TECHNICAL RESEARCH REPORT

Comparing Gradient Estimation Methods Applied to Stochastic Manufacturing Systems

by Praveen V. Mellacheruvu, Michael C. Fu,
Jeffrey W. Herrmann

T.R. 2000-1

Comparing gradient estimation methods applied to stochastic manufacturing systems

Praveen V. Mellacheruvu, Michael C. Fu and Jeffrey W. Herrmann
Institute for Systems Research
University of Maryland
College Park, MD 20742

December 30, 1999

Abstract

This paper compares two gradient estimation methods that can be used for estimating the sensitivities of output metrics with respect to the input parameters of a stochastic manufacturing system. A brief description of the methods used currently is followed by a description of the two methods: the finite difference method and the simultaneous perturbation method. While the finite difference method has been in use for a long time, simultaneous perturbation is a relatively new method which has been applied with stochastic approximation for optimization with good results. The methods described are used to analyze a stochastic manufacturing system and estimate gradients. The results are compared to the gradients calculated from analytical queueing system models. These gradient methods are of significant use in complex manufacturing systems like semiconductor manufacturing systems where we have a large number of input parameters which affect the average total cycle time. These gradient estimation methods can estimate the impact that these input parameters have and identify the parameters that have the maximum impact on system performance.

1 Introduction to gradient estimation

Gradient estimation is an important technique that can be utilized to estimate the impact of change in input parameters on output metrics in stochastic processes. If the response of the output metrics with respect to the input parameters is continuous in nature, then the gradient of the output metric is obtained as a partial derivative of the response function. Gradient estimation for applications like optimization and sensitivity analysis can be done through a number of methods [1, 3, 8].

Section 2 describes some of the methods for gradient estimation and introduces the two methods used to estimate the gradients. Section 3 describes the problem that is considered here for sensitivity analysis and an example manufacturing system. Section 4 describes the application of the finite difference method to the example. Section 5 shows how the simultaneous perturbation method has been applied to the example. Section 6 compares the results to those from an exact queueing system model. Section 7 concludes the paper.

2 Introduction to gradient estimation methods

Some of the methods for gradient estimation are finite difference method, perturbation analysis, likelihood ratio method, frequency domain methods, and simultaneous perturbation method. While some methods like the perturbation analysis method require knowledge of the system being simulated which requires obtaining output or change in the input when the simulation is in progress, other methods like the finite difference methods take a black-box type approach to the simulation system for estimating the gradient.

Perturbation Analysis [5] is further classified into many submethods like the Infinitesimal Perturbation Analysis (IPA) and Smoothed Perturbation Analysis (SPA). IPA [13] reformulates the problem of estimating gradient with respect to the input parameters as the problem of estimating the gradient of an expected value involving a random variable whose distribution does not depend on the input vector, θ. The likelihood ratio method is described in more detail in [4]. The frequency domain method [6] involves oscillating the value of the input parameter in a sinusoidal fashion during a single run which will give an output function, a superposition function of the different inputs. This output function can be used for gradient estimation. The two methods that are presented here are the finite difference (FD) method and the simultaneous perturbation (SP) method.

Let us consider a stochastic process that has a certain number of input parameters and output metrics, which help us determine the performance of the process. The output metrics are obtained either through experiments, simulation or some other process as depicted in Figure 1.

The sensitivity of the output metrics to the input processes is very helpful in determining the impact of the input parameters on the output processes. The output metric can be shown as a function of the input parameters.

$$f = f(\theta_1, \theta_2 ... \theta_n) \tag{1}$$

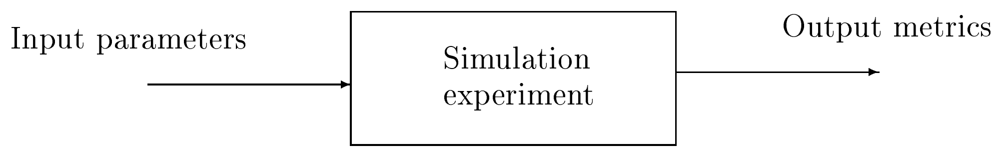

Figure 1: Simulation box

where f is the output metric written as a function of $\theta_i, i = 1, 2...n$, the input parameters.
The aim is to estimate the estimate g, the gradient, which is the partial derivative of the output metric.

$$g_i(\theta) = \frac{\partial f(\theta)}{\partial \theta_i} \tag{2}$$

gives the gradient of f with respect to the ith input parameter.

