$, and that the function $L(\boldsymbol{y}, \boldsymbol{v}_2)$ can be represented as $L(\boldsymbol{y}, \boldsymbol{v}_2) = \tilde{L}(\boldsymbol{y} + \boldsymbol{v}_2)$. We can then define $\boldsymbol{x} = \boldsymbol{y} + \boldsymbol{v}_2$, and obtain

$$\tilde{f}(\boldsymbol{x}, \boldsymbol{v}) = f(\boldsymbol{x} - \boldsymbol{v}_2, \boldsymbol{v}_1), \tag{2.4}$$

which is typically smooth in $\boldsymbol{v}$ (e.g., exponential family).

(b) **The "Push In" Technique**. This technique can be considered to be dual to "push out" in the sense that one searches for a transformation $\boldsymbol{y} = \boldsymbol{y}(\boldsymbol{x}, \boldsymbol{v}_1)$ such that the distribution of the corresponding random vector $\boldsymbol{x} = \boldsymbol{x}(\boldsymbol{Y}, \boldsymbol{v}_1)$ is *independent* of $\boldsymbol{v}_1$. In this case, $\ell(\boldsymbol{v})$ can be represented as

$$\ell(\boldsymbol{v}) = \int \dot{L}(\boldsymbol{x}, \boldsymbol{v}) \hat{f}(\boldsymbol{x}) d\boldsymbol{x} = \mathbb{E}_{\hat{f}}\{\dot{L}(\boldsymbol{X}, \boldsymbol{v})\}, \tag{2.5}$$

where $\dot{L}(\boldsymbol{x}, \boldsymbol{v}) = L[\boldsymbol{y}(\boldsymbol{x}, \boldsymbol{v}_1), \boldsymbol{v}_2]$, $\hat{f}(\boldsymbol{x})$ is independent of $\boldsymbol{v}_1$ and $\boldsymbol{v}_2$, and $\boldsymbol{v} = (\boldsymbol{v}_1, \boldsymbol{v}_2)$.

2.2 Sensitivity Analysis via the IPA Technique

The main observation of this subsection is that the *infinitesimal perturbation analysis* (IPA) method, pioneered by Ho and his co-workers (see, e.g., Ho et al. 1979, Glasserman 1991), corresponds to the "push in" technique. Furthemore, the case where $\hat{f} = 1$ on the interval $[0, 1)$ (i.e., X has uniform $[0, 1)$ distribution) is of particular interest. Letting $F(y, \boldsymbol{v}_1)$ be the cdf of the random variable Y, the transformation $\boldsymbol{x} = \boldsymbol{x}(\boldsymbol{y}, \boldsymbol{v}_1)$ reduces in this case to

$$x = x(y, \boldsymbol{v}_1) = F(y, \boldsymbol{v}_1), \tag{2.6}$$

or, equivalently,

$$y = F^{-1}(x, \boldsymbol{v}_1). \tag{2.7}$$

We point out that the original IPA approach is constrained by the fact that the transformation (2.6) typically leads to a sample performance function $\dot{L}(\boldsymbol{x}, \boldsymbol{v})$ which is *nondifferentiable* in $\boldsymbol{v}$ (see, e.g., Heidelberger 1986, Heidelberger et al. 1988, L'Ecuyer 1990). As a result, the required interchangeability conditions for expectation and differentiation do not hold. Furthermore, when $\dot{L}(\boldsymbol{v})$ is not available analytically, it typically allows estimation of the sensitivities at a fixed $\boldsymbol{v}$ *only*, whereas the SF method allows estimation of $\nabla^k \ell(\boldsymbol{v})$ *everywhere*. Consequently, when dealing with optimization problems, the IPA approach must rely on slowly-converging iterative algorithms of the stochastic approximation type, and thus on multiple simulations; see, e.g., L'Ecuyer (1992) and Kushner and Clark (1978). In contrast, its SF method counterpart solves an entire constrained optimization problem from a *single sample path* (see Rubinstein and Shapiro 1992). On the other hand, the IPA approach enjoys the advantage that its sensitivity estimators typically have a *smaller variance* than the standard SF estimators (assuming the interchangeability conditions hold).

