Using Multiple Runs in the Simulation of Stochastic Systems for Estimating Equilibrium Expectations

Winfried Grassmann
Department of Computer Science, University of Saskatchewan, Saskatoon, SK S7N 5C9, Canada

Keywords: Steady-state Simulation, Multiple Runs, Initialization Bias.

Abstract: In complex stochastic systems, Monte-Carlo simulation is often the only way to estimate equilibrium expectations. The question then arises what is better: a single run of length T, or n runs, each of length T/n. In this paper, it is argued that if there is a good state to start the simulation in, multiple runs may be advantageous. To illustrate this, we use numerical examples. These examples are obtained by using deterministic methods, that is, methods based on probability theory not using Monte-Carlo methods. The results of our numerical calculations forced us to make a sharp distinction between the time to reach equilibrium and the appropriate length of the warm-up period, and this distinguishes our study from earlier investigations.

1 INTRODUCTION

Monte Carlo methods are often the only way to estimate equilibrium expectations in complex stochastic systems, such as queueing networks. The question then arises whether a single long run is preferable to several shorter runs. Hence, instead of doing one run of length T, should we do n runs, each of length T/n? This is a question that has been addressed by several authors, including (Whitt, 1991), (Alexopoulos and Goldman, 2004) and (Kelton, 1989). In contrast to these studies, we make a sharp distinction between the time to reach equilibrium within a given tolerance, and the appropriate length of the warm-up period. Here, the warm-up period is a period during which the data obtained by the simulation is ignored. We were forced to do this because in our numerical experiments, the time to reach the equilibrium within $\pm 10\%$ was clearly much longer than the warm-up period that minimizes the mean squared error of the estimate as shown in (Grassmann, 2011). The problem of multiple runs is also related to the initialization bias problem, for details see (Pasupathy and Schmeiser, 2010), (Pawlikowski, 1990) and (Grassmann, 2014), and references therein.

Our investigation was motivated by a paper by Madansky (Madansky, 1976), who showed that when estimating the expected number of elements in an $M/M/1$ queue, the best starting state, judging by the mean squared error (MSE), is the empty state as long as T is not too short, a result confirmed by (Wilson and Prisker, 1978) and (Grassmann, 2008). Moreover, in (Grassmann, 2008) it was shown that if starting the $M/M/1$ queue empty, any warm-up period increases the MSE, that is, the optimal warm-up period is 0. However, when starting at 0, the expected queue length is far from its equilibrium value, that is, the time to reach equilibrium at a tolerance of $\pm$ 10% is far from 0. This observation forces us to make a sharp distinction between the time to reach equilibrium within a given tolerance and the appropriate length of a warm-up period, a distinction that is somewhat hazy in literature. As it turns out, the distinction between the time to reach equilibrium and the optimal length of the warm-up period forces us to re-evaluate the issue of multiple runs.

People may argue that the $M/M/1$ queue is special, and results derived from its study are invalid. This opinion leads to the philosophical issue as to how many counter-examples one needs in order to refute strongly held beliefs. Even though in mathematics, a single counter-example is sufficient, we believe more than one example is needed. However, (Grassmann, 2014) provides additional cases where it is crucial to distinguish between the time to reach equilibrium within a given tolerance and the warm-up period. Moreover, I feel that before making any conclusions about the generality of an observation, we have to understand the reasons why the time to reach equilibrium within a given tolerance is not necessarily a good indicator for the length of the warm-up period. Fortunately, relevant reasons are given in (Grassmann,

Grassmann W..
Using Multiple Runs in the Simulation of Stochastic Systems for Estimating Equilibrium Expectations.
DOI: 10.5220/0005535501380144
In *Proceedings of the 5th International Conference on Simulation and Modeling Methodologies, Technologies and Applications* (SIMULTECH-2015), pages 138-144
ISBN: 978-989-758-120-5

2014), a paper that provides a good understanding of the main issues, and without such an understanding, the question as to what models are exceptions, and why, cannot be addressed in any scientific way.