2.1 Introduction to finite difference method

In a one dimensional case, the derivative of a function f by first principles is given by

$$g(\theta) = \lim_{c \to 0} \frac{f(\theta + c) - f(\theta - c)}{2c} \tag{3}$$

When c, the step size is small, we can reasonably estimate the gradient by estimating the function f at $\theta + c$ and $\theta - c$.
The finite defference (FD) method of estimating the gradient is given by

$$\hat{g}_i(\theta) = \frac{\hat{f}(\theta + c_i e_i) - \hat{f}(\theta - c_i e_i)}{2c_i} \tag{4}$$

where

c_i = step size.

e_i = unit vector in the ith direction.

Thus we can estimate the gradient by conducting one simulation with input parameter $\theta + c_i e_i$ and obtain an estimate of $f(\theta + c_i e_i)$ and conduct another simulation at $\theta - c_i e_i$ and obtain an estimate of $f(\theta - c_i e_i)$. Equation 4 gives the gradient with respect to one input parameter. The gradient can be estimated for $i = 1, 2...p$ parameters by $2p$ simulations with step size c_i and unit vector e_i for $i = 1, 2...p$. One of the problems with the finite difference estimator is that when the step size is small, the variance of the estimators becomes large and when the step size increases, the bias of the estimate increases. So choice of the simulation parameters like number of replications and the choice of the estimator parameters like step size should be done carefully.

2.2 Introduction to simultaneous perturbation method

The simultaneous perturbation (SP) gradient estimation method uses just two simulations for estimating all the gradients.The SP gradient estimator for a process with p input parameters and one output metric, f is given as follows

$$\hat{g}_i(\theta) = \frac{\hat{f}(\theta + C\Delta) - \hat{f}(\theta - C\Delta)}{c_i \Delta_i} \tag{5}$$

where

Δ = a random p-dimensional perturbation vector.

C = A diagonal matrix with step sizes for the input parameters in the diagonal row. The reasoning behind the representation of the step size as a diagonal matrix is explained with the help of equation (6). c_i is the i-th diagonal element in C.

$$\hat{f}(\theta + C\Delta) = \hat{f}\left(\begin{bmatrix} \theta_1 \\ \theta_2 \\ .. \\ \theta_n \end{bmatrix} + \begin{bmatrix} c_1 & 0 & 0 & .. & 0 \\ 0 & c_2 & 0 & .. & 0 \\ .. & .. & .. & .. & .. \\ 0 & 0 & 0 & .. & c_n \end{bmatrix} \begin{bmatrix} \Delta_1 \\ \Delta_2 \\ .. \\ \Delta_n \end{bmatrix}\right) = \hat{f}\left(\begin{bmatrix} \theta_1 + c_1\Delta_1 \\ \theta_2 + c_2\Delta_2 \\ .. \\ \theta_n + c_n\Delta_n \end{bmatrix}\right) \tag{6}$$

Each element of Δ is independently generated from a probability distribution with mean zero. The second inverse moment of Δ should be not be infinite which means that Δ cannot be taken from an uniform or normal distribution. The rationale behind proper choice of Δ is explained in detail in [12]. The method differs from the FD method in that all the input parameters are simultaneously perturbed during a single simulation. In the two simulation runs which are needed to estimate the gradient using the SP method, the perturbation in parameters for one simulation run will be exactly vice versa compared to the other simulation run. Hence for gradient estimates for different input parameters, only the denominator of the above formula will be varying as Δ varies while the numerator will remain the same. Also here c_i, the step size, may remain the same for different input parameters or may be scaled for different input parameters, if the input parameters vary greatly in magnitude.