Example 2.1 Indicator functions. Suppose that we want to estimate $\nabla \ell(\boldsymbol{v})$, where $\ell(\boldsymbol{v}) = P_{\boldsymbol{v}_1}\{L(Y, \boldsymbol{v}_2) \leq \alpha\}$, L being, say, the waiting time in a $GI/G/1$ queue. The IPA approach will not work, since the inverse transformation (2.7) leads to a piecewise constant sample function (taking only 0 and 1 values). However, if we represent $\ell(\boldsymbol{v})$ as $\ell(\boldsymbol{v}) = \mathbb{E}_{\boldsymbol{v}_1}\{I_{(-\infty,0)}[L(\boldsymbol{v}_2) - \alpha]\}$, by combining the standard SF approach with the "push out" technique, we can estimate $\nabla \ell(\boldsymbol{v})$, $\boldsymbol{v} = (\boldsymbol{v}_1, \boldsymbol{v}_2)$, *everywhere from a single simulation run.*

We next show how to apply the IPA technique in order to estimate the derivative of the steady-state expected waiting time $\ell(\boldsymbol{v})$ in a $GI/G/1$ queue. To this end, we shall use Lindley's recursive (sample path) equation for the waiting time process L_t,

$$L_1 = 0, \quad L_{t+1} = \max\{0, L_t + Z_t\}, \; t \geq 1, \qquad (2.8)$$

where $Z_t = Y_{1t} - Y_{2(t+1)}$. Assume first that the input sequences are *independent*, and consider separately the SF and the IPA approaches. The naive SF approach based on the representation

$$\nabla^k \ell(\boldsymbol{v}) = \mathbb{E}_g\{L_t \nabla^k \widetilde{W}_t\} \qquad (2.9)$$

fails, since for large t, the variance of the associated SF estimators is very large (see, e.g., Rubinstein and Shapiro 1992). Here $\widetilde{W}_t = \prod_{j=1}^{t} W_j$, $W_j = f(\boldsymbol{Z}, \boldsymbol{v})/g(\boldsymbol{Z})$, and $\boldsymbol{Z} \sim g(\boldsymbol{z})$. Its IPA counterpart produces, however, an unbiased estimator of $\nabla \ell(\boldsymbol{v})$ with manageable variance. The corresponding IPA algorithm for estimating $\nabla \ell(\boldsymbol{v})$ is:

Algorithm 2.1 :

1. *Generate the output process* $\dot{L}_t(\boldsymbol{v})$ *by using Lindley's recursive equation*

$$\dot{L}_1 = 0, \quad \dot{L}_{t+1}(\boldsymbol{v}) = \max\{0, \dot{L}_t + \dot{Z}_t(\boldsymbol{v})\}, \; t \geq 1, \qquad (2.10)$$

where $\dot{Z}_t(\boldsymbol{v}) = F_1^{-1}(U_{1,t}, \boldsymbol{v}_1) - F_2^{-1}(U_{2,t+1}, \boldsymbol{v}_2)$.

2. *Differentiate* $\ell(\boldsymbol{v}) = \mathbb{E}\dot{L}_t(\boldsymbol{v})$ *with respect to* $\boldsymbol{v}$ *(by taking the derivative inside the expected value), where* $\dot{L}_t$ *is given in (2.10). Derive* $\nabla \ell(\boldsymbol{v}) = \mathbb{E}\nabla \dot{L}_t(\boldsymbol{v})$.

