Mathl Comput. Modelling, Vol. 12, No. 2, pp. 193–219, 1989
Printed in Great Britain

0895-7177/89 $3.00 + 0.00
Pergamon Press plc

SENSITIVITY ANALYSIS AND THE "WHAT IF" PROBLEM IN SIMULATION ANALYSIS

H. ARSHAM,[1] A. FEUERVERGER,[2] D. L. MCLEISH,[3] J. KREIMER[4] and R. Y. RUBINSTEIN[5]

[1]School of Business, University of Baltimore, Baltimore, MD 21201, U.S.A.

[2]Department of Statistics, University of Toronto, Toronto, Ontario M5S 1A1, Canada

[3]Department of Statistics, University of Waterloo, Waterloo, Ontario NZL 3G1, Canada

[4]Department of Industrial Engineering and Management, Ben-Gurion University of Negev, Beer-Sheva 84105, Israel

[5]Faculty of Engineering Management, Technion—Israel Institute of Technology, Technion City, Haifa 32000, Israel and Department of Operations Research, George Washington University, Washington, DC 20052, U.S.A.

(Received and accepted for publication March 1988)

Communicated by D. N. P. Murthy

Abstract—We discuss some known and some new results on the score function (SF) approach for simulation analysis. We show that while simulating a *single sample path* from the underlying system or from an associated system and applying the Radon–Nikodym measure one can: *estimate the performance sensitivities* (gradient, Hessian etc.) of the underlying system with respect to some parameter (vector of parameters); *extrapolate the performance* measure for different values of the parameters; *evaluate the performance* measures of queuing models working in *heavy* traffic by simulating an *associated* (auxilliary) queuing model working in *light* (lighter) traffic; *evaluate the performance* measures of stochastic models while simulating random vectors (say, by the inverse transform method) from an auxiliary probability density function rather than from the original one (say by the acceptance–rejection method). Applications of the SF approach to a broad variety of stochastic models are given.

1. INTRODUCTION

Let

$$l(\mathbf{v}) = E_f[L(\mathbf{Y})] = \int L(\mathbf{y}) f(\mathbf{v}, \mathbf{y})\, d\mathbf{y} \tag{1}$$

be the steady-state performance measure of a stochastic system, where f is a pdf (probability density function), $\mathbf{Y}$ is an RV (random vector) distributed $f(\mathbf{v}, \mathbf{y})$ and $\mathbf{v} \in \mathbf{V}$ is a vector of parameters.

We shall deal here with both DEDS (discrete events dynamic systems) and DESS (discrete events static systems). The main difference between DEDS and DESS is that while the first evolve over time the second do not [1]. Note that DESS assumes using a fixed number of RVs $\mathbf{Y}$, whereas DEDS may require a random number of Ys (e.g. regenerative simulation). Examples of DEDS are queuing networks, and examples of DESS are reliability systems and stochastic networks. For queuing networks, $L(\mathbf{Y})$ might be the time until a certain level is crossed, the mean sojourn time, utilization and throughput, and $f(\mathbf{v}, \mathbf{y})$ might be the multidimensional pdf of the interarrival times, service times or routing probabilities. For a PERT system, $L(\mathbf{Y})$ might be the shortest path and $f(\mathbf{v}, \mathbf{y})$ the multidimensional pdf of the duration of the activities.

It is further assumed that $l(\mathbf{v})$ is not available analytically (because of the complexity of the system) and we have to resort to Monte Carlo simulation.

In this paper we survey the main results from Rubinstein's work [1–6] on the *score function* (SF) approach and present some new results on this subject. More definitely we show that while simulating a *single sample path* from the underlying system or from an associated system and then using the *Radon–Nikodym measure* [7] we can

(i) *estimate simultaneously* the performance $l(\mathbf{v})$ and all its sensitivities (gradient $\nabla l(\mathbf{v})$, Hessian $\nabla^2 l(\mathbf{v})$ etc.);

(ii) *extrapolate the performance* $l(\mathbf{v})$ for different values $\mathbf{v} + \Delta\mathbf{v}^s$, $s = 1,2,\ldots$;
(iii) *evaluate the performance* l for *heavy traffic* queuing models by simulating *associated* queuing models working in *light* (*lighter*) traffic;
(iv) *evaluate the performance* l while generating a stream of RVs from an *auxiliary* pdf, say $g(\mathbf{y})$, from which RVs can be easily generated, rather than from the *original* pdf $f(\mathbf{y})$ from which RV generation is time-consuming.

It is important to point out that issues (i) and (ii) have been extensively treated by Ho and his collaborators using *perturbation analysis* (PA). The PA approach was proposed by Ho, Eyler and Chien in 1979 [8] (see also Refs [9–14] for further references on PA) and the SF approach was proposed by Rubinstein [2] (see also Refs [1, 3–6, 15, 17]). Glynn [18–20] and Reiman and Weiss [21] independently discovered the second approach. Glynn [18–20] made substantial contributions to sensitivity analysis. Rief *et al.* [22, 23] applied similar ideas for deriving sensitivities in radiation transport problems.

We assume further that the *batch means method* [e.g. 24] is applied for performance evaluation of DEDS. Application of our approach to the method of *independent replications*, the *regenerative method* etc. is quite similar [4, 16, 21].

Section 2 deals with sensitivity analysis (estimation of gradient, Hessian etc.) of both DESS and DEDS. Section 3 is devoted to performance extrapolation of $l(\mathbf{v})$; i.e. to evaluation of $l(\mathbf{v})$ for different values of $\mathbf{v} + \Delta\mathbf{v}^s$, $s = 1, 2, \ldots$. Section 4 shows how, using our approach, one can evaluate the performance of a queuing model working in *heavy traffic* while simulating an *associated* (*auxiliary*) queuing model working in *light* (*lighter*) *traffic*. In Section 5, similar ideas will be used for performance evaluation of both DESS or DEDS, while generating a stream of RVs from an *auxiliary* pdf (by using, say, the inverse transform method) rather than generating from the *original* pdf $f(\mathbf{y})$ (by using, say, the time-consuming acceptance–rejection method). In Section 6 we extend our results for stochastic models where both L and f depend on the vector of parameters $\mathbf{v}$. In Section 7, we introduce a nonlinear control random variable procedure for variance reduction. Finally, in Sections 8 and 9 concluding remarks and some ideas for future research are given, respectively.

2. SENSITIVITY ANALYSIS†

2.1. Sensitivity of DESS

Consider the model (1), assuming that the underlying cdf $F(\mathbf{v}, \mathbf{y})$ belongs to a family of absolutely continuous cdfs. The treatment where the $F(\mathbf{v}, \mathbf{y})$ belongs to a family of discrete or mixture distributions is similar. The partial derivative of $l(\mathbf{v})$ [see equation (1)] with respect to v_j, $j = 1, \ldots, n$, is

$$\begin{aligned}\frac{\partial l(\mathbf{v})}{\partial v_j} &= \frac{\partial}{\partial v_j}\int L(\mathbf{y}) f(\mathbf{v}, \mathbf{y})\, d\mathbf{y} \\ &= \int L(\mathbf{y}) \frac{\partial f(\mathbf{v}, \mathbf{y})}{\partial v_j}\, d\mathbf{y} \\ &= \int L(\mathbf{y}) \frac{\partial \ln f(\mathbf{v}, \mathbf{y})}{\partial v_j} f(\mathbf{v}, \mathbf{y})\, d\mathbf{y} \\ &= E\left[L(\mathbf{Y}) \frac{\partial \ln f(\mathbf{v}, \mathbf{Y})}{\partial v_j}\right]; \end{aligned} \tag{2}$$

provided that the operators differentiation and expectation (integration) are interchangeable, $\partial f(\mathbf{v}, \mathbf{y})/\partial v_j$ exists, and $f(\mathbf{v}, \mathbf{y})$ is positive $\forall \mathbf{v} \in \mathbf{V}$, where $\mathbf{V}$ is an open set.

The gradient of $l(\mathbf{v})$ can be written as

$$\nabla l(\mathbf{v}) = E[L(\mathbf{Y}) \nabla \ln f(\mathbf{v}, \mathbf{Y})]. \tag{3}$$

†A large portion of the material of this section is based on Refs [1, 3, 4].

Proceeding with equation (2), we obtain

$$\frac{\partial^2 l(\mathbf{v})}{\partial v_j \partial v_k} = \frac{\partial}{\partial v_k} \int L(\mathbf{y}) \frac{\partial \ln f(\mathbf{v}, \mathbf{y})}{\partial v_j} f(\mathbf{v}, \mathbf{y})\, d\mathbf{y}$$

$$= \int L(\mathbf{y}) \left[\frac{\partial^2 \ln f(\mathbf{v}, \mathbf{y})}{\partial v_j \partial v_k} + \frac{\partial \ln f(\mathbf{v}, \mathbf{y})}{\partial v_j} \frac{\partial \ln f(\mathbf{v}, \mathbf{y})}{\partial v_k} \right] f(\mathbf{v}, \mathbf{y})\, d\mathbf{y}$$

$$= E\left\{ L(\mathbf{Y}) \left[\frac{\partial^2 \ln f(\mathbf{v}, \mathbf{Y})}{\partial v_j \partial v_k} + \frac{\partial \ln f(\mathbf{v}, \mathbf{Y})}{\partial v_j} \frac{\partial \ln f(\mathbf{v}, \mathbf{Y})}{\partial v_k} \right] \right\}. \tag{4}$$

The Hessian of $l(\mathbf{v})$ is therefore

$$\nabla^2 l(\mathbf{v}) = E[L(\mathbf{Y}) H(\mathbf{v}, \mathbf{Y})], \tag{5}$$

$$H(\mathbf{v}, \mathbf{Y}) = \nabla^2 \ln f(\mathbf{v}, \mathbf{Y}) + \nabla \ln f(\mathbf{v}, \mathbf{Y}) \nabla' \ln f(\mathbf{v}, \mathbf{Y}). \tag{6}$$

Here $'$ denotes the transpose operator. Proceeding further with equations (4) and (5), one can readily obtain partial and mixed derivatives of higher order.

Let ϕ denote a linear operator, say differentiation or integration; then formulas (2)–(5) can be generalized as follows:

$$\phi(l(\mathbf{v})) = \phi \int L(\mathbf{y}) f(\mathbf{v}, \mathbf{y})\, d\mathbf{y} = \int L(\mathbf{y}) \phi(f(\mathbf{v}, \mathbf{y})) \frac{f(\mathbf{v}, \mathbf{y})}{f(\mathbf{v}, \mathbf{y})}\, d\mathbf{y} = E_{f(\mathbf{v})} \left[L(\mathbf{Y}) \frac{\phi(f(\mathbf{v}, \mathbf{Y}))}{f(\mathbf{v}, \mathbf{Y})} \right], \tag{7}$$

provided again that the operators ϕ and expectation (integration) are interchangeable. Note that the index $f(\mathbf{v})$ in the last term of formula (7) means that the expectation is taken with respect to $f(\mathbf{v}, \mathbf{y})$. It follows from formula (7) that if $\phi = \nabla$ we obtain formula (3), and if $\phi = \nabla^2$ we obtain formula (5). Note that formula (7) can be further generalized as

$$\phi(l(\mathbf{v})) = \int L(\mathbf{y}) \phi(f(\mathbf{v}, \mathbf{y})) \frac{g(\mathbf{y})}{g(\mathbf{y})}\, d\mathbf{y} = E_g \left[L(\mathbf{Z}) \frac{\phi(f(\mathbf{v}, \mathbf{Z}))}{g(\mathbf{Z})} \right], \tag{8}$$

where $\mathbf{Z}$ is distributed $g(\mathbf{z})$, $g(\mathbf{z})$ corresponds to a probability measure dominating the family of pdfs $\{f(\mathbf{v}, \mathbf{y}), \mathbf{v} \in \mathbf{V}\}$ in the absolute continuous sense, and index g in the last term of formula (8) indicates that the expectation is taken with respect to g. It is clear that in the particular case where $g(\mathbf{z}) = f(\mathbf{v}, \mathbf{z})$, we obtain formula (7). In this section, if not stated otherwise, we assume $g(\mathbf{z}) = f(\mathbf{v}, \mathbf{z})$ and ϕ is a differentiation operator.

Note that formulas (7) and (8) and therefore equations (2)–(5) assume that the operators integration and ϕ are interchangeable. For a rigorous treatment of these issues see Refs [1, 21, 25, 26].

An unbiased estimate of $\phi(l(\mathbf{v}))$ is

$$\bar{\phi}(l(\mathbf{v})) = \frac{1}{N} \sum_{i=1}^{N} L(\mathbf{Y}_i) \frac{\phi(f(\mathbf{v}, \mathbf{Y}_i))}{f(\mathbf{v}, \mathbf{Y}_i)} \tag{9}$$

where $\mathbf{Y}_i$ is distributed $f(\mathbf{v}, \mathbf{y})$. Denote $l(\mathbf{v}) = \nabla^0 l(\mathbf{v})$; then in the particular case $\phi = \nabla^0$, $\phi = \nabla$ and $\phi = \nabla^2$, we obtain

$$\bar{\nabla}^0 l_N(\mathbf{v}) = l_N(\mathbf{v}) = \frac{1}{N} \sum_{i=1}^{N} L(\mathbf{Y}_i(\mathbf{v})), \tag{10}$$

$$\bar{\nabla} l_N(\mathbf{v}) = \frac{1}{N} \sum_{i=1}^{N} L(\mathbf{Y}_i) \nabla \ln f(\mathbf{v}, \mathbf{Y}_i) \tag{11}$$

and

$$\bar{\nabla}^2 l_N(\mathbf{v}) = \frac{1}{N} \sum_{i=1}^{N} L(\mathbf{Y}_i) H(\mathbf{v}, \mathbf{Y}_i) \tag{12}$$

respectively, where $H(\mathbf{v}, \mathbf{Y})$ is given in formula (6).

Since the sensitivity $\bar{\nabla}^r l_N(\mathbf{v})$, $r = 1, 2, \ldots$, contain $\nabla \ln f(\mathbf{v}, \mathbf{y})$ (called in statistics the *efficient score* [e.g. 27, p. 107]), we shall call our method the *score function (SF) method.*

The *advantage* of the SF method is that:

(i) All the unknown quantities $l(\mathbf{v})$, $\nabla l(\mathbf{v})$, $\nabla^2 l(\mathbf{v})$ and higher order partial and mixed derivatives can be estimated *simultaneously* from a *single* simulation.
(ii) In order to find the sensitivities $\phi(l(\mathbf{v}))$ or $l(\mathbf{v})$ we *do not need to differentiate* the sample performance function $L[y(\mathbf{v})]$, which in many cases might not be a smooth (differentiable) function. What we *only need to know* is the sensitivities of $\ln f(\mathbf{v}, \mathbf{y})$; i.e. $\phi(\ln f(\mathbf{v}, \mathbf{y}))$ and $L(\mathbf{Y}_i)$.

In the following examples we shall find the efficient scores and the associated sensitivities for several standard distributions.