3 Problem statement

We consider the problem of estimating the sensitivity of the steady state average total cycle time (CT) to the processing times (PT) of each operation in the manufacturing system. The manufacturing system is a flow shop with no reentrant flow. The manufacturing system produces just one product. This problem is important is because the impact of processing times on total average cycle time will give the people who manage the system information on the importance of the process parameters. The manufacturing system has seven workstations. The seven workstations are Coater, Stepper, Developer, Exposer, Printer, Reader and Writer. Table 1 gives the number

Tool Group	Number of tools	Processing Time (in Hrs)
Coater	2	5
Stepper	1	1
Developer	2	5
Exposer	2	6
Printer	1	3
Reader	1	2
Writer	2	7

Table 1: Toolgroups in the model and their parameters

of tools at each workstation and the mean processing time of that operation at that workstation. The processing times at each workstation are exponentially distributed.

The product is a wafer, which enters the factory in lots of one unit each. The lots enter with an average interarrival time of four hours. The interarrival times are exponentially distributed. The example is depicted in Figure 2. We will use the Factory Explorer simulation tool [14] to simulate the system and obtain estimates of the average total cycle time.

4 Gradient estimate using finite difference method

Gradient measurement using FD method can be done through several sub-methods like the forward difference, backward difference and central difference methods. We will use the central difference method because the gradient estimate from the central difference method will usually have less bias than the forward or backward difference. The FD estimator follows:

$$(\hat{g}_i(\theta))_N = \frac{1}{N}\sum_{j=1}^{N}\left(\frac{\hat{f}_j(\theta + c_i e_i) - \hat{f}_j(\theta - c_i e_i)}{2c_i}\right) \tag{7}$$

Where

$\hat{g}_i$ = Estimate of the ith component of the gradient vector.

$\hat{f}_j$ = jth estimate of the function, which is obtained from simulation.

c_i = Step size for the ith parameter. Here $c_i = \theta_i/100$

θ = Vector of baseline input parameters.

N = Number of replications.

e_i = Unit vector in direction i.

The simulation tool used for conducting simulations considers the time duration for which the system is simulated rather than the number of customers, so we simulate the system for 87,600 hours (10 years). To obtain reasonable accuracy, we perform $N = 20$ replications. The cycle time and gradient estimates for one parameter over 20 replications are given in Table 4. Since the model

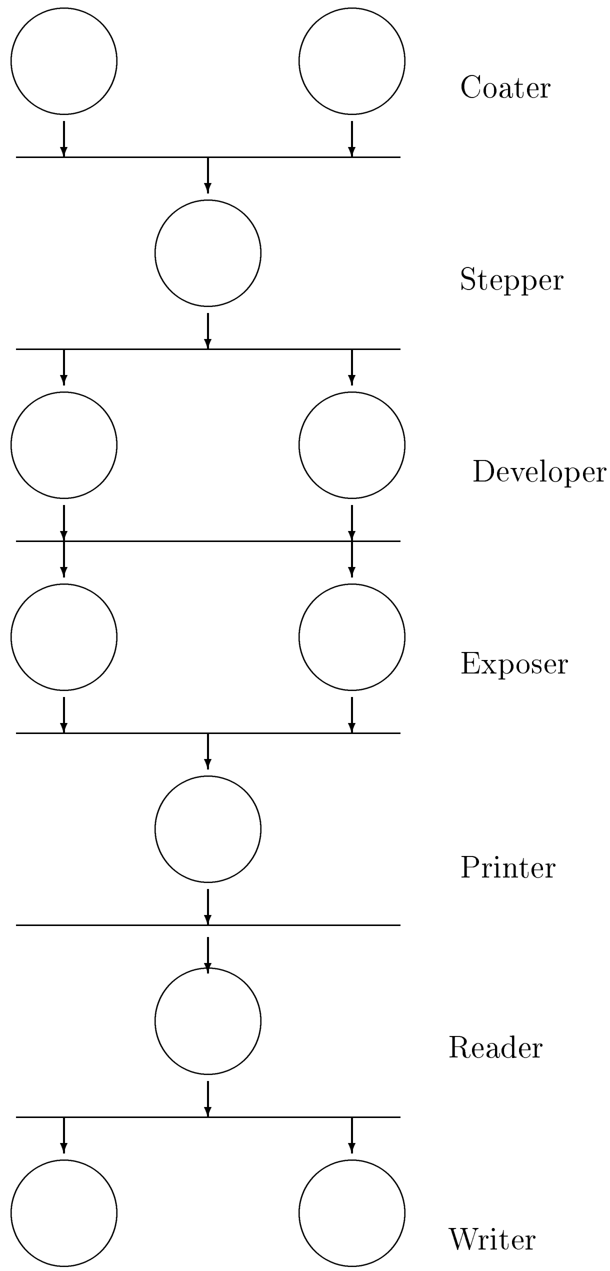

Figure 2: Input model - manufacturing system

has seven input parameters we need a total of 280 simulation runs.