3. *Simulate the* $GI/G/1$ *queue for* N *customers and estimate* $\nabla \ell(\boldsymbol{v})$ *as*

$$\nabla \hat{\ell}_N(\boldsymbol{v}) = N^{-1} \sum_{t=1}^{N} \nabla \dot{L}_t(\boldsymbol{v}). \qquad (2.11)$$

Consider now the *autocorrelated case.* The corresponding IPA algorithm for estimating $\nabla \ell(\boldsymbol{v})$ can be readily derived by modifying Algorithm 2.1. Specifically, for a TES sequence we have to replace

$$\dot{Z}_t(\boldsymbol{v}) = F_1^{-1}(U_{1,t}, \boldsymbol{v}_1) - F_2^{-1}(U_{2,t+1}, \boldsymbol{v}_2)$$

in step 2 of Algorithm 2.1 by

$$\begin{aligned} \dot{Z}_t(\boldsymbol{v}) \; &= \; F_1^{-1}[\eta_{1t}^{+}(U_{1,t}, \boldsymbol{v}_{11}), \boldsymbol{v}_{21}] \\ &- \; F_2^{-1}[\eta_{2t}^{+}(U_{2,t+1}, \boldsymbol{v}_{12}), \boldsymbol{v}_{22}], \end{aligned} \qquad (2.12)$$

where $\boldsymbol{v}_1 = (\boldsymbol{v}_{11}, \boldsymbol{v}_{12})$, $\boldsymbol{v}_2 = (\boldsymbol{v}_{21}, \boldsymbol{v}_{22})$, and all other data remain the same. A similar approach can be applied to other autocorrelated sequences.

Another situation where IPA estimators may outperform their standard SF counterparts, is a regenerative simulation with long regenerative cycles. The RSF (regenerative score function) estimators are useless here, since the variance of the associated likelihood ratio process $\widetilde{W}_t$ is large (see Rubinstein and Shapiro 1992). The IPA approach will produce estimators which have much lower variance (see Glasserman 1991, L'Ecuyer 1990), provided the interchangeability conditions hold. Note, however, that in this case we may use the following alternatives to IPA: the CSF (conditional score function) estimators of McLeish and Rollans (1992), the DSF (decomposable score function) and TSF (truncated score function) estimators (see Rubinstein 1992b). It is also important to note that the IPA estimators are *consistent*, while their CSF, DSF and TSF counterparts are slightly *biased.* Finally, we remark that when the output process is not regenerative, but stationary and ergodic, we can still use CSF, DSF, TSF and IPA estimators.

2.3 Sensitivity Analysis Via the "Push Out" Technique

In this subsection we apply the "push out" technique to autocorrelated sequences. Note first that for an autocorrelated sequence $\{X_t : t > 0\}$ of the form $X_t = x(X_{t-1}, Y_t, \boldsymbol{v}_2)$, $t = 1, 2, \ldots$, we simply have $\tilde{L}(\underline{x}_t) \equiv L(\underline{x}_t)$, where $\tilde{L}(\underline{x}_t) = L(\underline{y}_t, \boldsymbol{v}_2)$, (see (2.1) and (2.2)). Therefore, application of the "push out" technique just reduces to finding the joint t-dimensional pdf $\tilde{f}_t(x_1, \ldots, x_t, \boldsymbol{v})$ of the random vector $\underline{X}_t = (X_1, \ldots, X_t)$ in terms of the 1-dimensional pdf $f(y, \boldsymbol{v}_2)$ of the random variables Y_t. For Markovian processes $\{X_t : t > 0\}$, the joint pdf $\tilde{f}_t(x_1, \ldots, x_t, \boldsymbol{v})$ can be written as

$$\begin{aligned} &\tilde{f}_t(x_1, \ldots, x_t; \boldsymbol{v}) \; = \\ &\quad \tilde{f}(x_1; \boldsymbol{v}) \cdot \tilde{f}(x_2 | x_1; \boldsymbol{v}) \cdots \tilde{f}(x_t | x_{t-1}; \boldsymbol{v}). \end{aligned} \qquad (2.13)$$