The setup we use is like the one of (Whitt, 1991): The system is described by d state variables $X_1(t)$, $X_2(t)$, ..., $X_d(t)$, and we define $X(t) = [X_k(t), k = 1,2,\ldots,d]$. We are interested in the process $R(t) = f(X(t))$, and the problem is to find the expectation $\lim_{t\to\infty} \mathrm{E}(R(t) = \mathrm{E}(R)$. If $d = 1$, we use $X(t)$ instead of $[X_1(t)]$. We assume that the process is ergodic, in which case $\mathrm{E}(R)$ exists and it is unique. $R(t)$ could be, for instance, the number of jobs in a queueing network with 3 queues. In this case $f(X(t))$ becomes

$$f([X_1(t),X_2(t),X_3(t)]] = X_1(t) + X_2(t) + X_3(t).$$

We assume that we know $f(.)$. The question is how many runs should be made to find the estimate for $\mathrm{E}(R)$ with the lowest possible MSE. We assume that the runs are of equal length, and that the total time for all runs is T. Hence, if there are n runs, each run has a length of T/n. We use time averages to estimate $\mathrm{E}(R)$, that is:

$$\bar{R}_\nu(T/n) = \frac{1}{T}\int_0^{T/n} R(t)dt \qquad (1)$$

and we take the overall average

$$\bar{\bar{R}}(T) = \sum_{\nu=1}^{n} \bar{R}_\nu(T/n). \qquad (2)$$

If there is only 1 run, we use $\bar{R}(T)$ instead of $\bar{\bar{R}}(T)$. All runs start in the same state, which implies that all $R_\nu(T/n)$ have the same distribution. We will therefore sometimes omit the subscript. The common expectation will be denoted by $\mathrm{E}(\bar{R}(T/n))$, and the common variance by $\mathrm{Var}(\bar{R}(T/n))$. Equation (1) assumes continuous sampling, but similar formulas can easily derived if sampling is only done at certain epochs, or if the data sampled is discrete, as is the case when considering successive waiting times. Also note that for reasons discussed later, we use no warm-up period. If a warm-up period of length w is used, then the lower bound of the integral of (1) must be replaced by w, and $\frac{1}{T}$ by $\frac{1}{T-w}$. To judge the quality of $\bar{\bar{R}}(T)$, we use the mean squared error, that is

$$\begin{aligned}\mathrm{MSE}(\bar{\bar{R}}(T)) &= \mathrm{E}((\bar{\bar{R}}(T) - \mathrm{E}(R))^2) \\ &= \mathrm{Var}(\bar{\bar{R}}(T)) + \mathrm{Bias}^2(\bar{\bar{R}}(T)) \qquad (3)\end{aligned}$$

where

$$\mathrm{Bias}(\bar{\bar{R}}(T)) = \mathrm{E}(\bar{\bar{R}}(T)) - \mathrm{E}(R). \qquad (4)$$

When we use the term MSE, we always refer to $\mathrm{MSE}(\bar{\bar{R}}(T))$ or, if there is only one run, $\mathrm{MSE}(\bar{R}(T))$.

Our attention is restricted to point estimations, as opposed to interval estimations and confidence intervals (Alexopoulos and Goldman, 2004). Of course, multiple runs have the additional benefit to provide some idea about the reliability of an estimate.

The models we use are rather simple, such as the $M/M/1$ queue or a simple tandem queue. However, from such simple models, insights can be gained that allow us to form, and potentially prove, interesting conjectures when analyzing more complex models. It would be much more difficult to gain these insights by looking at complex systems. For simple models, there are very effective methods to find expectations and variances of time averages (Grassmann, 1987), methods that are deterministic in the sense that they do not use Monte Carlo techniques, but probability theory. Incidentally, finding variances by simulation with acceptable precision requires a large number of observations, and this leads to long execution time, making these method non-competitive for small models. However, since deterministic methods increase exponentially with the dimensionality of the model, simulation is often the only way to analyze models with many state variables.

2 WHY MULTIPLE RUNS CAN DECREASE THE MSE

Though it may sound contradictory, warm-up periods can increase the MSE. As it turns out, these are exactly the cases were multiple runs can be advantageous. In (Grassmann, 2014), a number of reasons have been given why warm-up periods can increase the MSE when initializing the system in certain states. We list these reason here to make the paper self-contained.

1. Consider the case where sampling is not done continuously, but only at time 0, τ, 2τ, 3τ, ..., $m\tau = T$. If τ is large enough, then the readings at these sample points are almost independent random variables. Now, if the state at time 0 is close to $\mathrm{E}(R)$, then it would make sense to include it in the time average, that is, instead of $\bar{R}(T) = \frac{1}{m}\sum_{\nu=1}^{m} R(\nu\tau)$, one uses $\frac{1}{m+1}\sum_{\nu=0}^{m} R(\nu\tau)$. In fact, this would amount to a weighted average with a weight of $\frac{1}{m+1}$ for $R(0)$ and a weight of $\frac{m}{m+1}$ for $\bar{R}(T)$, which makes sense, provided $R(0)$ is a reasonable estimate for $\mathrm{E}(R)$.
2. In many cases, $\mathrm{E}(R(t))$ is subject to a drift, that is, $\mathrm{E}(R(t))$ changes in a predictable fashion. In this case, starting in a state with $R(0)$ close to $\mathrm{E}(R)$ may not be a good starting state. Consider, for in-