An important parameter in the FD method is the step size. In the FD method, a large step size yields estimates with high bias, but a small step size yields estimates with high variances. For this model, we consider a step size $c_i = (0.01)\theta_i$. This step size is relatively small but using a higher number of replications can reduce the variance.

Based on the chosen values that define the logistics of the simulation and input parameters, the gradients for the average total cycle time with respect to the mean processing parameters are estimated. The confidence intervals are built for a confidence of 99%. The confidence interval is obtained by the following methodology. The standard error which is given by equation (8) is calculated first. Then the half width of the confidence interval is calculated as per equation (9).

$$S = \sqrt{\frac{\sum_{i=1}^{N} X_i^2 - N\bar{X}^2}{N-1}} \tag{8}$$

where

X_i = Gradient estimate at replication i.

$\bar{X}$ = Mean gradient estimate over N replications.($\bar{X} = (\sum_{i=1}^{N} X/N)$)

$$h = t_{N-1,1-\alpha/2}\frac{S}{\sqrt{N}} \tag{9}$$

where

t = a constant which is obtained from statistical tables depending on α and N.

$\alpha = 0.01$, if 99% is the confidence needed in the estimate.

The confidence interval is given by $(\bar{X} - h, \bar{X} + h)$. Table 4 gives the summary data for the FD method including the cycle time and gradient estimates. Figure 3 gives a graphical representation of the gradient results compared with the gradients from the analytical method.

An important conclusion which can be obtained from the finite difference method is that the changes in average cycle times of the machines whose mean processing times are varied consist of almost the total change in the average cycle time of the manufacturing system. This is facilitated by the finite difference method where we estimate the average cycle time on a per-parameter basis. This is reflected in the data in Table 4.

5 Gradient Estimate using Simultaneous Perturbation

The second method being discussed here is the SP method. Here the gradient is estimated for the i-th parameter as follows. g is the p-dimensional vector of gradients.

$$(\hat{g}_i(\theta))_N = \frac{1}{N}\sum_{j=1}^{N}\left(\frac{\hat{f}_j(\theta + C\Delta_j) - \hat{f}_j(\theta - C\Delta_j)}{2c_i\Delta_{ji}}\right) \tag{10}$$

Replication	Average cycle time-lower	Average cycle time-higher	Gradient
1	81.164	81.539	3.75
2	77.400	77.779	3.79
3	78.186	78.566	3.80
4	83.612	83.938	3.26
5	74.415	74.898	4.83
6	77.952	78.677	7.25
7	77.179	77.394	2.15
8	78.973	80.158	11.85
9	79.944	80.447	5.03
10	76.042	76.326	2.84
11	78.114	79.271	11.57
12	77.876	77.857	-0.19
13	73.183	73.015	-1.68
14	79.941	79.987	0.46
15	78.657	79.044	3.87
16	75.986	77.418	14.32
17	78.534	78.605	0.71
18	76.034	76.623	5.89
19	75.508	75.391	-1.17
20	76.261	76.851	5.90
		Average	4.411
		Standard error	0.955
	Half width of confidence interval		2.415
	Confidence interval (1.986,6.837)		

Table 2: Table showing cycle time and gradient estimates estimated by finite difference method for the process coater

where

$\hat{g}_i$ = Estimate of the gradient for the ith input parameter

p = number of input parameters

N = number of replications

θ = vector of baseline mean processing times

Δ_j = a random p-dimensional perturbation vector, which changes at each replication.

C = the diagonal matrix of step sizes

c_i = step size for the ith input parameter. Here $c_i = \theta_i/100$.