Example 1a

Let $\mathbf{Y}_k$, $k = 1, \ldots, m$, be independent RVs each distributed $G(\lambda_k, \beta_k)$, where G denotes a gamma distribution; i.e.

$$f(\lambda_k, \beta_k, y_k) = \frac{\lambda_k \, e^{-\lambda_k y_k} (\lambda_k y_k)^{\beta_k - 1}}{\Gamma(\beta_k)}, \qquad y_k > 0, \lambda_k > 0, \beta_k > 0,$$

and

$$f(\boldsymbol{\lambda}, \boldsymbol{\beta}, \mathbf{y}) = \prod_{k=1}^{m} f(\lambda_k, \beta_k, y_k).$$

Assume that we are interested in the sensitivity with respect to $\boldsymbol{\lambda}$ only. We have

$$\nabla \ln f(\boldsymbol{\lambda}, \boldsymbol{\beta}, \mathbf{y}) = \boldsymbol{\beta}\boldsymbol{\lambda}^{-1} - \mathbf{y}, \tag{13}$$

where

$$\boldsymbol{\lambda}^{-1} = (\lambda_1^{-1}, \ldots, \lambda_m^{-1})', \quad \boldsymbol{\beta} = (\beta_1, \ldots, \beta_m)', \quad \mathbf{y} = (y_1, \ldots, y_m)', \quad \boldsymbol{\lambda}\boldsymbol{\beta} = (\lambda_1\beta_1, \ldots, \lambda_m\beta_m)'$$

and

$$\{H(\boldsymbol{\lambda}, \mathbf{y})\}_{jk} = (\beta_j \lambda_j^{-1} - y_j)(\beta_k \lambda_k^{-1} - y_k) - \delta_{jk}\beta_k \lambda_k^{-2}, \qquad j, k = 1, \ldots, m, \tag{14}$$

where

$$\delta_{jk} = \begin{cases} 1, & j = k \\ 0, & j \neq k. \end{cases}$$

Finally,

$$\nabla l(\boldsymbol{\lambda}) = E[L(\mathbf{Y})(\boldsymbol{\beta}\boldsymbol{\lambda}^{-1} - \mathbf{Y})] \tag{15}$$

and

$$\nabla^2 l(\boldsymbol{\lambda}) = E[L(\mathbf{Y})H(\boldsymbol{\lambda}, \mathbf{Y})]. \tag{16}$$

Example 2a

Let $\mathbf{Y} \sim N(\mu, \Sigma)$; i.e.

$$f(\mu, \mathbf{y}) = \frac{1}{(2\pi)^{m/2} |\Sigma|^{1/2}} \exp\{-\tfrac{1}{2}[(\mathbf{y} - \mu)'\Sigma^{-1}(\mathbf{y} - \mu)]\},$$

where Σ is a positive defined matrix. We have

$$\nabla l(\mu) = E[L(\mathbf{Y})\Sigma^{-1}(\mathbf{Y} - \mu)]$$

$$\nabla^2 l(\mu) = E\{L(\mathbf{Y})[\Sigma^{-1}(\mathbf{Y} - \mu)(\mathbf{Y} - \mu)'\Sigma^{-1} - \Sigma^{-1}]\},$$

respectively.

Example 3a

Let $\mathbf{Y}_k, k = 1, \ldots, m$, be independent RVs each distributed Bernoulli with parameter p_k; i.e.

$$P(p_k, y_k) = p_k^{y_k}(1 - p_k)^{1 - y_k}, \qquad y_k = 0, 1 \quad \text{and} \quad P(p, y) = \prod_{k=1}^{m} P(p_k, y_k).$$

We have

$$[\nabla \ln P(p, \mathbf{y})]_k = \frac{y_k - p_k}{p_k(1 - p_k)},$$

$$[\nabla l(p)]_k = E\left[L(\mathbf{Y}) \frac{y_k - p_k}{p_k(1 - p_k)} \right]$$

and

$$\nabla^2 l(p) = E[L(\mathbf{Y})H(\mathbf{p}, \mathbf{Y})],$$

where

$$\{H(p, \mathbf{y})\}_{jk} = \frac{y_j - p_j}{p_j(1 - p_j)} \frac{y_k - p_k}{p_k(1 - p_k)} - \delta_{jk} \frac{y_k - 2p_k y_k + p_k^2}{p_k^2(1 - p_k)^2}, \qquad j, k = 1, \ldots, m.$$

Example 4a: exponential family

Suppose that $\mathbf{Y}$ has the pdf

$$f(\mathbf{v}, \mathbf{y}) = \exp[a(\mathbf{v})b(\mathbf{y}) + c(\mathbf{v}) + d(\mathbf{y})].$$

Then

$$\nabla \ln f(\mathbf{v}, \mathbf{y}) = \nabla a(\mathbf{v})b(\mathbf{y}) + \nabla c(\mathbf{v})$$

and

$$H(\mathbf{v}, \mathbf{y}) = \nabla^2 a(\mathbf{v})b(\mathbf{y}) + \nabla^2 c(\mathbf{v}) + [\nabla a(\mathbf{v})b(\mathbf{y}) + \nabla c(\mathbf{v})]^2.$$

As examples of DESS consider a reliability system and a stochastic PERT network.

(i) Reliability system. The mean lifetime of a coherent reliability system can be written [3, Sect. 1.1] as

$$l(\mathbf{v}) = E\left[\max_{j=1,\ldots,p} \min_{i \in L_j} \mathbf{Y}_i \right], \tag{17}$$

where L_j is the jth complete path from a source to a sink in the system, $\mathbf{Y}_i$, $i = 1, \ldots, m$, are the durations (lifetimes) of the components with cdfs $F(\mathbf{v}_i, \mathbf{y})$ depending on a parameter (vector) $\mathbf{v}_i$, $i = 1, \ldots, m$, and p is the number of complete paths in the system.

(ii) Stochastic network. The mean shortest path (the minimal project duration) in a stochastic PERT network can be written [3, Section 1.1] as

$$l(\mathbf{v}) = E\left[\min_{j=1,\ldots,p} \sum_{i \in L_j} \mathbf{Y}_i \right], \tag{18}$$

where p, L_j, $\mathbf{Y}_i$ and $\mathbf{v}_i$ have meanings similar to those in equation (17).

Clearly

$$\bar{l}_N(\mathbf{v}) = \frac{1}{N} \sum_{s=1}^{N} \left(\max_{j=1,\ldots,p} \min_{i \in L_j} \mathbf{Y}_{is} \right)$$

and

$$\bar{l}_N(\mathbf{v}) = \frac{1}{N} \sum_{s=1}^{N} \left(\min_{j=1,\ldots,p} \sum_{i \in L_j} \mathbf{Y}_{is} \right)$$

are unbiased estimates for equations (17) and (18), respectively.

We shall now derive the sensitivities $\nabla l(\mathbf{v})$, $\nabla^2 l(\mathbf{v})$ and their corresponding estimates $\bar{\nabla} l_N(\mathbf{v})$ and $\bar{\nabla}^2 l_N(\mathbf{v})$ for the reliability model (17). We have from equations (17), (3) and (5),

$$\nabla l(\mathbf{v}) = E\left[\left(\max_{j=1,\ldots,p} \min_{i\in L_j} \mathbf{Y}_i\right)\nabla \ln f(\mathbf{v}, \mathbf{Y})\right] \tag{19}$$

and

$$\nabla^2 l(\mathbf{v}) = E\left[\left(\max_{j=1,\ldots,p} \min_{i\in L_j} \mathbf{Y}_i\right)H(\mathbf{v}, \mathbf{Y})\right], \tag{20}$$

where $H(\mathbf{v}, \mathbf{Y})$ is given in equation (6).

Consider Example 1a; i.e. assume that $\mathbf{Y}_k$ are independent, each distributed $G(\lambda_k, \beta_k)$, $k = 1, \ldots, m$. Substituting equations (13) and (14) into equations (19) and (20), respectively, we obtain

$$\nabla l(\boldsymbol{\lambda}) = E\left[\max_{j=1,\ldots,p} \min_{i\in L_j} \mathbf{Y}_i(\boldsymbol{\beta}\boldsymbol{\lambda}^{-1} - \mathbf{Y})\right]$$

and

$$\nabla^2 l(\boldsymbol{\lambda}) = E\left[\max_{j=1,\ldots,p} \min_{i\in L_j} \mathbf{Y}_i H(\boldsymbol{\lambda}, \mathbf{Y})\right].$$

The estimates $\bar{\nabla} l(\boldsymbol{\lambda})$ and $\bar{\nabla}^2 l(\boldsymbol{\lambda})$ of $\nabla l(\boldsymbol{\lambda})$ and $\nabla^2 l(\boldsymbol{\lambda})$ are

$$\bar{\nabla} l_N(\boldsymbol{\lambda}) = \frac{1}{N}\sum_{s=1}^{N}\left[\max_{j=1,\ldots,p} \min_{i\in L_j} \mathbf{Y}_{is}(\boldsymbol{\beta}\boldsymbol{\lambda}^{-1} - \mathbf{Y}_s)\right] \tag{21}$$

and

$$\{\bar{\nabla}^2 l(\boldsymbol{\lambda})\}_{jk} = \frac{1}{N}\sum_{s=1}^{N}\left[\max_{j=1,\ldots,p} \min_{i\in L_j} \mathbf{Y}_{is}\{(\beta_j\lambda_j^{-1} - \mathbf{Y}_{js})(\beta_k\lambda_k^{-1} - \mathbf{Y}_{ks}) - \delta_{jk}\beta_k\lambda_k^{-2}\}\right], \quad j, k = 1, \ldots, m, \tag{22}$$

respectively.

2.2. *Sensitivity of DEDS*

Let $\{L_t: t > 0\}$ be the stochastic process under consideration. Assume that it is strictly stationary with $E(L_t^2) < \infty$, and depending on whether $(L_t, t > 0)$ is a continuous-time or discrete-time process, the steady-state mean $E(L_t)$ can be estimated by the time average as

$$\bar{l}_T = T^{-1}\int_0^T L_t \, dt$$

and

$$\bar{l}_T = T^{-1}\sum_{t=1}^{T} L_t,$$

respectively.

If we think of L_t as the queue length at time t, then $\bar{l}_T$ above would be a natural estimator of the steady-state mean queue length $E(L_t)$; if we think of L_t as the waiting time of the tth customer, then $\bar{l}_T$ in the alternative equation above would be an estimator of the steady-state mean waiting time. We also call L_t the *steady-state sample performance* at time t. For a typical DEDS, we write

$$L_t = L_t(\underline{\mathbf{Y}}_t) = L_t\{\mathbf{Y}_t, \mathbf{Y}_{t-1}, \ldots, \mathbf{Y}_0, \mathbf{Y}_{-1}, \mathbf{Y}_{-2}, \ldots\},$$

where $\underline{\mathbf{Y}}_t = (\mathbf{Y}_t, \mathbf{Y}_{t-1}, \ldots, \mathbf{Y}_0, \mathbf{Y}_{-1}, \mathbf{Y}_{-2}, \ldots), \mathbf{Y}_t, \mathbf{Y}_{t-1}, \ldots,$ is a sequence of i.i.d. RVs (input sequence) driving the output process L_t.

Example 5: GI/G/1 queue

Let l be the mean sojourn time of a customer in the GI/G/1 queue. Denote by X_t the steady-state waiting time of the tth customer in the queue. It is well-known [28] that

$$X_{t+1} = \max\{0, X_t - \mathbf{Y}_{1t} + \mathbf{Y}_{2t}\},$$

where $\mathbf{Y}_{1t}$ and $\mathbf{Y}_{2t}$ are the interarrival and the service times of the tth customer, respectively. In this case one can estimate l by

$$\bar{l}_T = \frac{1}{T}\sum_{t=1}^{T} X_t.$$

The mean steady-state performance can be written as

$$l(\mathbf{v}) = EL_t = EL_t(\underline{\mathbf{Y}}_t).$$

As we mentioned in the Introduction we shall apply here the SF approach to the batch means method. Its application to independent replication and the regenerative method can be found in Refs [4, 6, 21]. According to the batch means method we run the system until it reaches the steady-state, then we collect $M = NT$ observations (NT is, say, the number of customers commencing service at a particular service station), where N is the number of batches and T is the size of each batch, and estimate the steady-state performance

$$l(\mathbf{v}) = E(L_t) = E[L_t(\underline{\mathbf{Y}}_t)] \tag{23}$$

as

$$\bar{l}_{N,T}(\mathbf{v}) = \frac{1}{N}\frac{1}{T}\sum_{i=1}^{N}\sum_{t=1}^{T} L_{ti}(\underline{\mathbf{Y}}_{ti}), \tag{24}$$

where

$$\underline{\mathbf{Y}}_{ti} = (\mathbf{Y}_{ti}, \mathbf{Y}_{(t-1)i}, \ldots, \mathbf{Y}_{0i}, \mathbf{Y}_{-1i}, \mathbf{Y}_{-2i}, \ldots).$$

Clearly $\bar{l}_{N,T}(\mathbf{v})$ is an unbiased estimator of $l(\mathbf{v})$. Note that the batch size T is typically chosen [29] such that the correlation between L_t and L_{t+T} is negligible.

To derive the sensitivity estimators with the SF approach for the batch means we argue as follows:

(a) Define first

$$\underset{\sim}{l}(M, \mathbf{v}) = E\underset{\sim}{L}_t = E[\underset{\sim}{L}_t(\underline{\mathbf{Y}}_t(M))] = \int \underset{\sim}{L}_t(\underline{\mathbf{y}}_t)\, \underset{\sim}{f}_t(M, \mathbf{v}, \underline{\mathbf{y}}_t)\mathrm{d}\underline{\mathbf{y}}_t, \tag{25}$$

where

$$\underset{\sim}{f}_t(M, \mathbf{v}, \mathbf{y}_t) = \prod_{j=t-M+1}^{t} f(\mathbf{v}, \mathbf{y}_j), \tag{26}$$

$\underline{\mathbf{Y}}_t(M) = (\mathbf{Y}_t, \mathbf{Y}_{t-1}, \ldots, \mathbf{Y}_{t-M+1})$ and $\underset{\sim}{L}_t(\underline{\mathbf{Y}}_t(M))$ is the truncated version of $L_t(\underline{\mathbf{Y}}_t)$, and is called the *M-dependent process*. Note also that since $\underline{\mathbf{Y}}_t(M)$ presents a truncated version of $\underline{\mathbf{Y}}_t = (\mathbf{Y}_t, \mathbf{Y}_{t-1}, \ldots)$ in general $\underset{\sim}{l}(M, \mathbf{v}) \neq l(\mathbf{v})$ (for more details see below). We shall call $\underset{\sim}{L}_t$ the *truncated sample performance*.

As an estimator of $\underset{\sim}{l}(M, \mathbf{v})$ consider

$$\underset{\sim}{\bar{l}}_{N,T}(M, \mathbf{v}) = \frac{1}{N}\frac{1}{T}\sum_{i=1}^{N}\sum_{t=1}^{T} \underset{\sim}{L}_{ti}(\underline{\mathbf{Y}}_{ti}(M)).$$

Clearly

$$E[\underset{\sim}{\bar{l}}_{N,T}(M, \mathbf{v})] = \underset{\sim}{l}(M, \mathbf{v}) \neq l(\mathbf{v}),$$

i.e. $\underset{\sim}{\bar{l}}_{N,T}(M, \mathbf{v})$ is a biased estimator of $l(\mathbf{v})$. If, however, M is large enough, say, M is the proper batch size, i.e. $M = T$, then the truncation effect of $\underset{\sim}{f}_t(M, \mathbf{v}, \mathbf{y}_t)$

on the bias of $\underset{\sim}{\bar{l}}_{N,T}(M, \mathbf{v})$ is negligible. We shall bear in mind further that $M = T$. Note that $\underset{\sim}{\bar{l}}_{N,T}(M, \mathbf{v})$ is an auxiliary estimator, it is introduced for clarification purposes only and it is typically not available from simulation.

(b) Apply the linear operator ϕ to $\underset{\sim}{l}(M, \mathbf{v})$, i.e. [see formula (7)] write

$$\begin{aligned}\phi[\underset{\sim}{l}(M, \mathbf{v})] &= \phi\{E_{\underset{\sim}{f}(\mathbf{v})}[\underset{\sim}{L}_t(\underline{\mathbf{Y}}_t(M))]\} \\ &= \int \underset{\sim}{L}_t(\underline{\mathbf{y}}_t)\phi[\underset{\sim}{f}_t(M, \mathbf{v}, \underline{\mathbf{y}}_t)]\, \mathrm{d}\underline{\mathbf{y}}_t \\ &= E_{\underset{\sim}{f}(\mathbf{v})}\left[\underset{\sim}{L}_t(\underline{\mathbf{Y}}_t(M))\frac{\phi[\underset{\sim}{f}_t(M, \mathbf{v}, \underline{\mathbf{Y}}_t)]}{\underset{\sim}{f}_t(M, \mathbf{v}, \underline{\mathbf{Y}}_t)}\right] \\ &= E_{\underset{\sim}{f}(\mathbf{v})}[\underset{\sim}{L}_t \mathbf{V}_t(M)], \end{aligned} \tag{27}$$

provided interchangeability between the operator ϕ and the expectation is available. Here $\underset{\sim}{f}(\mathbf{v})$ indicates that the expectation is taken with respect to $\underset{\sim}{f}_t(M, \mathbf{v}, \underline{\mathbf{Y}}_t)$ and

$$\mathbf{V}_t(M) = \frac{\phi[\underset{\sim}{f}_t(M, \mathbf{v}, \underline{\mathbf{Y}}_t)]}{\underset{\sim}{f}_t(M, \mathbf{v}, \underline{\mathbf{Y}}_t)}.$$

As an unbiased estimator of $\phi[\underset{\sim}{l}(M, \mathbf{v})]$ consider

$$\phi[\underset{\sim}{\bar{l}}_{N,T}(M, \mathbf{v})] = \frac{1}{N}\frac{1}{T}\sum_{i=1}^{N}\sum_{t=1}^{T}\underset{\sim}{L}_{ti}\mathbf{V}_{ti}(M), \tag{28}$$

where

$$\underset{\sim}{L}_{ti} = \underset{\sim}{L}_{ti}[\underline{\mathbf{Y}}_{ti}(M)], \tag{29a}$$

$$\mathbf{V}_{ti}(M) = \mathbf{V}_{ti}(M, \mathbf{v}, \underline{\mathbf{Y}}_{ti}) = \frac{\phi[\underset{\sim}{f}_{ti}(M, \mathbf{v}, \underline{\mathbf{Y}}_{ti})]}{\underset{\sim}{f}_t(M, \mathbf{v}, \underline{\mathbf{Y}}_{ti})} \tag{29b}$$

and

$$\underset{\sim}{f}_{ti}(M) = \underset{\sim}{f}_{ti}(M, \mathbf{v}, \underline{\mathbf{Y}}_{ti}) = \prod_{j=t-M+1}^{t} f(\mathbf{v}, \mathbf{Y}_{ji}).$$

Clearly, $\phi[\underset{\sim}{\bar{l}}_{N,T}]$ is a biased estimator of $\phi[l(\mathbf{v})]$ for the same reason that $\underset{\sim}{\bar{l}}_{N,T}$ is a biased estimator of $l(\mathbf{v})$.