With (2.13) at hand, consider the "push out" estimators for $\nabla^k \ell_1(\boldsymbol{v})$, $\boldsymbol{v} = (\boldsymbol{v}_1, \boldsymbol{v}_2)$. Let $g_t(\underline{\boldsymbol{x}}_t) = g_t(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_t)$ be a pdf dominating the pdf $\tilde{f}_t(\underline{\boldsymbol{x}}_t, \boldsymbol{v}) = \tilde{f}_t(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_t; \boldsymbol{v})$, where $\tilde{f}_t(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_t; \boldsymbol{v})$ is given in (2.13). Assume that $\underline{\boldsymbol{Z}}_t = (\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_t) \sim g_t(\underline{\boldsymbol{z}}_t)$ and the components Z_j, $j = 1, \ldots, t$, of the vector $\underline{\boldsymbol{Z}}_t$ are *dependent*. In this case we can *approximate* $\ell_1(\boldsymbol{v})$ as $\ell_1^r(\boldsymbol{v})$ (the superscript r is the name of the approximate estimator),

$$\ell_1^r(\boldsymbol{v}) = \mathbb{E}_g\{\sum_{t=1}^{\tau} L_t(\underline{\boldsymbol{Z}}_t)\widetilde{W}_t^r(\underline{\boldsymbol{Z}}_t, \boldsymbol{v})\}, \qquad (2.14)$$

where

$$\widetilde{W}_t^r(\underline{\boldsymbol{z}}_t, \boldsymbol{v}) = \frac{\tilde{f}_t(\underline{\boldsymbol{z}}_t, \boldsymbol{v})}{g_t(\underline{\boldsymbol{z}}_t)} = \prod_{j=1}^{t} W_j$$

and

$$W_j = \frac{\tilde{f}(\boldsymbol{z}_j | \boldsymbol{z}_{j-1}; \boldsymbol{v})}{g(\boldsymbol{z}_j | \boldsymbol{z}_{j-1})}.$$

The estimation of $\nabla^k \ell_1^r(\boldsymbol{v})$, $k > 0$, is similar.

We now proceed to derive an expression for $\widetilde{W}_t^r$, driven by TES input sequences. It is straightforward to show that the conditional pdf's $\tilde{f}(x_t | x_{t-1}; \boldsymbol{v})$ of TES sequences can be written in terms of the unconditional pdf $f(y, \boldsymbol{v}_1)$, for $t > 1$, as

$$\tilde{f}_t(x_t | x_{t-1}, \boldsymbol{v}) = \frac{f(x_t, \boldsymbol{v}_1)}{R - L} I\{F(x_t, \boldsymbol{v}_1) \in C[F(x_{t-1}, \boldsymbol{v}_1), L, R]\}, \qquad (2.15)$$

where $\boldsymbol{v}_2 = (L, R)$, $I\{\cdot\}$ is an indicator function, and $C(x, L, R)$ is a circular interval defined in Jagerman and Melamed (1992a). In view of (2.15), the resulting $\widetilde{W}_t^r$ is

$$\widetilde{W}_t^r = \left(\frac{1/(R-L)}{1/(R_0 - L_0)}\right)^t \prod_{j=1}^{t} \frac{f(z_j, \boldsymbol{v}_1)}{f(z_j, \boldsymbol{v}_{01})} \times \prod_{k=1}^{t} I\{F(z_k, \boldsymbol{v}_1) \in C(F(z_{k-1}, \boldsymbol{v}_1), L, R)\}, \qquad (2.16)$$

where zero subscripts pertain to a reference system.