The implementation of the gradient estimator has been studied in depth in [11] as part of a study on stochastic optimization using simultaneous perturbation stochastic approximation. The gradient for all input parameters for one replication can be calculated with only two simulations.

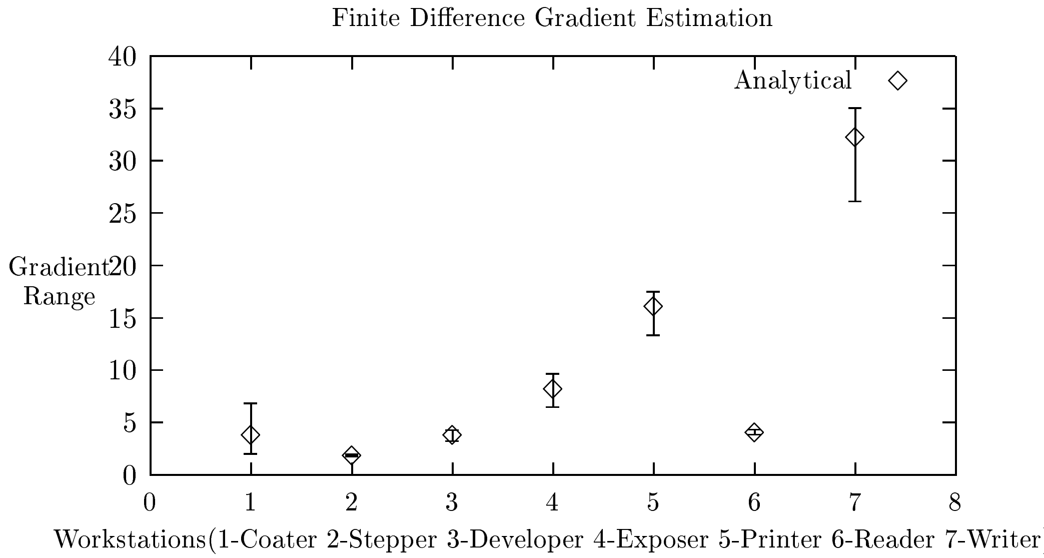

Figure 3: Distribution of the gradient obtained from Finite Difference method and comparison with the gradient obtained by analytical method for each workstation

Tool	Coater	Stepper	Developer	Exposer	Printer	Reader	Writer
Number of tools	2	1	2	2	1	1	2
Mean processing time-baseline	5	1	5	6	3	2	7
Processing time-lower	4.95	0.99	4.95	5.94	2.97	1.98	6.93
Average cycle time-lower	77.748	77.990	77.795	77.538	77.558	77.926	76.015
Mean processing time-upper	5.05	1.01	5.05	6.06	3.03	2.02	7.07
Average cycle time - upper	78.189	78.026	78.171	78.504	78.482	78.091	80.300
Gradient	4.411	1.840	3.762	8.052	15.398	4.109	30.602
Confidence Interval	(1.986, 6.837)	(1.753, 1.927)	(3.262, 4.262)	(6.447, 9.657)	(13.312, 17.483)	(3.852, 4.365)	(26.121, 35.083)

Table 3: Summary Data for Finite Difference method

Tool	Change in average cycle time at tool	Change in total average cycle time	Difference
Coater	0.381	0.441	0.060
Stepper	0.036	0.037	0.001
Developer	0.379	0.376	-0.002
Exposer	0.913	0.966	0.053
Printer	0.982	0.924	-0.058
Reader	0.161	0.164	0.003
Writer	4.285	4.284	0.000

Table 4: Table comparing the Change in CT at a tool and the change in total CT when the PT for that tool is varied between the upper and lower levels

Tools	Number of Tools	Mean processing time (Hours)			Gradient	Confidence Interval
		Lower		Upper		
Coater	2	4.95	5	5.05	5.574	(-4.780,15.928)
Stepper	1	0.99	1	1.01	-22.743	(-74.512,29.027)
Developer	2	4.95	5	5.05	3.809	(-6.545,14.162)
Exposer	2	5.94	6	6.06	11.631	(3.003,20.259)
Printer	1	2.97	3	3.03	13.664	(-3.592,30.921)
Reader	1	1.98	2	2.02	11.146	(-14.738,37.031)
Writer	2	6.93	7	7.07	34.979	(27.583,42.374)

Table 5: Summary Data for simultaneous perturbation method

The Δ vector considered here is obtained from a Bernoulli distribution. It consists of p i.i.d symmetric Bernoulli random variables X_i. $P\{X_i = 1\} = 0.5$. $P\{X_i = -1\} = 0.5$. The step size considered here is $c_i = (0.01)\theta$. Setting c_i as a function of θ_i takes care of the differences in the magnitudes of the processing times. Table 5 gives the summary data for the SP method including the cycle time and gradient estimates.