(c) Since the sample performance $L_t(\underline{\mathbf{Y}}_t)$ is available from simulation we can use instead of $\phi[\underset{\sim}{\bar{l}}_{N,T}(M, \mathbf{v})]$ the following estimator of $\phi[l(\mathbf{v})]$:

$$\phi[\bar{l}_{N,T}(M, \mathbf{v})] = \frac{1}{N}\frac{1}{T}\sum_{i=1}^{N}\sum_{t=1}^{T} L_{ti}(\underline{\mathbf{Y}}_{ti})\mathbf{V}_{ti}(M, \mathbf{v}, \underline{\mathbf{Y}}_t) = \frac{1}{N}\frac{1}{T}\sum_{i=1}^{N}\sum_{t=1}^{T} L_{ti}V_{ti}(M), \tag{30}$$

where $\mathbf{V}_{ti}(M) = \mathbf{V}_{ti}(M, \mathbf{v}, \underline{\mathbf{Y}}_t)$ is given by equation (29b). The estimator $\phi[\bar{l}_{N,T}(M, \mathbf{v})]$ will be used as our basic estimator for the sensitivities $\phi[l(\mathbf{v})]$. Note that since $E(L_{ti}) = l(\mathbf{v})$ and $E(\underset{\sim}{L}_{ti}) = \underset{\sim}{l}(M, \mathbf{v}) \neq l(\mathbf{v})$, the estimator $\phi[\bar{l}_{N,T}]$ is typically less biased than the estimator $\phi[\underset{\sim}{\bar{l}}_{N,T}]$.

In the particular case where $\phi = \nabla^0$, $\phi = \nabla$ and $\phi = \nabla^2$, we obtain

$$\bar{\nabla}^0 l_{N,T}(M, \mathbf{v}) = \bar{l}_{N,T}(\mathbf{v}) = \frac{1}{N}\frac{1}{T}\sum_{i-1}^{N}\sum_{t-1}^{T} L_{ti}(\underline{\mathbf{Y}}_{ti}), \tag{31}$$

$$\bar{\nabla} l_{N,T}(M, \mathbf{v}) = \frac{1}{N}\frac{1}{T}\sum_{i=1}^{N}\sum_{t=1}^{T} L_{ti}\mathbf{S}_{ti}(M) \tag{32}$$

and

$$\bar{\nabla}^2 l_{N,T}(M, \mathbf{v}) = \frac{1}{N}\frac{1}{T}\sum_{i=1}^{N}\sum_{t=1}^{T} L_{ti}\mathbf{H}_{ti}(M), \tag{33a}$$

where

$$\mathbf{S}_{ti}(M) = \frac{\nabla \underset{\sim}{f}_{ti}(M, \mathbf{v}, \underset{\sim}{\mathbf{Y}}_{ti})}{\underset{\sim}{f}_{ti}(M, \mathbf{v}, \underset{\sim}{\mathbf{Y}}_{ti})} = \sum_{j=t-M+1}^{t} \nabla \log f(\mathbf{v}, \mathbf{Y}_{ji}), \tag{33b}$$

and similarly $\mathbf{H}_{ti}(M)$.

Note that the estimator $\bar{\nabla}^0 l_{N,T}(\mathbf{v})$ coincides with the conventional batch means estimator $\bar{l}_{N,T}(\mathbf{v})$ [see equation (24)] and therefore is unbiased for $l(\mathbf{v})$. The reason is that for $\phi = \nabla^0$ we have $\mathbf{V}_{ti}(M) = 1$, $M = 1, 2, \ldots$.

It follows from Theorem 5.3 of Karlin and Taylor [30, p. 488] that if L_t is a mixing process then $L_t\mathbf{S}_t(M)$ and $L_t\mathbf{H}_t(M)$ are stationary and ergodic processes. In the particular case where $L_t(\underset{\sim}{\mathbf{Y}}_t) = \underset{\sim}{L}_t(\underset{\sim}{\mathbf{Y}}_t(M))$, i.e. L_t is by itself an M-dependent process, we have that $\phi[\underset{\sim}{l}(M, \mathbf{v})]] = \phi[l(\mathbf{v})]$ and $E\phi[\underset{\sim}{\bar{l}}_{N,T}(M, \mathbf{v})] = E\phi[\bar{l}_{N,T}(M, \mathbf{v})] = l(\mathbf{v})$, i.e. $\phi[\bar{l}_{N,T}(M, \mathbf{v})]$, is an unbiased estimator of $\phi[l(\mathbf{v})]$ provided the interchangeability conditions hold.

Note that when $M = 1$ the DEDS estimator $\phi[\bar{l}_{N,T}(M, \mathbf{v})]$ reduces to the DESS estimator $\phi[\bar{l}_{N,T}(M, \mathbf{v})] = \phi[\bar{l}_{N,T}(\mathbf{v})]$ [see equation (9)]. Thus, using the SF approach we can estimate *simultaneously* from a *single* simulation experiment *both* the performance measure $l(\mathbf{v})$ and its sensitivities $\phi[l(\mathbf{v})]$.

Noting that $E[\mathbf{S}_{ti}(M)] = 0$ we can rewrite $\nabla \underset{\sim}{l}(M, \mathbf{v})$ as

$$\nabla \underset{\sim}{l}(M, \mathbf{v}) = \text{Covar}[L_t, \mathbf{S}_t(M)] = E(L_t\mathbf{S}_t(M)) - E(L_t)ES_t(M)$$

and define

$$\tilde{\nabla} l_{N,T}(M, \mathbf{v}) = \bar{\nabla} l_{N,T}(M, \mathbf{v}) - \bar{l}\bar{\mathbf{S}}(M) \tag{34}$$

as an alternative to the estimator $\bar{\nabla} l_{N,T}(M, \mathbf{v})$. Here

$$\bar{l} = \frac{1}{N}\frac{1}{T}\sum_{i=1}^{N}\sum_{t=1}^{T} L_{ti} \quad \text{and} \quad \bar{\mathbf{S}}(M) = \frac{1}{N}\frac{1}{T}\sum_{i=1}^{N}\sum_{t=1}^{T} \mathbf{S}_{ti}(M).$$

Noting also that $E(\mathbf{H}_{ti}) = 0$, we can define in analogy to equation (34)

$$\tilde{\nabla}^2 l_{N,T}(M, \mathbf{v}) = \bar{\nabla}^2 l_{N,T}(M, \mathbf{v}) - \bar{l}\bar{\mathbf{H}} \tag{35}$$

as an alternative to the estimator $\bar{\nabla}^2 l_{N,T}(M, \mathbf{v})$. For simplicity of notation (if no misinterpretation occurs), we suppress further M; otherwise we write T instead of M since we have assumed that $M = T$.

Note that in order to compute $\bar{\nabla} l_{N,T}(\mathbf{v})$ and $\bar{\nabla}^2 l_{N,T}(\mathbf{v})$ one has to compute L_{ti}, $\mathbf{S}_{ti}$ and $\mathbf{H}_{ti}$ and then apply formulas (32) and (33) or (34) and (35), respectively. Note also that computation of $\mathbf{S}_{ti}$ and $\mathbf{H}_{ti}$ is generally less time-consuming than that of L_{ti}.

Now let us find $\mathbf{S}$ and $\mathbf{H}$ for the distributions in Examples 1a–4a.

Example 1b

Let $\mathbf{Y}_k$, $k = 1, \ldots, m$, be independent each distributed $G(\lambda_k, \beta_k)$. We have

$$\mathbf{S}_T = \sum_{t=1}^{T} \nabla \ln f(\boldsymbol{\lambda}, \boldsymbol{\beta}, \mathbf{Y}_t) = \boldsymbol{\lambda}^{-1} T\boldsymbol{\beta} - \sum_{t=1}^{T} \mathbf{Y}_t, \tag{36}$$

where

$$\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_m)', \quad \boldsymbol{\beta} = (\beta_1, \ldots, \beta_m)', \quad \boldsymbol{\lambda}^{-1} = (\lambda_1^{-1}, \ldots, \lambda_m^{-1})',$$

$$\boldsymbol{\lambda}\boldsymbol{\beta} = (\lambda_1\beta_1, \ldots, \lambda_m\beta_m)', \quad \mathbf{Y}_i = (Y_{i1}, \ldots, Y_{im})',$$

$$\{\mathbf{H}_T\}_{jk} = \left(\lambda_j^{-1} T\beta_j - \sum_{t=1}^{T} Y_{tj}\right)\left(\lambda_k^{-1} T\beta_k - \sum_{t=1}^{T} Y_{tk}\right) - \delta_{jk}\beta_k T\lambda_k^{-2}, \qquad j, k = 1, \ldots, m, \tag{37}$$

and

$$\delta_{jk} = \begin{cases} 1, & j = k \\ 0, & j \neq k. \end{cases} \tag{38}$$

Example 1c

As an example of the above consider an M-station closed (open) queuing network. Assume that $\mathbf{Y}_i \sim G(\lambda_i, \beta_i)$, $i = 1, \ldots, m$, and $\mathbf{Y}_i$ and λ_i denote the service time RV and the service rate, respectively, at station i. Let $\bar{l}_{N,T}(\mathbf{v})$ be the sample utilization at station $j, j = 1, \ldots, M$. Then $\bar{\nabla} l_{N,T}(\boldsymbol{\lambda})$ and $\bar{\nabla}^2 l_{N,T}(\boldsymbol{\lambda})$ [see equations (32) and (34) and equations (33) and (35), respectively] present sensitivity estimates of the utilizations at station j with respect to the service rate vector $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_m)$.

Example 2b

Let $\mathbf{Y} \sim N(\mu, \Sigma)$, For $N = 1$ we have

$$\mathbf{S}_T = \Sigma^{-1}\left(\sum_{t=1}^{T} \mathbf{Y}_t - \mu T\right)$$

and

$$\mathbf{H}_T = \left[\Sigma^{-1}\left(\sum_{t=1}^{T} \mathbf{Y}_t - \mu T\right)\left(\sum_{t=1}^{T} \mathbf{Y}_t - \mu T\right)'\Sigma^{-1} - \Sigma^{-1}\right],$$

respectively.

Example 3b

Let $\mathbf{Y}_k$, $k = 1, \ldots, m$, be independent RVs each distributed Bernoulli with parameter p_k. We have

$$[\mathbf{S}_T]_k = \left[\nabla \ln \prod_{t=1}^{T} P(\mathbf{p}, y_t)\right]_k = \frac{\sum_{t=1}^{T} y_{tk} - p_k T}{p_k(1-p_k)},$$

where $\mathbf{y} = (y_1, \ldots, y_m)'$, $\mathbf{p} = (p_1, \ldots, p_m)$ and

$$\{\mathbf{H}_T\}_{jk} = \frac{\sum_{t=1}^{T} y_{tj} - p_j T}{p_j(1-p_j)} \frac{\sum_{t=1}^{T} y_{tk} - p_k T}{p_k(1-p_k)} - \delta_{jk}\left[\sum_{t=1}^{T} (y_{tk} - 2p y_{tk}) + p_k^2 T\right], \qquad j, k = 1, \ldots, m.$$

Example 4b: exponential family

For $f(\mathbf{v}, \mathbf{y})$ given in Example 4a and $N = 1$ we have

$$\mathbf{S}_T = \nabla a(\mathbf{v}) \sum_{t=1}^{T} b(\mathbf{Y}_t) + T\nabla c(\mathbf{v})$$

and

$$\mathbf{H}_T = \nabla^2 a(\mathbf{v}) \sum_{t=1}^{T} b(\mathbf{Y}_t) + \nabla^2 c(\mathbf{v}) + \left[\nabla a(\mathbf{v}) \sum_{t=1}^{T} b(\mathbf{Y}_t) + \nabla c(\mathbf{v})\right]'\left[\nabla a(\mathbf{v}) \sum_{t=1}^{T} b(\mathbf{y}_t) + \nabla c(\mathbf{v})\right],$$

respectively.

Assuming that $\mathrm{Var}[L_t(\mathbf{Y}_t)] < C < \infty, t \geqslant 0$ and taking into account equations (32) and (33a), it is readily seen [4–6] that typically for $N = 1$ we have

$$\mathrm{Var}\, \bar{\nabla} l_T(\mathbf{v}) = O(T) \tag{39}$$

and

$$\mathrm{Var}\, \bar{\nabla}^2 l_T(\mathbf{v}) = O(T). \tag{40}$$

It follows from equations (39) and (40) that the SF method will not be efficient for large T (e.g. queuing network in heavy traffic). In this case one can use efficient variance-reduction techniques [5, 6, 31] or the cross-spectral method [32].

Let $\mathbf{v}$ now be a discrete rather than a continuous parameter. Assuming for simplicity $\mathbf{v} \in \mathbf{V} \subset \mathbf{R}^1$, we can use the following finite-difference analog of $\nabla l(\mathbf{v})$

$$\nabla l(\mathbf{v}) = \frac{1}{\Delta \mathbf{v}} [l(\mathbf{v}') - l(\mathbf{v})], \tag{41}$$

where

$$\left.\begin{aligned} l(\mathbf{v}) &= EL_t = E_{f(\mathbf{v})}[L_t(\underline{\mathbf{Y}}_t)]; \\ l(\mathbf{v}') &= E_{f(\mathbf{v}')}[L_t(\underline{\tilde{\mathbf{Y}}}_t)]; \\ \mathbf{v}' &= \mathbf{v} + \Delta \mathbf{v}; \\ \underline{\tilde{\mathbf{Y}}}_t &= (\tilde{\mathbf{Y}}_t, \tilde{\mathbf{Y}}_{t-1}, \ldots); \\ \tilde{\mathbf{Y}}_j &\sim f(\mathbf{v}', \mathbf{y}), \\ j &= t, t-1, \ldots \end{aligned}\right\}. \tag{42}$$

Note that the subscript $f(\mathbf{v}')$ in formulas (42) means that the expectation is taken with respect to $f_t(\mathbf{v}', \underline{\mathbf{Y}}_t)$, $\underline{\mathbf{Y}}_t = (\mathbf{Y}_t, \mathbf{Y}_{t-1}, \ldots)$.

We shall call $L_t(\underline{\mathbf{Y}}_t)$ and $L_t(\underline{\tilde{\mathbf{Y}}}_t)$ in formulas (23) and (42) the *nominal* and the *perturbed sample performance* functions, respectively.