Now let $\boldsymbol{Z}_{11}, \ldots, \boldsymbol{Z}_{\tau_1 1}, \ldots, \boldsymbol{Z}_{1N}, \ldots, \boldsymbol{Z}_{\tau_N N}$ be an autocorrelated sequence of N busy (not necessarily regenerative) cycles generated from the pdf $g(\boldsymbol{z})$. We can estimate $\nabla^k \ell_1(\boldsymbol{v})$, $k = 0, 1, \ldots$, from a *single simulation* as follows:

$$\nabla^k \tilde{\ell}_{1N}^r(\boldsymbol{v}) = N^{-1} \sum_{i=1}^{N} \sum_{t=1}^{\tau_i} L_t(\underline{\boldsymbol{Z}}_{ti}) \nabla^k \widetilde{W}_t^r(\underline{\boldsymbol{Z}}_{ti}, \boldsymbol{v}). \qquad (2.17)$$

It is crucial to realize that the estimators $\nabla^k \tilde{\ell}_{1N}^r(\boldsymbol{v})$ above will be biased. The reason is that the output process $\{L_t\}$ is driven by an autocorrelated sequence $\boldsymbol{Z}_{11}, \ldots, \boldsymbol{Z}_{\tau_1 1}, \ldots, \boldsymbol{Z}_{1N}, \ldots, \boldsymbol{Z}_{\tau_N N}$, and therefore, is not regenerative.

Consider the special case when this sequence is *independent*, i.e., when the dominating density $g_t(\underline{\boldsymbol{z}}_t)$ can be writen as $g_t(\underline{\boldsymbol{z}}_t) = \prod_{j=1}^{t} g(\boldsymbol{z}_j)$. Notice that in this case, the processes $\{L_t(\underline{\boldsymbol{Z}}_{ti})\}$ and $g_t(\underline{\boldsymbol{z}}_t)$ are regenerative, because they are driven by an independent sequence $\boldsymbol{Z}_{11}, \ldots, \boldsymbol{Z}_{\tau_1 1}, \ldots, \boldsymbol{Z}_{1N}, \ldots, \boldsymbol{Z}_{\tau_N N}$. However, the process $\tilde{f}_t(\underline{\boldsymbol{z}}_t, \boldsymbol{v})$ is still not regenerative. We argue, heuristically, that the estimator $\nabla^k \tilde{\ell}_{1N}^r(\boldsymbol{v})$ from (2.17) will typically be only slightly biased. The bias can be further reduced, if the dominating pdf will generate long cycles (this happens when the reference queueing system is simulated in heavy traffic).

The algorithm for estimating the parameter vector $\nabla^k \ell_1(\boldsymbol{v})$, $k > 0$, *everywhere* in $\boldsymbol{v} = (\boldsymbol{v}_1, \boldsymbol{v}_2)$ using the "push out" method can be written as:

Algorithm 2.2 :

1. *Generate an independent sequence* $\boldsymbol{Z}_{11}, \ldots, \boldsymbol{Z}_{\tau_1 1}, \ldots, \boldsymbol{Z}_{1N}, \ldots, \boldsymbol{Z}_{\tau_N N}$ *of* N regenerative *cycles from the dominating pdf* $g_t(\underline{\boldsymbol{z}}_t)$.
2. *Generate the output processes* $L_t(\underline{\boldsymbol{Z}}_{ti})$ *and* $\nabla \widetilde{W}_t^r(\underline{\boldsymbol{Z}}_{ti})$.
3. *Calculate* $\nabla^k \tilde{\ell}_{1N}^r(\boldsymbol{v})$ *from (2.17).*

The algorithm for generating estimators of $\nabla^k \ell(\boldsymbol{v})$ with the "push out" method is similar.