The SP gradient estimation is done for $N = 140$ replications. This facilitates comparison between SP and FD. In the FD method, the gradient is estimated by changing the numerator and keeping the denominator constant. But in the SP method, the numerator is kept constant and the denominator is changed for calculating the gradients of different parameters. Hence while SP method requires two simulations to estimate the gradient for seven parameters, the FD method needs 14 simulations. Hence when we have $N = 20$ replications for the finite difference method, we can have $N = 140$ replications for the SP method.

Figure 4 gives a graphical representation of the gradient results compared with the gradients from the analytical method.

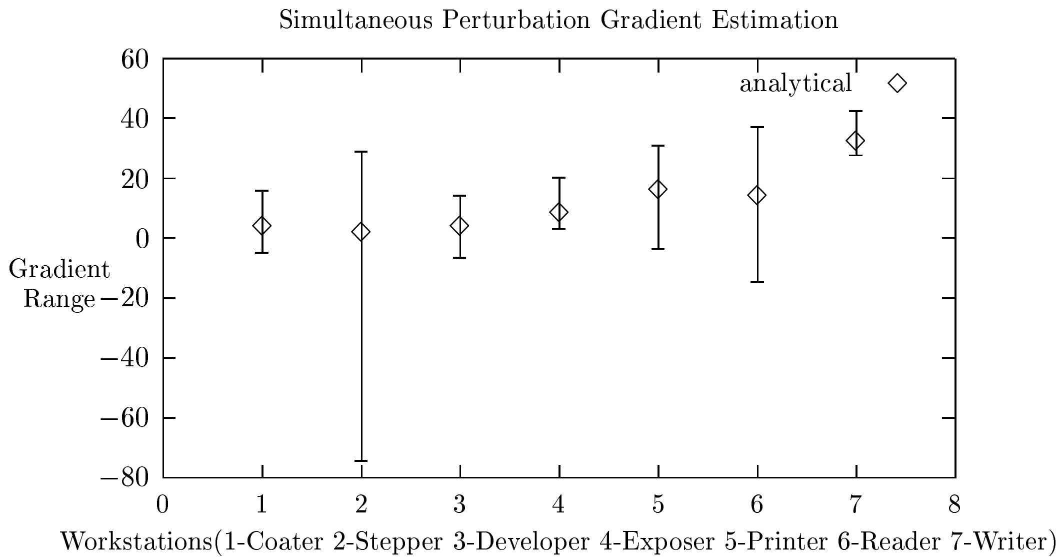

Figure 4: Plot showing the distribution of the gradient obtained from Simultaneous Perturbation method for each workstation and comparison with the gradient obtained by analytical methods

6 Analytical verification

The gradients obtained by the FD and SP methods were compared to the partial derivatives calculated exactly from queueing system models. In the example considered, we have workstations with one or two tools each. Each station acts as an M/M/1 or M/M/2 queueing system, since the interarrival times and processing times are exponentially distributed. The cycle times at each tool can be calculated using exact models for M/M/1 and M/M/2 systems. Expressions for the utilization, average cycle time and gradient for the M/M/1 and M/M/2 queueing systems are given below [2, 10].