To estimate $\hat{\nabla} l(\mathbf{v})$ we estimate first $l(\mathbf{v})$ [see also formula (24)] and $l(\mathbf{v}')$ as

$$\bar{l}_{N,T}(\mathbf{v}) = \bar{l}_{f(\mathbf{v})}(\mathbf{v}) = \frac{1}{N} \frac{1}{T} \sum_{i=1}^{N} \sum_{t=1}^{T} L_{ti}(\underline{\mathbf{Y}}_{ti}) \tag{43}$$

and

$$\bar{l}_{N,T}(\mathbf{v}') = \bar{l}_{f(\mathbf{v}')}(\mathbf{v}') = \frac{1}{N} \frac{1}{T} \sum_{i=1}^{N} \sum_{t=1}^{T} L_{ti}(\underline{\tilde{\mathbf{Y}}}_{ti}), \tag{44}$$

respectively. With $\bar{l}_{f(\mathbf{v})}$ and $\bar{l}_{f(\mathbf{v}')}$ at our disposal we can define now the following two estimates for $\hat{\nabla} l(\mathbf{v})$:

$$\nabla l(\mathbf{v})_{\mathrm{N1}} = \frac{1}{N \Delta \mathbf{v}} \sum_{i=1}^{N} \sum_{t=1}^{T} L_{ti}(\underline{\tilde{\mathbf{Y}}}_{ti}) - \sum_{i=n+1}^{2N} \sum_{t=1}^{T} L_{ti}(\underline{\mathbf{Y}}_{ti}) \tag{45}$$

and

$$\nabla l(\mathbf{v})_{\mathrm{N2}} = \frac{1}{N \Delta \mathbf{v}} \sum_{i=1}^{N} \sum_{t=1}^{N} L_{ti}(\underline{\tilde{\mathbf{Y}}}^{\mathrm{c}}_{ti}) - L_{ti}(\underline{\mathbf{Y}}_{ti}), \tag{46}$$

where it is assumed that for each i the components of the random vectors $\underline{\mathbf{Y}}_{ti}$ and $\underline{\tilde{\mathbf{Y}}}_{ti}$ in formula (45) are uncorrelated, while the corresponding components of $\underline{\mathbf{Y}}_{ti}$ and $\underline{\tilde{\mathbf{Y}}}^{\mathrm{c}}_{ti}$ in formula (46) are correlated. More specifically, it is assumed that the random vectors $\underline{\tilde{\mathbf{Y}}}^{\mathrm{c}}_{ti}$ and $\underline{\mathbf{Y}}_{ti}$ use CRN (common random numbers); i.e.

$$\{\underline{\mathbf{Y}}^{\mathrm{c}}_t, \underline{\mathbf{Y}}_t\} = \{F^{-1}(\mathbf{v} + \Delta \mathbf{v}, \underline{\mathbf{U}}_t), F^{-1}(\mathbf{v}, \underline{\mathbf{U}}_t)\}.$$

Here $\underline{\mathbf{U}}_t = \{U_j, j = t - T + 1, \ldots, t\}$, where U_j are iid random variables each distributed $\mathbf{U}(0, 1)$ and F^{-1} is the inverse of the cdf F. We shall call $\nabla l(\mathbf{v})_{\mathrm{N1}}$ and $\nabla l(\mathbf{v})_{\mathrm{N2}}$ the CMC (crude Monte Carlo) and CRN (common random numbers) estimates of $\hat{\nabla} l(\mathbf{v})$, respectively.

It is shown in Refs [11, 14, 25, 26] that generally the CRN estimate (46) is more accurate than its CMC counterpart (45) in the sense that

$$\operatorname{Var} \nabla l(\mathbf{v})_{\mathrm{N2}} < \operatorname{Var} \nabla l(\mathbf{v})_{\mathrm{N1}}.$$

Note that both estimates (45) and (46) require two simulations: one with $\mathbf{v}$ and another with $\mathbf{v} + \Delta \mathbf{v}$. Clearly, when $\mathbf{v} \in \mathbf{R}^n$, each estimate requires at least $(n + 1)$ simulations and therefore, for large n, calculation of $\nabla l(\mathbf{v})_{\mathrm{N2}}$ and $\nabla l(\mathbf{v})_{\mathrm{N2}}$ can be very time-consuming.

Now we introduce an alternative to estimates (45) and (46) which in analogy to the SF estimate $\bar{\nabla} l(\mathbf{v})$, assumes simulation of *the nominal system only*. We argue as follows:

(a) Write $\underset{\sim}{l}(M, \mathbf{v})$ [see formula (25)], with $\mathbf{v}'$ instead of $\mathbf{v}$, as

$$\begin{aligned}\underset{\sim}{l}(M, \mathbf{v}') = E_{f(\mathbf{v}')} L_t[(\underset{\sim}{\mathbf{Y}}_t, M))] &= \int \underset{\sim}{L}_t(\mathbf{y}_t) \underset{\sim}{f}_t(M, \mathbf{v}', \underset{\sim}{\mathbf{y}}_t)\, d\underset{\sim}{\mathbf{y}}_t \\ &\times \int L_t(\underset{\sim}{\mathbf{y}}_t) \frac{\underset{\sim}{f}_t(M, \mathbf{v}', \underset{\sim}{\mathbf{y}}_t)}{\underset{\sim}{f}_t(M, \mathbf{v}, \underset{\sim}{\mathbf{y}}_t)} \underset{\sim}{f}_t(M, \mathbf{v}, \underset{\sim}{\mathbf{y}}_t)\, d\underset{\sim}{\mathbf{y}}_t \\ &= E_t(\mathbf{v})[\underset{\sim}{L}_t(\underset{\sim}{\mathbf{Y}}_t(M)) W_t(M, \underset{\sim}{\mathbf{Y}}_t)] \\ &= E_{f(\mathbf{v})}[L_t W_t(M)], \end{aligned} \tag{47}$$

where

$$W_t(M) = W_t(M, \underset{\sim}{\mathbf{Y}}_t) = \frac{\underset{\sim}{f}_t(M, \mathbf{v}', \underset{\sim}{\mathbf{Y}}_t)}{\underset{\sim}{f}_t(M, \mathbf{v}, \underset{\sim}{\mathbf{Y}}_t)}$$

and

$$\underset{\sim}{f}_t(M, \mathbf{v}', \underset{\sim}{\mathbf{Y}}_t) = \prod_{j=t-M+1}^{t} f(\mathbf{v}', \mathbf{Y}_j).$$

It is important to note that the expectation in the second term of formula (47) is taken with respect to $\underset{\sim}{f}(M, \mathbf{v}', \underset{\sim}{\mathbf{y}}_t)$, while the expectation in the last term of formula (47) is taken with respect to $\underset{\sim}{f}_t(M, \mathbf{v}, \underset{\sim}{\mathbf{y}}_t)$. It follows directly from formula (47) that changing the measure from $\underset{\sim}{f}_t(M, \mathbf{v}', \underset{\sim}{\mathbf{y}}_t)$ to $\underset{\sim}{f}_t(M, \mathbf{v}, \underset{\sim}{\mathbf{y}}_t)$ we can express it as expectation with respect to pdf $\underset{\sim}{f}_t(M, \mathbf{v}, \underset{\sim}{\mathbf{y}}_t)$. Note that formula (47) presents a particular case of the Radon–Nikodym derivative [7].

As an unbiased estimator of formula (47) [see also formula (28)] consider the following auxiliary estimator:

$$\bar{l}_{N,T}(M, \mathbf{v}') = \frac{1}{N} \frac{1}{T} \sum_{i=1}^{N} \sum_{t=1}^{T} \underset{\sim}{L}_{ti} W_{ti}(M), \tag{48}$$

where, in analogy to formula (29),

$$W_{ti}(M) = \frac{\underset{\sim}{f}_{ti}(M, \mathbf{v}', \underset{\sim}{\mathbf{Y}}_{ti})}{\underset{\sim}{f}_{ti}(M, \mathbf{v}, \underset{\sim}{\mathbf{Y}}_{ti})}. \tag{49}$$

Clearly, $\underset{\sim}{l}_{N,T}(M, \mathbf{v}')$ is a biased estimator of $l(\mathbf{v}^s)$ for the same reason that $\phi[\underset{\sim}{l}_{N,T}(M, \mathbf{v})]$ [see formula (28)] is a biased estimator of $\phi l(\mathbf{v})$.

(b) Since the sample performance $L_t = L_t(\underset{\sim}{\mathbf{Y}}_t)$ is available from simulation, we can use instead of $\bar{l}_{N,T}(M, \mathbf{v}')$ [see also formula (30)] the following estimator for $l(\mathbf{v}')$:

$$\tilde{l}_{N,T}(M, \mathbf{v}') = \tilde{l}_{\mathbf{v}}(M, \mathbf{v}') = \frac{1}{N} \frac{1}{T} \sum_{i=1}^{N} \sum_{t=1}^{T} L_{ti} W_{ti}(M). \tag{50}$$

The estimator $\tilde{l}_{N,T}$ will be used in Section 3 as our basic estimator for performance extrapolation. Note that $\tilde{l}_{N,T}$ is typically less biased than $\underset{\sim}{l}_{N,T}$ for the same reason that $\phi[\underset{\sim}{l}_{N,T}]$ is less biased than $\phi[\bar{l}_{N,T}]$. Also, as for $\phi[\bar{l}_{N,T}]$, we assume in formula (50) that $M = T$, where T is the proper batch size which is chosen such that the truncation effect of $W_{ti}(M)$ on the bias of $\tilde{l}_{N,T}(M, \mathbf{v}')$ is negligible.

(c) Finally, as an estimator of $\nabla l(\mathbf{v})$ [an alternative to estimators (45) and (46)] we define

$$\begin{aligned}\nabla l(\mathbf{v})_{N3} = \nabla l(M, \mathbf{v})_{N3} &= \frac{1}{\Delta \mathbf{v}} [\tilde{l}_{N,T}(M, \mathbf{v}') - \bar{l}_{N,T}(\mathbf{v})] \\ &= \frac{1}{\Delta \mathbf{v}} \frac{1}{N} \frac{1}{T} \sum_{i=1}^{N} \sum_{t=1}^{T} L_{ti}[W_{ti}(M) - 1]. \end{aligned} \tag{51}$$

Note that if L_t is an M-dependent process then the estimator $\tilde{l}_{N,T}(M, \mathbf{v}')$ is unbiased for $l(\mathbf{v}')$. Note again that in the particular case where $M = 1$ DEDS estimators reduce to DESS estimators.

· As before we shall use further T instead of M and if no misinterpretation occurs we shall suppress M completely.

Returning to the case where $\mathbf{v}$ is a continuous parameter, we readily obtain

$$\lim_{\Delta\mathbf{v}\to 0} \hat{\nabla} l(\mathbf{v}) = \lim_{\Delta\mathbf{v}\to 0} \frac{1}{\Delta\mathbf{v}} E_{f(\mathbf{v})}\{L_t(\underline{\mathbf{Y}}_t)[W_t(\underline{\mathbf{Y}}_t) - 1]\} = E_{f(\mathbf{v})}[L_t(\underline{\mathbf{Y}}_t)\tilde{\mathbf{S}}_t(M, \mathbf{v}, \underline{\mathbf{Y}}_t)] = \nabla l(M, \mathbf{v}) \tag{52}$$

and

$$\lim_{\Delta\mathbf{v}\to 0} \nabla l(\mathbf{v})_{\mathrm{N3}} = \frac{1}{N}\frac{1}{T}\sum_{i=1}^{N}\sum_{t=1}^{T} L_{ti}\mathbf{S}_{ti}(M) = \bar{\nabla} l_{N,T}(M, \mathbf{v}), \tag{53}$$

provided that the operators lim and expectation in formula (52) are interchangeable, and $f(\mathbf{v}, \mathbf{y})$ is differentiable with respect to $\mathbf{v}$.

Thus, when $\Delta\mathbf{v}\to 0$, the finite-difference estimate $\hat{\nabla} l(\mathbf{v})$ and the LR estimate $\nabla l(\mathbf{v})_{\mathrm{N3}}$ converge to $\bar{\nabla} l(M, \mathbf{v}) = E[\bar{\nabla} l_{N,T}(M, \mathbf{v})]$ and to its estimate $\bar{\nabla} l(M, \mathbf{v})$, respectively.

Note that the estimates $\nabla l(\mathbf{v})_{\mathrm{N3}}$ and $\tilde{l}_{f(\mathbf{v})}(\mathbf{v} + \Delta\mathbf{v})$ can be readily generalized as

$$\nabla l(\mathbf{v})_{\mathrm{N4}} = \frac{1}{\Delta\mathbf{v}}\frac{1}{N}\frac{1}{T}\sum_{i=1}^{N}\sum_{t=1}^{T} L_{ti}(\underline{\mathbf{Z}}_{ti})\frac{f_{ti}(\mathbf{v} + \Delta\mathbf{v}, \underline{\mathbf{Z}}_{ti}) - f_{ti}(\mathbf{v}, \underline{\mathbf{Z}}_{ti})}{g_{ti}(\underline{\mathbf{Z}}_{ti})} \tag{54}$$

$$\tilde{l}_g(\mathbf{v} + \Delta\mathbf{v}) = \frac{1}{N}\frac{1}{T}\sum_{i=1}^{N}\sum_{t=1}^{T} L_{ti}(\underline{\mathbf{Z}}_{ti})\frac{f_{ti}(\mathbf{v} + \Delta\mathbf{v}, \underline{\mathbf{Z}}_{ti})}{g_{ti}(\underline{\mathbf{Z}}_{ti})} \tag{55}$$

where

$$g_t(\underline{\mathbf{Z}}_t) = \prod_{j=t-T+1}^{t} g(\mathbf{Z}_j) \quad \text{and} \quad \mathbf{Z}_j \sim g(\mathbf{z}).$$

Note also that the pdf $g(\mathbf{z})$ can be chosen with a view to variance reduction; i.e. to achieve $\mathrm{Var}\,\nabla l(\mathbf{v})_{\mathrm{N4}} < \mathrm{Var}\,\nabla l(\mathbf{v})_{\mathrm{N3}}$ and similarly in equation (55). Such a choice of $g(\mathbf{z})$ [e.g. 33] is called *importance sampling*. In Sections 4 and 5 we shall present two alternative applications of the estimate $\tilde{l}_g$ one for estimating performance of queuing networks in heavy traffic and the other to avoid using the acceptance–rejection method for perform evaluation.

It is important to note that in contrast to $\nabla l(\mathbf{v})_{\mathrm{N1}}$ and $\nabla l(\mathbf{v})_{\mathrm{N2}}$, the LR estimate $\nabla l(\mathbf{v})_{\mathrm{N3}}$ [and $\nabla l(\mathbf{v})_{\mathrm{N4}}$] is based on simulation of a *single sample path* $L_t(\underline{\mathbf{Y}}_t)$ [and $L_t(\underline{\mathbf{Z}}_t)$] *only*: it does not assume *any transformation or reconstruction* of it.

3. THE "WHAT IF" PROBLEM (PERFORMANCE EXTRAPOLATION)

This section is based on work by Rubinstein [4–6]. It deals with the so-called "*what if*" problem which can be formulated as follows. What will be the value of the performance measure $l(\mathbf{v})$ of the model (1) if the vector $\mathbf{v}$ is perturbed by $\Delta\mathbf{v}^s$, $s = 1, \ldots, r$? We show that using the SF approach one can estimate (extrapolate) *simultaneously* all the values $l(\mathbf{v} + \Delta\mathbf{v}^s)$, $s = 1, \ldots, r$, from a *single simulation*.

We shall consider the following two approaches: the Radon–Nikodym measure approach and the Taylor series expansion approach

3.1. Radon–Nikodym Measure Approach

This approach is based on the representation of $l(M, \mathbf{v}')$ and $\tilde{l}_{N,T}(M, \mathbf{v}')$, as per formulas (47) and (50), respectively.

Assuming without loss of generality that $N = 1$, we can rewrite formula (50), with $\mathbf{v}$ replaced by $\mathbf{v}^s$, as

$$\tilde{l}_{N,T}(M, S) = \tilde{l}_{f(\mathbf{v})}(\mathbf{v}^s) = \frac{1}{T}\sum_{t=1}^{T} L_t(\underline{\mathbf{Y}}_t)W_t(M, \mathbf{v}, \underline{\mathbf{Y}}_t), \tag{56}$$

where

$$W_t(M, \mathbf{v}, \underset{\sim}{\mathbf{Y}}_t) = W_t(T, \mathbf{v}, \underset{\sim}{\mathbf{Y}}_t) = \frac{f_t(T, \mathbf{v}^s \underset{\sim}{\mathbf{Y}}_t)}{f_t(T, \mathbf{v}, \underset{\sim}{\mathbf{Y}}_t)}, \tag{57}$$

$$f_t(T, \mathbf{v}^s, \underset{\sim}{\mathbf{Y}}_t) = \prod_{j=t-T+1}^{T} f(\mathbf{v}^s, \mathbf{Y}_j).$$

Similarly to the sensitivity estimators $\bar{\nabla}^r l_{N,T}(\mathbf{v})$ we have here:

(i) For $T < \infty$ the estimators $\tilde{l}_{f(\mathbf{v})}(\mathbf{v}^s)$, $s = 1, \ldots, r$, are typically biased since $f_t(\mathbf{v}, \underset{\sim}{\mathbf{Y}}_t)$ and $f_t(\mathbf{v}^s, \underset{\sim}{\mathbf{Y}}_t)$ present truncated versions of $f_t(\mathbf{v}, \underset{\sim}{\mathbf{Y}}_t)$ and $f_t(\mathbf{v}^s, \underset{\sim}{\mathbf{Y}}_t)$, respectively. The bias of $\tilde{l}_{f(\mathbf{v})}$ increases as the batch size T decreases.