Assume that the process $L_t(\underline{\boldsymbol{X}}_t)$ is stationary and ergodic but not regenerative. In this case we can estimate $\nabla^k \ell(\boldsymbol{v})$, $k = 0, 1$, *simultaneously for different values of* $\boldsymbol{v}_1$ and $\boldsymbol{v}_2$ by using the TSF (truncated score function) estimators

$$\nabla^k \tilde{\ell}_N^{tr}(\boldsymbol{v}, m) = N^{-1} T^{-1} \sum_{i=1}^{N} \sum_{t=1}^{T} \tilde{L}_{ti}(\underline{\boldsymbol{Z}}_{ti}) \nabla^k \widehat{W}_{ti}^{tr}(\underline{\boldsymbol{Z}}_{ti}, \boldsymbol{v}, m). \qquad (2.18)$$

Here the superscripts tr denote quantities associated with the truncated estimators, $\widehat{W}_t^{tr}(\underline{\boldsymbol{Z}}_{ti}, \boldsymbol{v}, m) = \prod_{j=t-m+1}^{t} W_j(Z_{ji}, \boldsymbol{v})$, m is the truncation parameter, N is the number of batches, T is the size of the batch, and $Z_{11}, \ldots, Z_{TN}$ is a random sample from the pdf $g(z)$ which for each $t > 0$ dominates the conditional pdf $\tilde{f}_t(x_t | x_{t-1}, \boldsymbol{v})$.

3 NUMERICAL RESULTS

This section presents simulation results employing direct TSF estimators for the autocorrelated counterparts of the standard $M/G/1$ queue. We shall use the the subscript c to signify that arrivals or services are autocorrelated, e.g., $M_c/G/1$, $M/G_c/1$ and $M_c/G_c/1$. In particular M_c specifies an autocorrelated sequence with exponential marginals.

In our examples, we considered an $M_c/G/1$ queue with gamma-distributed service times (denoted $\Gamma(\mu,\beta)$). We chose an arrival rate $\lambda = 1$, and set the gamma shape parameter to $\beta = 2$. The desired sensitivities were estimated with respect to the gamma scale parameter μ. The performance $\ell(\boldsymbol{v})$ was chosen as the expected steady-state waiting time, where $v_1 = \mu$. The reference parameter was selected to be $\mu_0 = 2$ (corresponding to traffic intensity $\rho_0 = 0.5$), and the truncation parameter was set to $m = 15$. We ran Monte Carlo simulations at the reference parameter above, and used the SF method to do a change-of-measure estimation of $\nabla^k_{v_1}\ell(\boldsymbol{v})$, $k = 0, 1$, at the parameter value $\mu = 5$ ($\rho = 0.4$). In order to compare our results with those obtained from the Crude Monte-Carlo method (without change of measure), we ran additional simulations of the $M_c/G/1$ queue with $\mu = 5$, ($\lambda = 1$, $\beta = 2$).

Table 3.1 in the Appendix presents the point estimators $\nabla^k_{v_1}\bar{\ell}^{tr}_N(\boldsymbol{v},\boldsymbol{v},m)$, $\nabla^k_{v_1}\bar{\ell}^{tr}_N(\boldsymbol{v},\boldsymbol{v}_0,m)$, $k = 0, 1$ (based on $N = 10^5$ customers), and the corresponding half-length (in %) of the relative confidence intervals, denoted $w^k(\boldsymbol{v},\boldsymbol{v},m)$, $k = 0, 1$, and $w^k(\boldsymbol{v},\boldsymbol{v}_0,m)$, $k = 0, 1$, as functions of $\alpha = R - L$. The α parameter is an indication of the magnitude of autocorrelations present in a TES model, and the parameterization (α,ϕ), where $\phi = (R+L)/\alpha$, is equivalent to the parameterization (L, R); see Jagerman and Melamed (1992a) for a detailed discussion of the meaning of α and ϕ. Here $\boldsymbol{v} = (v_1, v_2)$, $v_1 = \mu$, $\boldsymbol{v}_2 = \alpha$ ($\phi = 0$), and similarly for $\boldsymbol{v}_0$. We mention that the lag-1 autocorrelation is a quadratic function of α and ϕ, and the case $\phi = 0$ (equivalently, $L = R$) corresponds to a TES process without drift (see Jagerman and Melamed 1992a).