For the M/M/1 system:

$$u_i = r_a t_i \tag{11}$$

$$CT_i = \frac{t_i}{(1-u_i)} \tag{12}$$

$$Partialderivative = \frac{\partial(CT_i)}{\partial(t_i)} = \frac{1}{(1-u_i)^2} \tag{13}$$

where

u = Utilization of the tool at workstation i

r_a = Arrival rate = (1/mean interarrival time)

t_i = Mean processing time at workstation i

Tool	Arrival Rate (Jobs/Hour)	Number of tools	Mean Processing Time(Hours)	Utilization	Average Cycle time (Hours)	Partial Derivative
Coater	0.25	2	5	62.5	8.205	3.745
Stepper	0.25	1	1	25.0	1.333	1.778
Developer	0.25	2	5	62.5	8.205	3.745
Exposer	0.25	2	6	75.0	13.714	8.163
Printer	0.25	1	3	75.0	12.000	16.000
Reader	0.25	1	2	50.0	4.000	4.000
Writer	0.25	2	7	87.5	29.867	32.142

Table 6: Summary Data for Analytical Method

For the M/M/2 system:

$$u_i = \frac{r_a t_i}{2} \tag{14}$$

$$CT_i = \frac{t_i}{(1-u_i^2)} \tag{15}$$

$$Partialderivative = \frac{\partial(CT_i)}{\partial(t_i)} = \frac{(1+u_i^2)}{(1-u_i^2)^2} \tag{16}$$

where

u = Utilization of the tool

r_a = Arrival rate = (1/mean interarrival time)

t_i = Mean processing time at workstation i

Table 6 lists the exact average cycle times and gradients.

7 Summary

The FD method provided reasonably good estimates for the gradient of average total cycle time with respect to the mean processing times while the SP method could not perform as well as the FD method. It gave poor confidence limits for the gradients though the mean gradient was quite accurate for some of the parameters. This could be due to the fact that the estimate for one value depends on the way in which one variable affects the others during cycle time estimation, which results in the high noise levels in the measurements of the gradients. When we compare the gradient estimates of both the methods against the exact method we can see that, the FD method has significantly performed better than SP.

These gradient methods are of significant use in complex manufacturing systems like semiconductor manufacturing systems where we have a large number of input parameters which affect the average total cycle time. These gradient estimation methods can estimate the impact that these input parameters have and identify the parameters that have the maximum impact on system performance.

References

[1] Banks Jerry, Ed., *Handbook of simulation*, Wiley Interscience., 1998.

[2] Cassandras, Christos G., *Discrete event systems : modeling and performance analysis*, Irwin, Homewood, IL., 1993.

[3] Fu, Michael C., "Optimization via simulation : a review", *Annals of operations research*, Vol.53, pp.199-247, 1994.

[4] Glynn, P. W., "Likelihood ratio gradient estimation or stochastic systems", *Communications of the ACM*, Vol. 33, pp.75-84, 1990.

[5] Ho, Y.-C., and X.-R.Cao, *Perturbation analysis of discrete event dynamical systems*, Kluwer, Norwell, Mass., 1991.

[6] Jacobson, S.H., "Convergence results for harmonic gradient estimators", *ORSA journal of computing*, Vol.6, pp.381-397, 1994.

[7] L'Ecuyer,P., N.Giroux, and P.W.Glynn., "Stochastic optimization by simulation: numerical experiments with the M/M/1 queue", *Management science*, Vol. 40, pp.1245-1261, 1994.

[8] L'Ecuyer,P., "An overview of derivative estimation", *Proceedings of the 1991 winter simulation conference*, pp.207-217, 1991.

[9] Law Averill M., and Kelton W. David., *Simulation modeling and analysis*, Mc-Graw Hill, 1991.

[10] Panico, Joseph A., *Queueing theory : a study of waiting lines for business, economics and science*, Prentice-Hall Inc., Englewood Cliffs, N.J., 1969.

[11] Spall, J.C., "Implementation of the simultaneous perturbation algorithm for stochastic optimization" *IEEE transactions on aerospace and electronic systems*, Vol. 34, No. 3, pp.817-823, 1998.

[12] Spall, J.C., "Multivariate stochastic approximation using a simultaneous perturbation gradient approximation" *IEEE Transactions on automatic control*, Vol. 37, pp. 332-341, 1992.

[13] Suri, Rajan., "Perturbation analysis: The state of art and research issues explained via the GI/G/1 queue", *Proceedings of the IEEE*, Vol. 77,No. 1, pp.114-137, IEEE, 1989.

[14] *Factory Explorer, user's manual* Wright Williams and Kelley, 1998.