(ii) For fixed T and $\Delta\mathbf{v}^s$ the bias of $\tilde{l}_{f(\mathbf{v})}$ increases with the traffic intensity.

(iii) For fixed k and $\Delta\mathbf{v}^s$ the correlation between $L_t W_t$ and $L_{t+k} W_{t+k}$ is typically less than the correlation between L_t and L_{t+k}, i.e.

$$\rho(L_t W_t, L_{t+k} W_{t+k}) < \rho(L_t, W_t).$$

This actually means that the batch size in the estimator $\tilde{l}_{f(\mathbf{v})}(\mathbf{v}^s)$ can be chosen smaller than in the CMC estimator

$$l_{N,T}(M, \mathbf{v}) = \bar{l}_{f(\mathbf{v})}(\mathbf{v}) = \frac{1}{T} \sum_{t=i}^{T} L_t(\underset{\sim}{\mathbf{Y}}_t).$$

Or, in other words, by putting more weights on L_t (multiplying L_t by W_t) one can reduce the batch size.

It follows from equations (56) and (57) that in order to estimate the performance $l(\mathbf{v}^s)$ for different values $\mathbf{v}^s = \mathbf{v} + \Delta\mathbf{v}^s$, $s = 1, \ldots, r$, we only have to perform some simple calculations with L_t and W_t. Note that the extra computation of W_t is usually *small* relative to L_t. Note also that *all* the estimates $\tilde{l}_{f(\mathbf{v})}(\mathbf{v}^s)$, $s = 1, \ldots, r$, are computed *simultaneously* from a *single* simulation of the nominal system and *no transformation* of the nominal sample path is needed: it is taken in its original form. It is crucial to understand that the goal of the Radon–Nikodym measure used in equation (56) is to *transform all perturbed sample paths* $L_t(\underset{\sim}{\mathbf{Y}}_t)$ associated with different values of $\mathbf{v}^S = \mathbf{v} + \Delta\mathbf{v}^s$, $s = 1, \ldots, r$, [see equation (44)] to the *nominal one* $L_t(\underset{\sim}{\mathbf{Y}}_t)$ with $\mathbf{v}^s = \mathbf{v}$. We shall call $\tilde{l}_{f(\mathbf{v})}$ the *LR* (*likelihood ratio*) estimate of $l(\mathbf{v}^s)$.

Example 6: GI/G/1 queue

Let $L_t(\underset{\sim}{\mathbf{Y}}_t)$ be the sample sojourn time of the tth customer in the nominal GI/G/1 queue. In order to estimate the mean sojourn time $l(\mathbf{v} + \Delta\mathbf{v}^s)$ of a customer in the perturbed GI/G/1 queue, we can apply directly equations (56) and (57), with $L_t = X_t$ (see Example 5) and W_t depending on the choice of the interarrival and service time pdfs.

Consider now DESS. Since DESS can be considered as a particular case of DEDS, i.e. with $T = 1$, we have from equation (51),

$$\tilde{l}_{f(\mathbf{v})} = \frac{1}{N} \sum_{i=1}^{N} \left[L(\mathbf{Y}_i) \frac{f(\mathbf{v}^s, \mathbf{Y}_i)}{f(\mathbf{v}, \mathbf{Y}_i)} \right], \qquad s = 1, \ldots, r. \tag{58}$$

Example 7

Consider a coherent reliability system with mean lifetime as in equation (17). We obtain from equation (58),

$$\tilde{l}_{f(\mathbf{v})} = \frac{1}{N} \sum_{k=1}^{N} \left[\left(\max_{j=1,\ldots,p} \min_{i \in L_j} Y_{ik} \right) \frac{f(\mathbf{v}^s, \mathbf{Y}_k)}{f(\mathbf{v}, \mathbf{Y}_k)} \right]. \tag{59}$$

consider now a stochastic PERT network as in equation (18). We have

$$\tilde{l}_{f(\mathbf{v})} = \frac{1}{N} \sum_{k=1}^{N} \left[\left(\min_{j=i,\ldots,p} \sum_{i \in L_j} Y_{jk} \right) \frac{f(\mathbf{v}^s, \mathbf{Y}_k)}{f(\mathbf{v}, \mathbf{Y}_k)} \right]. \tag{60}$$

For another application of the LR estimate $\tilde{l}_{f(\mathbf{v})}$ assume that we want to optimize $l(\mathbf{v})$ with respect to $\mathbf{v}$. The conventional approach uses the CMC estimates (44) and (45), as estimates of $l(\mathbf{v}^s)$ and $\nabla' l(\mathbf{v})$, respectively [e.g. 34]. Clearly, using the estimates $\tilde{l}_{f(\mathbf{v})}$ and $\bar{\nabla}' l(\mathbf{v})$ [see formulas (56), (57), (32) and (33), respectively] instead of the conventional CMC ones we can obtain *tremendous* computational savings while optimizing $l(\mathbf{v})$.

Estimating $l(\mathbf{v}+\Delta\mathbf{v})$ with the LR estimates instead of CMC ones yields computational savings, but reduces precision; i.e. the variance of $\tilde{l}_{f(\mathbf{v})}(\mathbf{v}^s)$ is usually larger than that of $\tilde{l}_{f(\mathbf{v}^s)}(\mathbf{v}^s)$, its CMC counterpart. It is not difficult to show [4–6] that under rather mild conditions on L_t and $W_t(T)$:

$$\text{(i)}\quad \operatorname{Var} \tilde{l}_{f(\mathbf{v})}(\mathbf{v}^s) = \operatorname{Var} L_t W_t(T) \to \infty \text{ as } T \to \infty; \tag{61}$$

$$\text{(ii)}\quad \operatorname{Var} \tilde{l}_{f(\mathbf{v})}(\mathbf{v}^s) = O(\operatorname{Var} W_t(T)); \tag{62}$$

$$\text{(iii)}\quad \operatorname{Var} W_t = E(W_t)^2 - E(W_t)]^2 = \left[1 + O\left(\max_{i,j} \alpha_i^s \alpha_j^s\right)\right]^T - 1,$$

$$\alpha_i = |\Delta v_i / v_i|, \quad i = 1, \ldots, m. \tag{63}$$

It follows from formulas (62) and (63) that both Var W_t and Var $\tilde{l}_{f(\mathbf{v})}$ increase exponentially in t.

The situation is not so bad as one might think from the first glance. The reasons being:

(i) As we pointed out earlier for fixed k the correlation between $L_t W_t$ and $L_{t+k} W_{t+k}$ typically decreases as the variability of W_t increases. This suggests obtaining shorter batches by putting more weight W_t on L_t (multiplying L_t by W_t).

(ii) There are many efficient variance-reduction techniques (see Refs [1, 4–6] and Section 7 of this paper) to improve the accuracy of the estimator $\tilde{l}_{\mathbf{v}}$.

(iii) For small α_i and α_i, $i, j = 1, \ldots, m$, namely, when $T \max_{ij} \alpha_i \alpha_j \leqslant O(1) \ll O(T)$, we readily obtain from formula (63),

$$\operatorname{Var} W_t \approx O\left(T \max_{i,j} \alpha_i \alpha_i\right). \tag{64}$$

We shall discuss the last issue in more detail. Let

$$f(\mathbf{v}, \mathbf{y}) = \prod_{i=1}^{m} f(\mathbf{v}_i, \mathbf{y}_i),$$

where

$$f(\lambda_i, \beta_i, \mathbf{y}) = \frac{\lambda_i^{\beta_i} \exp(-\lambda_i \mathbf{y}) \mathbf{y}^{(\beta_i - 1)}}{\Gamma(\beta_i)}, \tag{65}$$

i.e. $\mathbf{Y}_i \sim G(\mathbf{v}^i)$, $\mathbf{v}_i = (\lambda_i, \beta_i)$, $i = 1, \ldots m$; G denotes a gamma distribution and consider the following two cases:

(a) only a single parameter, say λ_i, out of $2m$ parameters of $\mathbf{v} = (v_1, \ldots, v_m)$ is perturbed.

(b) k out of $2m$ parameters of $\mathbf{v} = (v_1, \ldots, v_m)$ are perturbed.

Case (a). Proposition 1. For $f(\mathbf{v}, \mathbf{y})$ as per formulas (65),

$$E[W_t^2(T)] = E(W_t^2) = \left(1 + \frac{\Delta\lambda^2}{\lambda^2 + 2\lambda\,\Delta\lambda}\right)^{T\beta}. \tag{66}$$

Proof. We have from formulas (57) and (65).

$$W_t(T) = \prod_{t=1}^{T} \frac{f(\boldsymbol{\lambda} + \Delta\boldsymbol{\lambda}, \boldsymbol{\beta}, \mathbf{Y}_t)}{f(\boldsymbol{\lambda}, \boldsymbol{\beta}, \mathbf{Y}_t)}$$

$$= \prod_{t=1}^{T} \frac{(\lambda + \Delta\lambda)^{\beta} \exp[-(\boldsymbol{\lambda} + \Delta\boldsymbol{\lambda})\mathbf{Y}_t]}{\boldsymbol{\lambda}^{\boldsymbol{\beta}} \exp(-\boldsymbol{\lambda}\mathbf{Y}_t)}$$

$$= \left(1 + \frac{\Delta\boldsymbol{\lambda}}{\boldsymbol{\lambda}}\right)^{T\boldsymbol{\beta}} \exp\left(-\Delta\boldsymbol{\lambda} \sum_{t=1}^{T} \mathbf{Y}_t\right). \tag{67}$$

Note that in formula (67) and further on we use for convenience W_T instead of W_t and λ and $\Delta\lambda$ instead of λ_1 and $\Delta\lambda_1$. The second moment of W_r is

$$E(W_r) = E\left[\prod_{t=1}^{T}\left(1+\frac{\Delta\lambda}{\lambda}\right)^{2\beta}\exp(-2\Delta\lambda Y_t)\right]$$

$$= \int\prod_{t=1}^{T}\left(1+\frac{\Delta\lambda}{\lambda}\right)^{2\beta}\exp(-2\Delta\lambda y_t)\lambda^{\beta}\exp(-\lambda y_t)\, y_t^{\beta-1}/\Gamma(\beta)\,dy_t$$

$$= \left(1+\frac{\Delta\lambda^2}{\lambda^2+\lambda\Delta\lambda}\right)^{T\beta}. \tag{68}$$

Q.E.D.

Note also that formula (68) matches with formula (63).

Taking into account that $E(W_r) = 1$, we have

$$\text{Var } W_r = E(W_T^2) - (EW_r)^2 = \left(1+\frac{\Delta\lambda^2}{\lambda^2+2\lambda\,\Delta\lambda}\right)^{T\beta} - 1.$$

Assuming further, without loss of generality, that $\lambda = 1$, we obtain

$$\text{Var } W_r = \left(1+\frac{\Delta\lambda^2}{1+2\Delta\lambda}\right)^{T\beta} - 1. \tag{69}$$

Consider for simplicity the case where $L_r = 1$, i.e. where $L_t W_t = W_t$. Assume that

$$\Delta\lambda^2 T \leqslant O(1), \tag{70}$$

from which it follows the Var $W_r \leqslant O(1)$.

Table 1 presents $E(W_r)^2$ as a function of T for different $\Delta\lambda$ and $\beta = 2$.

We shall restrict ourselves for concreteness to the case where Var $W_r \leqslant 25$. It follows from formula (70) and Table 1 that our method works satisfactorily if the perturbations in $\Delta\lambda$ are:

(i) *Small* ($0 < \Delta\lambda \leqslant 0.01$) and T is *rather large* ($T \approx 10^4$)
example: $\Delta\lambda = 0.001$, $T = 10^4$, Var $W_r = 6.10$.

(ii) *Rather small* ($0.01 \leqslant \Delta\lambda \leqslant 0.04$ and T is *moderate* ($T \approx 10^3$)
example: $\Delta\lambda = 0.04$, $T = 10^3$, Var $W_r = 18.31$.

(iii) *Moderate*: ($0.04 < \Delta\lambda \leqslant 0.1$) and T is *small* ($T \approx 10^2$)
example: $\Delta\lambda = 0.1$, $T = 10^2$, Var $W_r = 4.26$.

(iv) *Rather large* ($0.1 < \Delta\lambda \leqslant 0.5$) and T is *small* ($T \approx 10$)
example: $\Delta\lambda = 0.5$, $T = 10$, Var $W_r = 9.55$.

In all other cases this method *does not work* because the resulting estimate of $l(\lambda + \Delta\lambda)$ has high variance. Take, for example, $\Delta\lambda = 0.03$ and $T = 10^4$. We have (see Table 1)

$$\text{Var } W_r = 5.46^{10} - 1.$$

Case (b). We shall consider only the case where k out of m parameters of the vector $\lambda = (\lambda_1, \ldots, \lambda_m)$ are perturbed. Arguing as for formulas (67) and (69) we readily obtain

$$W_{rk} = \prod_{t=1}^{T}\prod_{i=1}^{k}\left[\left(1+\frac{\Delta\lambda_i}{\lambda}\right)^{T\beta}\exp(-\Delta\lambda_i Y_{ti})\right] \tag{71}$$

$$\text{Var } W_{rk} = \prod_{i=1}^{k}\left(1+\frac{\Delta\lambda_i^2}{\lambda^2+2\lambda\,\Delta\lambda_i}\right)^{T\beta} - 1. \tag{72}$$

For the particular case where $\lambda_i = \lambda$, and $\Delta\lambda_i = \Delta\lambda$, $i = 1, \ldots, k$, we obtain

$$\text{Var } W_{rk} = \left(1+\frac{\Delta\lambda^2}{\lambda^2+2\lambda\,\Delta\lambda}\right)^{kT\beta} - 1. \tag{73}$$

Table 1. $E[W_r^2]$ as a function of T for different $\Delta\lambda$ and $\beta = 2$

$\Delta\lambda$ \ T	1	10^1	10^2	10^3	10^4	10^5
0.01	1.0002	1.002	1.02	1.21	7.1	7.10^{10}
0.02	1.0008	1.008	1.08	2.16	2.16^{10}	2.16^{100}
0.03	1.00168	1.017	1.183	5.46	5.46^{10}	5.46^{100}
0.04	1.003	1.03	1.34	19.31	19.31^{10}	19.31^{100}
0.05	1.0046	1.046	1.57	93.72	93.72^{10}	93.72^{100}
0.1	1.0166	1.181	5.26	5.26^{10}	5.26^{100}	5.26^{1000}
0.2	1.058	1.756	1.75^{10}	1.75^{100}	1.75^{1000}	$1.75^{10,000}$
0.5	1.2656	10.55	10.5^{10}	10.5^{100}	10.5^{1000}	$10.5^{10,000}$
1	1.777	1.77^{10}	1.77^{100}	1.77^{1000}	$1.7^{10,000}$	$1.77^{100,000}$

It follows from formulas (69) and (72) that $E(W_{rk}^2) = (E(W_{r1})^2)^k$, where $W_{r1} = W_r$ [see formula (67)]. Clearly, one can find Var W_{rk} by replacing the column label r by kr with all other data remaining the same.

Example 8

Consider again the M-station queuing network of Example 1c, with $\mathbf{Y}_i \sim G(\lambda_i, \beta_i)$ and $l(\boldsymbol{\lambda})$ being the utilization at station j, $j = 1, \ldots, m$. Assume that k of the m service rate parameters, say $\lambda_1, \ldots, \lambda_k$, are perturbed: each by $\Delta\lambda_i$, $i = 1, \ldots, k$. Substituting W_{rk} [see formula (71)] in formula (56) we can estimate $l(\boldsymbol{\lambda} + \Delta\boldsymbol{\lambda}^s)$ for *many* values of $\boldsymbol{\lambda} + \Delta\boldsymbol{\lambda}^s$, $s = 1, \ldots, r$, from a *single simulation* of the nominal queuing network.