Table 3.2 in the Appendix presents similar data for an AR(1) model of arrivals. Here the autocorrelation parameter v_2 is the lag-1 autocorrelation of the autoregressive model.

The results of these tables are self-explanatory, showing a good performance of the statistics under change of measure. The results also give us a glimpse of the effect of autocorrelation on waiting time statistics. As expected, increasing autocorrelations give rise to increased waiting times (for TES models, the α magnitude is inversely related to the magnitude of autocorrelation). Furthermore, the effect is more dramatic for TES than for AR(1), since apparently TES autocorrelations decay more slowly. See also Livny et al. (1993) for a similar study.

Our extensive simulation studies of sensitivity and optimization in open queueing networks lead us to make the following recommendations.

[i] Use the direct SF estimators, provided (a) the variances (confidence intervals) are reasonable, (e.g., when the expected number of customers in a busy cycle is, say, less than 10), and (b) the sample performance function $L(\boldsymbol{v}_2)$ is differentiable with respect to $\boldsymbol{v}_2$. (Clearly (b) holds when L does not depend on $\boldsymbol{v}_2$).

[ii] Use low variance DSF (decomposable score function) and TSF (truncated score function) estimators (e.g., Rubinstein 1992b), or the CSF (conditional score function) method of McLeish and Rollans (1992) if (a) fails (the variance is large). Our numerical studies clearly indicate that DSF and TSF estimators are highly efficient when optimizing DES.

[iii] Use the "push out" technique if (b) fails and the required transformation $\boldsymbol{x} = \boldsymbol{x}(\boldsymbol{y},\boldsymbol{v}_2)$ is available.

ACKNOWLEDGMENT

Reuven Y. Rubinstein was supported in this research by the L. Edelstein Research Fund of the Technion - Israel Institute of Technology.

APPENDIX

Table 3.1 TSF Estimators $\nabla_{v_1}^k \bar{\ell}_N^{tr}(\boldsymbol{v}, \boldsymbol{v}, m)$, $\nabla_{v_1}^k \bar{\ell}_N^{tr}(\boldsymbol{v}, \boldsymbol{v}_0, m)$, $k = 0, 1$, as Functions of the Autocorrelation Parameters $\boldsymbol{v}_2 = (\alpha, \phi = 0)$ in a TES Model.

α	without change of measure				with change of measure			
	$\bar{\ell}_N^{tr}(\boldsymbol{v})$	$w^0(\boldsymbol{v})$	$\nabla_{v_1}\bar{\ell}_N^{tr}(\boldsymbol{v})$	$w^1(\boldsymbol{v})$	$\bar{\ell}_N^{tr}(\boldsymbol{v}_0)$	$w^0(\boldsymbol{v}_0)$	$\nabla_{v_1}\bar{\ell}_N^{tr}(\boldsymbol{v}_0)$	$w^1(\boldsymbol{v}_0)$
0.1	0.348	4.38	-0.159	18.83	0.319	4.36	-0.150	18.96
0.2	0.209	4.18	-0.107	15.63	0.208	4.66	-0.095	15.25
0.3	0.176	4.14	-0.087	16.83	0.174	4.43	-0.083	14.55

Table 3.2 TSF Estimators $\nabla_{v_1}^k \bar{\ell}_N^{tr}(\boldsymbol{v}, \boldsymbol{v}, m)$, $\nabla_{v_1}^k \bar{\ell}_N^{tr}(\boldsymbol{v}, \boldsymbol{v}_0, m)$, $k = 0, 1$, as Functions of the Autocorrelation Parameter v_2 in an AR(1) Model.