As a numerical example, let $T = 10^2$, $k = 10$, $\beta = 2$, $\lambda_i = \lambda = 1$, $i = 1, \ldots m$, and $\Delta\lambda_i = \Delta\lambda = 0.04$. In this case we have from Table 1 [see also formula (73)] that

$$\text{Var } W_{rk} = \left(1 + \frac{(0.04)^2}{1 + 2 \cdot 0.04}\right)^{2 \cdot 10^3} - 1 = 19.31 - 1 = 18.31. \tag{74}$$

It follows from formulas (61) and (62) that

$$\text{Var } \tilde{l}_{f(\mathbf{v})}(\mathbf{v}^s) = O(\text{Var } W_{rk}) = C \text{ Var } W_{rk}. \tag{75}$$

If $C \approx 1$ then substituting equation (74) into equation (75), we conclude that the variance of the estimate $\tilde{l}_{f(\mathbf{v})}$ is ≈ 18 times greater than that of its CMC counterpart $\bar{l}_{f(\mathbf{v})}$.

Note that if T were equal to 10^3 instead of 10^2 or k equal to 10^2 instead of 10, this method would not work, since

$$\text{Var } W_T = \left(1 + \frac{(0.04)^2}{1 + 2 \cdot 0.04}\right)^{2 \cdot 10^4} - 1 = (19.31)^{10} - 1.$$

Example 9

Consider the reliability model (17) with $f(\mathbf{v}, \mathbf{y})$ given as in formulas (65). Substituting equation (71) in formula (59) and taking into account that for DESS, $T = 1$, we obtain

$$\tilde{l}_{f(\boldsymbol{\lambda})}(\boldsymbol{\lambda}^s) = \frac{1}{N} \sum_{m=1}^{N} \left[\left(\max_{j=1,\ldots,p} \min_{i \in L_j} Y_{im} \right) \prod_{r=1}^{k} \left(1 + \frac{\Delta\lambda^s}{\lambda_r}\right)^{\beta} \exp(-\Delta\lambda_r^s Y_{rm}) \right]. \tag{76}$$

Let $\lambda_r = \lambda = 1$, $\Delta\lambda_r = \Delta\lambda = 0.04$, $\beta = 2$ and $k = 10^3$. In this case our method works since (see Table 1)

$$\text{Var } W_{1k} = E(W_{1k})^2 - 1 = 19.31 - 1 = 18.31.$$

Note that to obtain $\tilde{l}_{f(\boldsymbol{\lambda})}$ for the stochastic network (18) we have to replace "max min" in equation (76) by "min Σ", with all other data remaining the same.

As we pointed out, to improve the accuracy of the estimators $\tilde{l}_{f(\mathbf{v})}$ [see formulas (56)–(58) we can use variance-reduction techniques (see Section 7 and Refs [1, 4–6]).

Until now we have assumed that the vector $\mathbf{v}$, and therefore the nominal system, is chosen in advance. In most cases, however, the choice of the vector $\mathbf{v}$ is at our disposal. In this case it is natural to choose $\mathbf{v}$ as

$$\bar{\mathbf{v}} = r^{-1} \sum_{s=1}^{r} \mathbf{v}^s. \tag{77}$$

We shall now show (see Example 10 below) that in choosing v as per the above equation one can increase the efficiency of our approach. To see this, assume for simplicity that only one out of $2m$ parameters $\mathbf{v}_i = \lambda_i, \beta_i, i = 1, \ldots, m$, in formulas (65) is perturbed.

Example 10

Consider Table 1. Let $T = 10^3, \boldsymbol{\lambda} = 1, \mathbf{S} = 3, \Delta\boldsymbol{\lambda}^1 = 0.04, \Delta\boldsymbol{\lambda}^2 = 0.08$ and $\Delta\boldsymbol{\lambda}^3 = 0.12$. It follows that in this case the SF approach works *only* for the case $\Delta\boldsymbol{\lambda}^1 = 0.04$, since (see Table 1) $\text{Var } W_t(\Delta\lambda_1) = \text{Var } W_{10^3}(0.04) = 18.31$, while $\text{Var } W_{10^3}(0.08) \geqslant R$ and $\text{Var } W_{10^3}(0.12) \geqslant R$, where R is a large number. If, however, we choose $\bar{\boldsymbol{\lambda}} = 1.08$ [see equation (77)] instead of $\boldsymbol{\lambda} = 1$ then it readily follows from formulas (69) and (70) that our approach works for *all* three cases.

Consider now the finite-difference LR estimate $\nabla l(\mathbf{v})_{\text{N3}}$. The variance of $\nabla l(\mathbf{v})_{\text{N3}}$ [see formulas (49) and (50)] is

$$\text{Var } \nabla l(\mathbf{v})_{\text{N3}} = \frac{1}{\Delta\mathbf{v}^2} \text{Var}\left[T^{-1} \sum_{t=1}^{T} (L_t W_t - L_t) \right] = \frac{1}{\Delta\mathbf{v}^2} \text{Var}(\bar{l}_{f(\mathbf{v})} - \bar{l}_{f(\mathbf{v})}), \tag{78}$$

provided $N = 1$.

Since for large T, in general, $\text{Var}(L_t W_t) \gg \text{Var}(L_t)$ we readily obtain from equation (78) that

$$\text{Var } \nabla l(\mathbf{v})_{\text{N3}} = \frac{1}{(\Delta\mathbf{v})^2} O(\text{Var } \bar{l}_{f(\mathbf{v})}). \tag{79}$$

Taking into account equation (62), we have

$$\text{Var } \nabla l(\mathbf{v})_{\text{N3}} = \frac{1}{(\Delta\mathbf{v})^2} O(\text{Var } W_t) = O(\text{Var } V_t), \tag{80}$$

where $V_t = 1/\Delta\mathbf{v}\, W_t$. Finally, it follows from formulas (39) and (53) that for small $\Delta\mathbf{v}$ and $N = 1$

$$\text{Var } \nabla l(\mathbf{v})_{\text{N3}} \approx \text{Var } \bar{\nabla} l_{1,T}(\mathbf{v}) = O(T). \tag{81}$$

3.2. Taylor Series Approach

Assume that the sensitivities $\bar{\nabla}^r l_{N,T}(\mathbf{v}), r = 1, 2, \ldots,$ [see formulas (32) and (33)] are available. We consider separately both DESS and DEDS.

(*i*) *DESS.* For simplicity, let $\mathbf{v} \in \mathbf{V} \subset \mathbf{R}^1$ and consider the following Taylor series expansion:

$$l(\mathbf{v} + \Delta\mathbf{v}) = l(\mathbf{v}) + \Delta\mathbf{v}\nabla l(\mathbf{v}) + \frac{(\Delta\mathbf{v})^2}{2} \nabla^2 l(\mathbf{v}) + O(\Delta\mathbf{v})^2. \tag{82}$$

Replacing $\nabla^r l(\mathbf{v}), r = 0.1, \ldots,$ where $\nabla^0 l(\mathbf{v}) \equiv l(\mathbf{v})$, by their corresponding point estimates $\bar{\nabla}^r_N l(\mathbf{v})$ [see formulas (11) and (12)] we can extrapolate $l(\mathbf{v} + \Delta\mathbf{v})$ as follows:

$$\bar{l}^{\text{e}}_N(\mathbf{v} + \Delta\mathbf{v}) = \bar{l}_N(\mathbf{v}) + \Delta\mathbf{v}\bar{\nabla} l_N(\mathbf{v}) + \frac{(\Delta\mathbf{v})^2}{2} \bar{\nabla}^2 l_N(\mathbf{v}) + O_{\text{p}}(\Delta\mathbf{v})^2. \tag{83}$$

Here e in $\bar{l}^{\text{e}}_N(\mathbf{v} + \Delta\mathbf{v})$ denotes the extrapolated value of $l(\mathbf{v} + \Delta\mathbf{v})$, and p in $O_{\text{p}}(\Delta\mathbf{v}^2)$ means that $O_{\text{p}}(\Delta\mathbf{v}^2) = O(\Delta\mathbf{v}^2)$ in a probabilistic sense. It can be readily shown [1] that if

(a) $$|\Delta\mathbf{v}| < 1$$

and

(b) $$\text{Var } \bar{\nabla}^r l_N(\mathbf{v}) = \frac{C_r}{N} < \frac{C}{N}, \quad r = 1, 2, \ldots, \tag{84}$$

where $C_r = \text{Var } \bar{\nabla}^r l_1(\mathbf{v}) < \infty$, then

$$\text{Var}[\bar{l}^{\text{e}}_N(\mathbf{v} + \Delta\mathbf{v})] \leqslant \frac{C}{N} \frac{1}{(1 - \Delta\mathbf{v})^2} = O\left(\frac{1}{N}\right) \tag{85}$$

and therefore

$$\lim_{N \to \infty} \bar{l}^{\text{e}}_N(\mathbf{v} + \Delta\mathbf{v}) = l(\mathbf{v} + \Delta\mathbf{v}) \text{ in the mean square.} \tag{86}$$

Thus, if the sensitivities $\bar{\nabla}^r l_N(\mathbf{v})$, $r = 1, 2, \ldots$, are available, $|\Delta\mathbf{v}| < 1$, and equation (80) holds, then using the Taylor expansion (83) we can extrapolate *simultaneously* all the performance measures $l(\mathbf{v} + \Delta\mathbf{v}^s)$ from a *single* simulation experiment.

Example 11: reliability model

Consider the reliability model (17). Assume that the components $\mathbf{Y}_i$ have gamma lifetimes; i.e. $\mathbf{Y}_i \sim G(\lambda_i, \beta_i)$. Then the extrapolated value of the mean lifetime of the system can be estimated as [see formulas (21) and (22)]:

$$\bar{l}^e_N(\boldsymbol{\lambda} + \Delta\boldsymbol{\lambda}) = \frac{1}{N}\sum_{s=1}^{N}\left(\max_{j=1,\ldots,p}\min_{i\in L_j} Y_{is}\right)[1 + \Delta\boldsymbol{\lambda}'\nabla \ln f(\boldsymbol{\lambda}, \mathbf{Y}) + \tfrac{1}{2}\Delta\boldsymbol{\lambda}' H(\boldsymbol{\lambda}, \mathbf{Y})\,\Delta\boldsymbol{\lambda}] + O_p(\Delta\boldsymbol{\lambda})^2,$$

where $\nabla \ln f(\boldsymbol{\lambda}, \mathbf{Y})$ and $H(\boldsymbol{\lambda}, \mathbf{Y})$ are given in formulas (13) and (14), respectively.

(*ii*) *DEDS.* The treatment of DEDS is similar to that of DESS. In analogy to formula (83) we obtain ($\mathbf{v} \in \mathbf{V} \subset \mathbf{R}^1$)

$$\bar{l}^e_{N,T}(\mathbf{v}) + \Delta\mathbf{v}) = \bar{l}_{N,T}(\mathbf{v}) + \Delta\mathbf{v}\bar{\nabla} l_{N,T}(\mathbf{v}) + \frac{(\Delta\mathbf{v})^2}{2}\bar{\nabla}^2 l_{N,T}(\mathbf{v}) + O_p(\Delta\mathbf{v})^2. \tag{87}$$

Since $\operatorname{Var}\bar{\nabla}^r l_T(\mathbf{v}) = O(T)$, $r = 1, 2, \ldots$, [see equations (39) and (40)], we obtain

$$\operatorname{Var}[\bar{l}^e_{N,T}(\mathbf{v} + \Delta\mathbf{v})] = O\left(\frac{T}{N}\right), \tag{88}$$

provided $|\Delta\mathbf{v}| < 1$, and condition (84) holds.

Thus, for DEDS the Taylor method might work if the batch size T is small relative to N, $|\Delta\mathbf{v}| < 1$ and condition (84) holds.

It would be interesting to compare the efficiencies of the Radon–Nikodym and Taylor series approaches: for example, to find conditions under which for given T, $\Delta\mathbf{v}$, ($|\Delta\mathbf{v}| < 1$), and a given class of sample functions $L_t(\underline{\mathbf{Y}}_t)$,

$$E[\bar{l}^e_{N,T}(\mathbf{v} + \Delta\mathbf{v}) - l(\mathbf{v} + \Delta\mathbf{v})]^2 \lesssim E[\tilde{l}_{f(v)}(\mathbf{v} + \Delta\mathbf{v}) - l(\mathbf{v} + \Delta\mathbf{v})]^2. \tag{89}$$

4. SIMULATION OF QUEUING NETWORKS IN HEAVY TRAFFIC

It is well-known [e.g. 24] that estimating performance measures of queuing networks in heavy and even moderate traffic is a rather difficult and time-consuming task. In this section we show that using the Radon–Nikodym theorem we can estimate the performance measure of the *original* heavy traffic queuing model by simulating an *associated* model working in *lighter* traffic. The approach used here is the same as for performance extrapolation and is based on the LR estimate (55). More specifically, let us write $\underline{l}$ [see equation (25)] as

$$\underline{l} = E_f[\underline{L}_t(\underline{\mathbf{Y}}_t)] = \int \underline{L}_t(\underline{\mathbf{y}}_t)\frac{f_t(\underline{\mathbf{y}}_t)}{g_t(\underline{\mathbf{y}}_t)} g_t(\underline{\mathbf{y}}_t)\,d\underline{\mathbf{y}}_t = E_g\left[\underline{L}_t(\underline{\mathbf{Z}}_t)\frac{f_t(\underline{\mathbf{Z}}_t)}{g_t(\underline{\mathbf{Z}}_t)}\right], \tag{90}$$

where $\underline{L}_t(\underline{\mathbf{Y}}_t)$ and $\underline{L}_t(\underline{\mathbf{Z}}_t)$ are the truncated sample performances of the original and the *associated* queues, respectively. Note that the expectation in the second term of equation (90) is taken with respect to the pdf f while the expectation in the last term of equation (90) is taken with respect to the pdf g. Note also that changing the probability measure from $f(\mathbf{y})$ to $g(\mathbf{y})$ we *transform the original sample path to an alternative one* which can be generated by using an associated queuing model working in lighter traffic.

In analogy to equations (24) and (56) we have the following alternative estimates:

(i) the CMC estimator,

$$\bar{l}_f = \frac{1}{N}\frac{1}{T}\sum_{i=1}^{N}\sum_{t=1}^{T} L_{ti}(\underline{\mathbf{Y}}_{ti}); \tag{91}$$

and

(ii) the LR estimator,

$$\tilde{l}_g = \frac{1}{N}\frac{1}{T}\sum_{i=1}^{N}\sum_{t=1}^{T} L_{ti}(\underset{\sim}{\mathbf{Z}}_{ti})\frac{\underset{\sim}{f}_{ti}(\underset{\sim}{\mathbf{Z}}_{ti})}{g_{ti}(\underset{\sim}{\mathbf{Z}}_{ti})}, \tag{92}$$

where

$$\underset{\sim}{f}_{ti}(\underset{\sim}{\mathbf{Z}}_{ti}) = \prod_{j=t-T+1}^{t} f(\mathbf{Z}_{ji})$$

and similarly g_{ti}.

It follows from estimate (92) that to evaluate the performance measure l in heavy traffic, we can simulate an associated queuing system working in lighter traffic and then perform simple calculations with the sample performance $L_t(\underset{\sim}{\mathbf{Z}}_t)$ and the LR $\underset{\sim}{f}_t(\mathbf{Z}_t)/g_t(\mathbf{Z}_t)$. Note that generally, the extra computation of $\underset{\sim}{f}_t(\underset{\sim}{\mathbf{Z}}_t)/g_t(\underset{\sim}{\mathbf{Z}}_t)$ is small relative to the computation of $L_t(\underset{\sim}{\mathbf{Z}}_t)$.

Let us now apply our approach to the GI/G/1 queue.

Example 12: GI/G/1 queue

In this case the CMC and the LR estimates are

$$\bar{l}_f = \frac{1}{T}\sum_{t=1}^{T} X_t$$

and

$$\tilde{l}_g = \frac{1}{\tilde{T}}\sum_{t=1}^{\tilde{T}} \tilde{X}_t \frac{\underset{\sim}{f}_t(\underset{\sim}{\mathbf{Z}}_t)}{g_t(\underset{\sim}{\mathbf{Z}}_t)}, \tag{93}$$

respectively. Here X_t is given in Example 5 and

$$\tilde{X}_{t+1} = \max\{0, \tilde{X}_t - \mathbf{Z}_{1t} + \mathbf{Z}_{2t}\}, \tag{94}$$

where $\mathbf{Z}_{1t}$ and $\mathbf{Z}_{2t}$ are the interarrival and the service times of the tth customer in the associated GI/G/1 having a joint pdf $g(\mathbf{z})$, $\mathbf{z} = (z_1, z_2)$, and $\tilde{T}$ is the batch size for the associated queue.