v_2	without change of measure				with change of measure			
	$\bar{\ell}_N^{tr}(\boldsymbol{v})$	$w^0(\boldsymbol{v})$	$\nabla_{v_1}\bar{\ell}_N^{tr}(\boldsymbol{v})$	$w^1(\boldsymbol{v})$	$\bar{\ell}_N^{tr}(\boldsymbol{v}_0)$	$w^0(\boldsymbol{v}_0)$	$\nabla_{v_1}\bar{\ell}_N^{tr}(\boldsymbol{v}_0)$	$w^1(\boldsymbol{v}_0)$
0.1	0.164	6.05	-0.099	17.38	0.164	6.18	-0.096	13.65
0.2	0.141	6.90	-0.095	17.91	0.137	6.72	-0.091	13.15
0.3	0.128	7.05	-0.091	18.33	0.125	7.89	-0.088	15.74

REFERENCES

Glasserman, P. 1991. *Gradient estimation via perturbation analysis*, Kluwer.

Heidelberger, P. 1986. Limitations of infinitesimal perturbation analysis, Manuscript, IBM T.J. Watson Research Center, Yorktown Heights, New York.

Heidelberger, P., Cao, X.R., Zazanis, M.A. and Suri, R. 1988. Convergence properties of infinitesimal perturbution analysis estimates, *Management Science* 34:1281–1302.

Ho, Y.C., Eyler, M.A. and Chien, T.T. 1979. A gradient technique For general buffer strorage design in a serial production line, *Int. J. on Production Research* 17:557–580.

Jagerman, D.L. and Melamed, B. 1992a. The transition and autocorrelation structure of TES processes part I: general theory, *Stochastic Models* (to appear).

Jagerman, D.L. and Melamed, B. 1992b. The transition and autocorrelation structure of TES processes part II: special cases, *Stochastic Models* (to appear).

Kushner, H.I. and Clarke, D.S. 1978. *Stochastic approximation methods for constrained and unconstrained systems*, Springer-Verlag, Applied Math. Sciences, Vol. 26.

L'Ecuyer, P. 1990. A unified version of the IPA, SF, and LR gradient estimation techniques, *Management Science* 36:1364–1383.

L'Ecuyer, P. 1992. Convergence rates for steady-state derivative estimators. *Annals of Operations Research* (to appear).

Livny M., Melamed B. and Tsiolis A.K. 1993. The impact of autocorrelation on queuing systems. *Management Science* (to appear).

McLeish, D.L. and Rollans, S. 1992. Conditioning for variance reduction in estimating the sensitivity of simulations. *Annals of Operations Research* (to appear).

Melamed, B. 1991. TES: a class of methods for generating autocorrelated uniform variates, *ORSA J. on Computing* 3:317-329.

Rubinstein, R.Y. 1992a. The 'Push Out' method for sensitivity analysis of discrete event systems, *Annals of Operations Research* (to appear).

Rubinstein, R.Y. 1992b. Decomposable score function estimators for sensitivity analysis and optimization of queueing networks, *Annals of Operations Research* (to appear).

Rubinstein, R.Y. and Shapiro, A. 1992. *Discrete event systems: sensitivity analysis and stochastic optimization by the score function method*, Wiley, New York.

AUTHOR BIOGRAPHIES

BENJAMIN MELAMED is Head of the Performance Analysis Department at the C&C Research Laboratories, NEC USA, Inc., in Princeton, NJ. His research interests include systems modeling and analysis, simulation, stochastic processes, and visual modeling environments. Melamed received his Ph.D. in Computer Science from the University of Michigan. Before coming to NEC, he taught in the Department of Industrial Engineering and Management Science at Northwestern University. He later joined the Performance Analysis Department at Bell Laboratories, where he became an AT&T Bell Laboratories Fellow in 1988. He is a Senior Member of IEEE.

REUVEN Y. RUBINSTEIN is a Professor in the Faculty of Industrial Engineering and Management at the Technion — Israel Institute of Technology. He received a Ph.D. degree in Operations Research from the Riga Politechnical Institute (USSR). His research interests are focused on stochastic modeling and optimization, applied probability and discrete event systems.