Note that the associated GI/G/1 queue works in *lighter* traffic than the original one if

$$\rho = \frac{E(\mathbf{Y}_2)}{E(\mathbf{Y}_1)} > \tilde{\rho} = \frac{E(\mathbf{Z}_2)}{E(\mathbf{Z}_1)}. \tag{95}$$

Clearly, if formula (95) holds then [e.g. 24] the estimate $\tilde{l}_g$ possesses the following properties:

(A) It has a shorter transient period than its counterpart $\bar{l}_f$. This means that using $\tilde{l}_g$ one can start collecting the steady-state data earlier than with $\bar{l}_f$.

(B) It uses a smaller batch size than $\bar{l}_f$, i.e. $\tilde{T} < T$.

The main drawback of the estimate $\tilde{l}_g$ (as with any of the LR estimates considered earlier) is that

(C) Typically $\operatorname{Var} \tilde{l}_g > \operatorname{Var} \bar{l}_f$.

For more details on this issue see Ref. [15] and Section 7 below.

5. PERFORMANCE EVALUATION AND THE ACCEPTANCE–REJECTION METHOD

In this section we shall show that the LR estimate $\tilde{l}_g$ [see equation (92)] can be used for performance evaluation without resorting to the acceptance–rejection method. We will begin by explaining the acceptance–rejection method and why it is sometimes cumbersome.

Let $F(\mathbf{y})$, $\mathbf{y} \in \mathbf{R}^m$, be a multidimensional cdf with dependent components. It is known [e.g. 33] that for general cdfs it is difficult (computationally) to apply the well-known inverse

transform method; i.e. to generate a vector $\mathbf{Y}$ while solving the following system of nonlinear equations,

$$\left.\begin{aligned} F_1(\mathbf{Y}_1) &= U_1 \\ F_2(\mathbf{Y}_2|\mathbf{Y}_1) &= U_2 \\ &\vdots \\ F_m(\mathbf{Y}_m|\mathbf{Y}_1,\ldots,\mathbf{Y}_{m-1}) &= U_m, \end{aligned}\right\} \tag{96}$$

with respect to $\mathbf{U} = (U_1, \ldots, U_m)$. Here $U_1, \ldots, U_m$ are iid random variables (random numbers) each distributed $\mathbf{U}(0, 1)$. In this case the acceptance–rejection method is often used.

According to the acceptance–rejection method [33] one presents the pdf $f(\mathbf{y})$ as

$$f(\mathbf{y}) = ch(\mathbf{y})g(\mathbf{y}), \tag{97}$$

where $c \geqslant 1$, $0 < h(\mathbf{y}) \leqslant 1$ and $g(\mathbf{y})$ is a pdf from which random vectors $\mathbf{Z} = (\mathbf{Z}_1, \ldots, \mathbf{Z}_m)$ can be readily generated [say,

$$g(\mathbf{z}) = \prod_{i=1}^{m} g_i(\mathbf{z}_i)$$

and each RV $\mathbf{Z}_i \sim g_i(\mathbf{z}_i)$ can be readily generated by the inverse transform method]. Then one executes the following:

Acceptance–rejection algorithm

1. Generate $\mathbf{Z}$ from $g(\mathbf{z})$.
2. Generate $\mathbf{U}$ from $\mathbf{U}(0, 1)$.
3. If $U_i \leqslant f(\mathbf{z}_i)/ch(\mathbf{z}_i)$, $\forall i = 1, \ldots, m$, accept $\mathbf{Z}$ as a random vector generated from $f(\mathbf{y})$.
4. Go to Step 1.

The main drawback of the acceptance–rejection method is that the number of trials needed to generate a point from $f(\mathbf{y})$ increases explosively with the dimensionality of $\mathbf{Y} \in \mathbf{R}^m$. For more details, see Ref. [33, p. 51].

Let us now compare briefly the efficiencies of both the CMC estimate $\bar{l}_f$ [see equation (91)] and the LR estimate $\tilde{l}_g$ [see equation (92)] for performance evaluation.

If m is large and $\mathbf{Y} = (\mathbf{Y}_1, \ldots, \mathbf{Y}_m)$ *must* be generated by using the acceptance–rejection method then clearly $\bar{l}_f$ is practically unrealizable. Deriving the alternative estimate $\tilde{l}_g$, however, might be much easier and less time-consuming, especially if we use the inverse transform method for generation from $g(\mathbf{z})$. Note that while using $\tilde{l}_g$ instead of $\bar{l}_f$ one has to take into account (see Section 4.2) that usually

$$\operatorname{Var} \tilde{l}_g > \operatorname{Var} \bar{l}_f,$$

and $\operatorname{Var} \tilde{l}_g$ increases with m ($\operatorname{Var} W_{Tm_1} > \operatorname{Var} W_{Tm_2}$ [see formulas (71)–(73)] if $m_1 > m_2$). Note also that g can be chosen not only with the view of avoiding generation from f but with the view of variance reduction (application of importance sampling) as well.

Example 15

Consider the reliability model (17). The CMC and the LR estimates can be written as

$$\bar{l}_f = \frac{1}{N} \sum_{s=1}^{N} \max_{j=1,\ldots,p} \min_{i \in L_j} \mathbf{Y}_{is} \tag{98}$$

and

$$\tilde{l}_g = \frac{1}{N} \sum_{s=1}^{N} \left(\max_{j=1,\ldots,p} \min_{i \in L_j} \mathbf{Z}_{is} \right) \frac{f(\mathbf{Z}_s)}{g(\mathbf{Z}_s)}, \tag{99}$$

respectively.

The expressions for $\bar{l}_f$ and $\tilde{l}_g$ for the stochastic PERT network (18) are similar: we have to replace the operator "max min" by the "min Σ" operator, respectively.

It is important to note that in both cases here and for heavy traffic performance evaluation, the estimate $\tilde{l}_g$ is based on probability measure transformation from f to g. Note, however, that in the first case g is introduced with the view of simulating an associated queue working in light traffic, while in the second case it is introduced with the view of avoiding generation from $f(\mathbf{y})$ by the acceptance–rejection method.

It is clear that there might be cases wherein $\tilde{l}_g$ can achieve both goals simultaneously to evaluate heavy traffic performance while simulating a light (lighter) traffic queue, and to avoid generating RVs from $f(\mathbf{y})$ by using the acceptance–rejection method.

Finally, it is important to note that the CMC estimate $\bar{l}_f$ is based on simulation of the *original* system while the LR estimate $\tilde{l}_g$ is based on simulation of an *artificial* (*auxiliary*) system.

6. EXTENSION OF THE MODEL

In many applications not only the density f but the sample performance L depends on $\mathbf{v}$. In this case,†

$$l(\mathbf{v}) = E_{f(\mathbf{v})}[L(\mathbf{v}, \mathbf{Y})] = \int L(\mathbf{v}, \mathbf{y}) f(\mathbf{v}, \mathbf{y})\, d\mathbf{y}. \tag{100}$$

Clearly, model (1) is a particular case of model (100) with $L(\mathbf{v}, \mathbf{Y}) = L(\mathbf{Y})$.

We shall show that in this case one can again

(i) estimate the sensitivities,
(ii) extrapolate the performance,
(iii) evaluate the heavy traffic performance

and

(iv) avoid generation from $f(\mathbf{v}, \mathbf{y})$

by using a single simulation and probability measure transformation. Consider cases (i)–(iv) separately.

(i) As before, let ϕ be a differentiation operator. Consider first DESS. In analogy to equation (7) we have

$$\phi(l(\mathbf{v})) = \phi \int L(\mathbf{v}, \mathbf{y}) f(\mathbf{v}, \mathbf{y})\, d\mathbf{y} = E_{f(\mathbf{v})}\left[\frac{\phi(L(\mathbf{v}, \mathbf{Y}) f(\mathbf{v}, \mathbf{Y}))}{f(\mathbf{v}, \mathbf{Y})}\right], \tag{101}$$

provided that both L and f are differentiable with respect to $\mathbf{v}$ and the operators ϕ and integration are interchangeable.

For the particular case where $\phi = \nabla$, we readily obtain

$$\nabla l(\mathbf{v}) = E_{f(\mathbf{v})}[L(\mathbf{v}, \mathbf{Y})(\nabla \ln f(\mathbf{v}, \mathbf{Y}) + \nabla \ln L(\mathbf{v}, \mathbf{Y}))] = E_{f(\mathbf{v})}[\nabla L(\mathbf{v}, \mathbf{Y}) + L(\mathbf{v}, \mathbf{Y}) \nabla \ln f(\mathbf{v}, \mathbf{Y})]. \tag{102}$$

Unbiased estimates of $l(\mathbf{v})$ and $\nabla l(\mathbf{v})$ are

$$\bar{l}_N(\mathbf{v}) = \frac{1}{N} \sum_{i=1}^{N} L(\mathbf{v}, \mathbf{Y}_i) \tag{103}$$

and

$$\bar{\nabla} l_N(\mathbf{v}) = \frac{1}{N} \sum_{i=1}^{N} [L(\mathbf{v}, \mathbf{Y}_i) \nabla \ln f(\mathbf{v}, \mathbf{Y}_i) + \nabla L(\mathbf{v}, \mathbf{Y}_i)], \tag{104}$$

respectively.

†For examples of such models, see Refs [3, Chap. 3; 18–20].

The extension to DEDS is similar. In the analogy to equations (31) and (32) we have

$$l_{N,T}(\mathbf{v}) = \frac{1}{N}\frac{1}{T}\sum_{i=1}^{N}\sum_{t=1}^{T} L_{ti}(\mathbf{v}, \underline{\mathbf{Y}}_{ti}) \tag{105}$$

and

$$\nabla l_{N,T}(M, \mathbf{v}) = N^{-1}T^{-1}\sum_{i=1}^{N}\sum_{t=1}^{T}[L_{ti}(\mathbf{v}, \underline{\mathbf{Y}}_{ti})\mathbf{S}_{ti}(M, \mathbf{v}, \underline{\mathbf{Y}}_{ti}) + \nabla L_{ti}(\mathbf{v}, \underline{\mathbf{Y}}_{ti})], \tag{106}$$

respectively. Higher order sensitivity estimates $\bar{\nabla}^r l_{N,T}(M, \mathbf{v})$, $r = 2, 3, \ldots$, can be readily obtained from equation (105) as well.

Thus, again using the SF approach, we can estimate *simultaneously* from a *single* simulation the performance $l(\mathbf{v})$ and *all* its sensitivities.

Let us turn now to performance extrapolation and avoiding simulation from $f(\mathbf{v}, \mathbf{y})$.

(ii) In analogy to equation (56) we have for $N = 1$,

$$\tilde{l}_{f(\mathbf{v})}(\mathbf{v}^s) = T^{-1}\sum_{t=1}^{T} L_t(\mathbf{v}^s, \underline{\mathbf{Y}}_t)W_t(\mathbf{v}, \underline{\mathbf{Y}}_t).$$

(iii) and (iv) In analogy to equation (92) the heavy traffic estimate and the estimate which avoids generation from f can be written for $N = 1$ as

$$\tilde{l}_g(\mathbf{v}') = \frac{1}{T}\sum_{t=1}^{T} L_t(\underline{\mathbf{Z}}_t)\frac{f_t(\underline{\mathbf{Z}}_t)}{g_t(\underline{\mathbf{Z}}_t)},$$

where

$$g_t(\underline{\mathbf{Z}}_t) = \prod_{j=t-T+1}^{t} g(\mathbf{Z}_j),\ \mathbf{Z}_j \sim g(\mathbf{z}), \qquad \text{and} \qquad g(\mathbf{z})$$

is chosen either with a view to generating a sample path $L_t(\mathbf{v}, \underline{\mathbf{Z}}_t)$ from an associated queue working under light traffic, or with a view to generating a random sample from $g(\mathbf{z})$ by the inverse transform method, say, to avoid generating from $f(\mathbf{y})$ by the acceptance–rejection method, say.

Thus we have shown that for model (100) all four of the above-mentioned issues (i)–(iv) can be treated *simultaneously* while using a single simulation and some probability measure transformations.

7. VARIANCE REDUCTION AND NUMERICAL RESULTS

We shall consider here briefly application of linear and nonlinear control random variables to improve the accuracy of the "what if" estimator $\tilde{l}_{f(\mathbf{v})}$. Their application to the sensitivity estimator $\bar{\nabla} l_{N,T}(\mathbf{v})$ can be found in Refs [1, 4, 6, 15].

(a) Linear control random variable procedure

Taking into acount that $E(W_t) = 1$, the linear control random variable procedure for

$$\tilde{l}_{f(\mathbf{v})} = \frac{1}{T}\sum_{t=1}^{T} L_t W_t$$

can be written [e.g. 35] as

$$\tilde{l}(\boldsymbol{\beta}) = \frac{1}{T}\sum_{t=1}^{T} L_t W_t + \boldsymbol{\beta}\left(\frac{1}{T}\sum_{t=1}^{T} W_t - 1\right) = X - \boldsymbol{\beta}(C - 1),$$

where

$$X = \frac{1}{T}\sum_{t=1}^{T} L_t W_t \quad \text{and} \quad C = \frac{1}{T}\sum_{t=1}^{T} W_t.$$

The value $\boldsymbol{\beta}^*$, which minimizes $\mathrm{Var}[\tilde{l}(\boldsymbol{\beta})]$, is

$$\boldsymbol{\beta}^* = \frac{\mathrm{Covar}(X, C)}{\mathrm{Var}(C)}.$$

In practice $\boldsymbol{\beta}^*$ is unknown and must be estimated from simulation.

(*b*) *Nonlinear control random variable procedure*

Since $E(\tilde{W}_t) = 1$, we can write $l(\mathbf{v}^s)$ [see equation (47)] as

$$l(\mathbf{v}^s) = \frac{E_{f(\mathbf{v})}(L_t W_t)}{E_{f(\mathbf{v})}(W_t)}$$

and consider

$$\tilde{l}^c = \frac{\sum_{i=1}^{N}\sum_{t=1}^{T} L_{ti} W_{ti}}{\sum_{i=1}^{T}\sum_{t=1}^{T} W_{ti}} = \frac{\sum_{i=1}^{N} X_i}{\sum_{i=1}^{N} C_i},$$

where

$$X_i = \frac{1}{T}\sum_{t=1}^{T} L_{ti} W_{ti}$$

$$C_i = \frac{1}{T}\sum_{t=1}^{T} W_{ti},$$

as an alternative to the LR estimator $\bar{l}_{f(\mathbf{v})}$ [see equation (56)].

Note that, following Glynn and Whitt [36], $\tilde{l}^c$ can be called the nonlinear control random variable (NCRV) estimator of ratio type.

It is readily shown [e.g. 5] that a $100(1-\delta)\%$ confidence interval for $l(\mathbf{v}^s) = E(X_i)/E(C_i)$ is

$$\tilde{l}^c \pm \frac{z_{1-\delta/2}^{S}}{\bar{C}N^{1/2}},$$

where $z_{1-\delta} = \Phi(1-\delta)$, Φ denotes the standard normal distribution function

$$\bar{C} = \frac{1}{N}\sum_{i=1}^{N} C_i \text{ and } S^2$$

is the sample variance of $\sigma^2 = \mathrm{Var}(X_i) - 2l(\mathbf{v}^s)\mathrm{Covar}(X^i, C_i) + [l(\mathbf{v}^s)]^2\,\mathrm{Var}(C_i)$.

Note that for DESS, $\tilde{l}^c$ reduces to

$$\tilde{l}^c = \frac{\sum_{i=1}^{N} L_i W_i}{\sum_{i=1}^{N} W_i} = \frac{\sum_{i=1}^{N} X_i}{\sum_{i=1}^{N} C_i},$$

where

$$X_i = L_i W_i, \qquad C_i = W_i, \qquad W_i = \frac{f(\mathbf{v}, \mathbf{Y}_i^s)}{f(\mathbf{v}, \mathbf{Y}_i)}.$$

Table 2 presents numerical results for the "what if" problem assuming that $l(\mathbf{v})$ $(\mathbf{v} = (\lambda, \mu))$ is the steady-state waiting time of a customer in an M/M/1 system with interarrival rate λ, service rate μ and traffic intensity $\rho = \lambda/\mu$. We simulated the M/M/1 queue for $\rho = 0.5$, starting at the origin and using the batch mean method. We assumed that $\lambda = 1, \mu = 2$, choose the number of batches $N = 100$ and we deleted the first 200 transient customers. More specifically. Table 2 presents the theoretical values of

$$l(\lambda + \alpha_1\lambda, \mu + \alpha_2\mu)$$

Table 2. Point estimators, sample variances and 95% confidence intervals for the steady-state mean waiting time in an M/M/1 queue with the LR estimator $\tilde{l}$ and the NCRV estimator $\tilde{l}^c$ for $\rho = 0.5$, $T = 50$, $M = 50$ and $N = 100$ replications

Parameters			Theoret. value	Point estimators		Sample variances			95% Confidence intervals	
$\alpha_1(\%)$	$\alpha_2(\%)$	ρ	l	$\tilde{l}$	$\tilde{l}^c$	S_w^2	$S_{\tilde{l}}^2$	$S_{\tilde{l}^c}^2$	$\tilde{l} \pm \frac{z_{1-\delta/2} S_{\tilde{l}}}{N^{1/2}}$	$\tilde{l}^c \pm \frac{z_{1-\delta/2} S_{\tilde{l}^c}}{N^{1/2}}$
0.0	0.0	0.500	1.000	1.007	1.007	0	0.134	0.134	0.936, 1.079	0.936, 1.079
+5.0	0.0	0.525	1.053	1.036	1.045	0.065	0.280	0.140	0.932, 1.139	0.971, 1.118
+10.0	0.0	0.550	1.111	1.073	1.105	0.272	0.755	0.250	0.903, 1.243	1.007, 1.203
+20.0	0.0	0.600	1.250	1.134	1.235	1.356	3.433	1.102	0.771, 1.497	1.003, 1.441
+30.0	0.0	0.650	1.429	1.173	1.350	3.909	9.851	4.116	0.558, 1.789	0.947, 1.753
+3.0	−3.0	0.531	1.099	1.099	1.091	0.080	0.413	0.207	0.973. 1.225	1.002. 1.181
+5.0	−5.0	0.553	1.176	1.183	1.157	0.256	0.947	0.308	0.992, 1.373	1.048, 1.265
+10.0	−10.0	0.611	1.429	1.505	1.356	1.669	6.407	0.972	1.009, 2.001	1.162, 1.549
+15.0	−15.0	0.676	1.818	1.921	1.570	6.018	24.813	1.816	0.862, 2.774	1.298, 1.841
+20.0	−20.0	0.750	2.500	2.654	1.881	17.101	100.38	4.739	0.690, 4.617	1.457, 2.307

where

$$\alpha_1 = \frac{\Delta\lambda}{\lambda} \quad \text{and} \quad \alpha_2 = \frac{\Delta\mu}{\mu},$$

point estimators, sample variances and confidence intervals while using the LR estimator $\tilde{l}$ and the NCRV estimator $\tilde{l}^c$ for different α_1 and α_2. We choose $T = M = 50$ and use the same stream of random numbers for both $\tilde{l}$ and $\tilde{l}^c$.

It follows from the results of Table 2 that the NCRV estimator $\tilde{l}^c$ is more accurate than the LR estimator $\tilde{l}$. It also follows from the results of Table 2 that using the NCRV estimator $\tilde{l}^c$ we can get meaningful results while perturbing λ by 30% with $\Delta\mu = 0$ or while perturbing both λ and μ by 15%, respectively. In other words, while simulating the M/M/1 system with traffic intensity $\rho = 0.5$ (low traffic) and perturbing either $\lambda(\Delta\mu = 0)$ or both λ and μ we can extrapolate with $\tilde{l}^c$ meaningfully the steady-state waiting time for $\rho = 0.65$ and $\rho = 0.676$ (higher traffic), respectively. Clearly, one has to take into consideration that a higher percentage of extrapolation with $\tilde{l}^c$ is associated with higher variance.

Table 3 presents simulation results for $\nabla l(\lambda) = \partial l(\mathbf{y})/\partial\lambda$ $(\mathbf{v} = (\lambda, \mu))$, where $l(\mathbf{v})$ is the expected waiting time of a customer in the M/M/1 queue. We choose for $\rho = 0.1$, $T = M = 10$, and for $\rho = 0.5$, $T = M = 30$; and $N = 100$. It follows from Table 3 that the estimator $\tilde{\nabla} l(\lambda)$ is more accurate than its counterpart $\bar{\nabla} l(\lambda)$.

8. CONCLUDING REMARKS

In this paper we have shown that using the SF approach one can:

(i) *Estimate simultaneously* from a *single* simulation run the performance measure $l(\mathbf{v})$ and *all* the sensitivities $\nabla^r l(\mathbf{v})$, $r = 1, 2, \ldots$, for both DESS and DEDS.

(ii) *Extrapolate from the same simulation run the performance* $l(\mathbf{v})$ (to solve the "what if" problem) for different values of $\mathbf{v} + \Delta\mathbf{v}^s$.

(iii) *Evaluate the performance* of a simple queuing model working in *heavy* traffic while simulating an *associated* queuing model working in *lighter* traffic. The positive features of the proposed estimate $\tilde{l}_g$ relative to the conventional CMC estimate $\bar{l}_f$ are:

(a) shorter transient period,
(b) shorter batch size;

its negative feature is its variance,

(c) usually $\operatorname{Var} \tilde{l}_g > \operatorname{Var} \bar{l}_f$ and $\operatorname{Var} \tilde{l}_g$ increases with the traffic intensity $\tilde{\rho}$ of the associated queue.

(iv) *evaluate the performance* l of stochastic models while generating a stream of RVs from an *auxiliary* pdf, say, g rather than from the original one f. The advantage of such an estimate relative to its CMC counterpart can be substantial

Table 3. Point estimators, sample variance and 95% confidence intervals for $\nabla l(\lambda) = \partial l/\partial \lambda$ [$l(\lambda, \mu)$ is waiting time in the M/M/1 queue) with the estimators $\bar{\nabla} l(\lambda)$ and $\tilde{\nabla} l(\lambda)$ for $N = 100$ and $\rho = 0.1$ ($M = 10$) and $\rho = 0.5$ ($M = 30$)

		Theoret. values		Point estimators			Sample Variance			95% Confidence intervals	
ρ	M	$l(\lambda, \mu)$	$\nabla l(\lambda)$	$\bar{l}(\lambda, \mu)$	$\bar{\nabla} l(\lambda)$	$\tilde{\nabla} l(\lambda)$	$S^2_{\bar{l}}$	$S^2_{\bar{\nabla} l}$	$S^2_{\tilde{\nabla} l}$	with $\bar{\nabla} l(\lambda)$	with $\tilde{\nabla} l(\lambda)$
0.1	5.0	0.0111	0.0123	0.0111	0.0115	0.0132	0.000	0.0002	0.0001	0.005, 0.018	0.008, 0.019
0.5	25.0	0.500	1.000	0.5171	0.9516	1.0343	0.053	4.911	3.266	0.517, 1.386	0.700, 1.368

if, say, it is assumed that f must be generated by the acceptance–rejection method, and g can be generated by the inverse transform method, respectively, and if f is a multi-dimensional density.

It is important to note that all four estimates mentioned above, namely:

(i) sensitivity $\bar{\nabla}^r l(\mathbf{v})$ [and $l(\mathbf{v})_{N3}$];
(ii) performance extrapolation $l_{f(\mathbf{v})}(\mathbf{v} + \Delta\mathbf{v}^s)$, $s = 1, \ldots, r$;
(iii); (iv) heavy traffic and one based on avoiding generation from f (both denoted $\tilde{l}_g$);

require a *single* simulation experiment either with the *original* [cases (i) and (ii)] or with the *associated* [cases (iii) and (iv)] model and some *measure transformations* (application of the Radon–Nikodym theorem).

Note that to improve the accuracy of the estimators variance-reduction techniques and especially NCRVs can be efficiently used (see Section 7 and Refs [1, 4–6]).

Implementation of the SF approach in all four cases (i)–(iv) is rather simple. For performance extrapolation and performance evaluation in heavy traffic we also proposed estimates (alternative to $\tilde{l}_f$ and $\tilde{l}_g$, respectively) based on the Taylor series expansion and discussed conditions under which they can give satisfactory results.

9. SOME IDEAS FOR FURTHER RESEARCH

We suggest the following directions for future research:

(i) Apply efficient variance-reduction techniques (control variates, antihetic variates, importance sampling etc) for further improvement of the accuracy of the sensitivity estimates $\bar{\phi}(l(\mathbf{v}))$ and the "what if" estimates $\tilde{l}_{f(\mathbf{v})}(\mathbf{v}^s)$ (and $\tilde{l}_g$).
(ii) Apply the sensitivity analysis estimates $\bar{\phi} l_T(\mathbf{v})$ and $\bar{G} l(\mathbf{v})_{N3}$ (and in particular $\bar{\nabla} l(\mathbf{v})$, $\bar{\nabla}^2 l(\mathbf{v})$ and $\nabla l(\mathbf{v})_{N3}$) to a broad variety of queuing networks.
(iii) Apply the LR estimates $\tilde{l}_{f(\mathbf{v})}$ and $\tilde{l}_{g(\mathbf{v})}$ for performance extrapolation and performance evaluation of computer simulation models.
(iv) Apply the sensitivity estimates for optimizing DESS and DEDS.
(v) Compare the efficiencies of Ho *et al.*'s [37] lower variance biased PA estimate $\nabla l(\mathbf{v})_{N5}$ with the higher variance, unbiased SF estimate $\nabla l(\mathbf{v})_{N3}$.
(vi) Compare the efficiencies (say, in the sense of the mean square error) of the LR estimate $\tilde{l}_{f(\mathbf{v})}$ (and $\tilde{l}_g$) with its counterpart Taylor series expansion estimates $\bar{l}^e_{N,T}$.
(vii) Investigate empirically and analytically how fast decreases the correlation between $L_t W_t$ and $L_{t+k} W_{t+k}$ (k is fixed) the batch size T as a function of W_t. Clearly, one has to take into consideration the following trade-off while using the LR estimator $\tilde{l}_v(\mathbf{v}^s)$ instead of the CMC estimator: the LR estimator $\tilde{l}_v$ which uses shorter batches than its CMC counterpart has in general larger variance.

Acknowledgements—We would like to express our indebtedness to Herman Chernoff from Harvard University, Andrew Rukhin from University of Massachusetts, Muni Srivastava from the University of Toronto for several valuable suggestions and to the members of the Department of Theoretical and Applied Mechanics, University of Auckland for their hospitality and secretarial assistance.

REFERENCES

1. R. Y. Rubinstein, Sensitivity analysis of computer simulation models via the score function. *Ops Res.* (in press).
2. R. Y. Rubinstein, A Monte Carlo method for estimating the gradient in a stochastic network. Unpublished manuscript, Technion, Haifa, Israel (1976).

3. R. Y. Rubinstein, *Monte Carlo Optimization, Simulation and Sensitivity of Queuing Networks*. Wiley, New York (1986).
4. R. Y. Rubinstein, The score function approach for sensitivity analysis of computer simulation models. *Maths Comput. Simuln* **28**, 1–29 (1986).
5. R. Y. Rubinstein, The "what if" problem in simulation analysis. Unpublished manuscript, George Washington Univ., Washington, D.C. (1987).
6. R. Y. Rubinstein, Modified importance sampling for performance evaluation and sensitivity analysis of computer simulation models. Unpublished manuscript. Technion, Haifa, Israel (1987).
7. E. L. Lehmann, *Theory of Point Estimation*. Wiley, New York (1983).
8. Y. C. Ho, M. A. Eyler and T. T. Chien, A gradient technique for general buffer storage design in a serial production line. *Int. J. Prod. Res.* **17**, 557–580 (1979).
9. P. Heidelberger, X. R. Cao, M. A. Zazanis and R. Suri, Convergence properties of infinitesimal perturbation analysis estimates. *Mgmt Sci.* (in press).
10. Y. C. Ho and X. R. Cao, Perturbation analysis and optimization of queuing networks. *J. optimizn Theory Applic.* **40**, 559–582 (1983).
11. X. R. Cao, Convergence of parameter sensitivity estimates in a stochastic environment. *IEEE Trans. autom. Control* **30**, 845–853 (1985).
12. X. R. Cao, On the sample functions of queuing networks with applications to perturbation analysis. *Ops Res.* (in press).
13. R. Suri, Infinitesimal perturbation analysis of discrete event dynamic systems: a general theory. In *Proc. IEEE Decision and Control Conf.*, San Antonio, Tex. (1983).
14. M. Zazanis and R. Suri, Comparison of perturbation analysis with conventional sensitivity estimates for regenerative stochastic systems. *Ops Res.* (in press).
15. M. Ahmed, H. Arsham and R. Y. Rubinstein, Computational experiments with the score function method for sensitivity analysis and performance extrapolation of computer simulation models (in preparation).
16. X. R. Cao, Sensitivity estimates based on one realization of a stochastic system. *J. statist. Comput. Simuln* **27**, 211–232 (1987).
17. J. Kreimer, Stochastic optimization—an adaptive approach. Ph.D. Dissertation, Technion, Haifa, Israel.
18. P. Glynn, Stochastic approaximation for Monte Carlo optimization. In *Proc. Winter Simulation Conf.*, Washington, D.C. (1986).
19. P. Glynn, Optimization of stochastic systems. In *Proc. Winter Simulation Conf.*, Washington, D.C. (1986).
20. P. Glynn, Sensitivity analysis for stationary probabilities of a Markov chain. In *Proc. 4th Army Conf. on Applied Mathematics and Computers* (1986).
21. M. I. Reiman and A. Weiss, Sensitivity analysis via likelihood ratios. In *Proc. Winter Simulation Conf*, Washington, D.C. (1986).
22. H. Rief, Monte Carlo uncertainty analysis. In *Handbook of Uncertainty Analysis* (Edited by Y. Ronen). CRC Press, Boca Raton, Fla (1987).
23. H. Rief, E. M. Gelbard, R. W. Shaefer and K. S., Smith, Review of Monte Carlo techniques for analyzing reactor perturbations. *Nucl. Sci. Engng* **92**, 289–297 (1985).
24. A. M. Law and W. D. Kelton, *Simulation Modeling and Analysis*. McGraw-Hill, New York (1982).
25. R. Y. Rubinstein and F. Szidarovszky, Convergence of perturbation analysis estimates for continuous sample functions: a general approach. *Maths Comput. Simuln* (in press).
26. R. Y. Rubinstein and F. Szidarovszky, Convergence of perturbation analysis estimates for discontinuous sample functions: a general approach. *Appl. Probabil.* (1988).
27. D. R. Cox and D. V. Hinkley, *Theoretical Statistics*. Chapman & Hall, London (1974).
28. D. Gross and C. M. Harris, *Fundamentals of Queuing Theory*. Wiley, New York (1985).
29. G. S. Fishman, *Principles of Discrete Event Digital Simulation*. Wiley, New York (1978).
30. S. Karlin and M. Taylor, *A First Course in Stochastic Processes*. Academic Press, New York (1975).
31. P. Glynn and D. I. Iglehar, Importance sampling for stochastic simulations. *Mgmt Sci.* (in press).
32. A Feuerverger, D. L. McLeish and R. Y. Rubinstein, A cross-spectral method for sensitivity analysis of computer simulation models. *C. r. math. Rep. Acad. Sci. R. Soc. Can.* **VIII**, No. 5 (1986).
33. R. Y. Rubinstein, *Simulation and the Monte Carlo Methods*. Wiley, New York (1981).
34. W. E. Billes and J. J. Swain, Mathematical programming and the optimization of computer simulations. *Math Program. Stud.* **11**, 189–207 (1979).
35. R. Y. Rubinstein and R. Marcus, Efficiency of multivariate control variates in Monte Carlo simulation. *Ops Res.* **33**, 661–677 (1985).
36. P. Glynn and W. Whitt, Indirect estimation via $L = \lambda W$. *Ops Res.* (in press).
37. Y. C. Ho, R. Suri, X. R. Cao, G. W. Dielhl, J. W. Dille and M. Zazanis, Optimization of large multiclass (non-product form) queuing networks using perturbation analysis. *J. Large Scale Syst.* **7**, 165–180 (1984).