stance, an $M/M/1$ queue with arrival rate λ and service rate μ, with $\lambda < \mu$. If the system starts in a fixed non-empty state, then in a small interval of length h, the expected number of arrivals is $\lambda h + o(h)$, and the expected number of departures is $\mu h + o(h)$, which means that the expected queue length decreases at a rate $\lambda - \mu$, a rate that is negative. Similar effects were observed by (Kelton and Law, 1985), using numerical experiments. Hence, before the $M/M/1$ queue reaches the empty state, the expectation will always decrease, precluding thus to be in an equilibrium. In other words, no reasonable equilibrium estimate can be obtained before reaching 0. This may explain the findings of Mandansky (Madansky, 1976) that the smallest MSE for the expected number of elements in an $M/M/1$ queue is obtained when starting the system empty, provided the simulation is of a reasonable length.

3. Most textbooks (Tocher, 1963), (Banks et al., 2005), (Law and Kelton, 2000) suggest to start the system in a typical state. This allows for different interpretations: One could call a state typical if it has a high equilibrium probability, or if it is frequently visited while the process is stationary. If the system starts in such a typical state, warm-up periods may be detrimental.

Choosing an initial state according items 1 to 3 is only possible if some prior information about the system is known. However, if prior information is known, its weight can be increased by doing multiple runs. Moreover, the better the information, the more runs should be done.

To find the states where a warm-up period can reduce the MSE, consider the integral given by (1), with the lower bound replaced by w, and with $\frac{1}{T}$ changed accordingly. Suppose there is only 1 run. Let $R(T,w)$ be the resulting expression. We now form the derivatives of $\mathrm{Bias}(\bar{R}(T,w))$, $\mathrm{Var}(\bar{R}(T,w))$ and $\mathrm{MSE}(\bar{R}(T,w))$ with respect to w for $w = 0$ to get according to (Grassmann, 2014):

$$\begin{aligned} \mathrm{Bias}'(\bar{R}(w,T))_{w=0} &= \mathrm{E}'(\bar{R}(w,T))_{w=0} \\ &= \frac{1}{T}(E(\bar{R}(T)) - f(a)) \quad (5) \\ \mathrm{Var}'(\bar{R}(w,T))_{w=0} &= \frac{2}{T}\mathrm{Var}(\bar{R}(T)) \quad (6) \\ \mathrm{MSE}'(\bar{R}(w,T))_{w=0} &= \frac{2}{T}(\mathrm{Var}\bar{R}(T) + \\ &\quad \mathrm{Bias}(\bar{R}(T))(\mathrm{E}(\bar{R}(t)) - f(a))). \quad (7) \end{aligned}$$

Here, $f(a) = f(X(0)) = R(0)$ as defined in the introduction. If the derivative of the MSE is negative, any warm-up period decreases the MSE, that is, a warm-up period is beneficial. On the other hand, if the derivative of the MSE is positive, a warm-up period of zero is a local minimum, and it is reasonable to expect that this minimum is also global. In this case, any warm-up period is detrimental.

States with a positive derivative are obviously good states to start a simulation in, and the benefits of starting in these states could potentially be increased by doing multiple runs. In fact, we observed that for most cases with positive derivatives with respect to w, multiple runs proved to be beneficial. Surprisingly, we even found some cases with a negative derivative for the MSE where multiple runs are optimal.

Next, we investigate as to what happens if T increases by a factor of k. Consider first the case where there is only one run. If T is large enough, then according to (Asmussen and Glynn, 2007), $\mathrm{Var}(\bar{R}(T))$ and $\mathrm{Bias}(\bar{R}(T))$ both decrease by a factor of k as T increases by a factor of k. This relation was originally suggested in (Grassmann, 1982), and (Grassmann, 2008) shows numerically that convergence to the limiting form can be quite fast. Because of (3), this implies that in the limit, the contribution of the bias decreases by k^2, whereas the variance decreases by k. Hence, as T increases, and T is large enough, the contribution of the bias diminishes.

If there are multiple runs, then we conclude from equations (2) and (1):

$$\mathrm{MSE}(\bar{\bar{R}}(T)) = \frac{1}{n}\mathrm{Var}(\bar{R}(T)) + \mathrm{Bias}^2(\bar{R}(T)). \quad (8)$$

It follows that compared to a single run, the importance of the bias increases.

If n, the number of runs, is too large, T/n is very small, and $\bar{\bar{R}}(T)$ is essentially an average of the initial conditions, making the contribution of the simulation almost meaningless. This follows from (8), which implies that the relative importance of the variance decreases with n, approaching 0 as $n \to \infty$. This means that as n increases, starting states close to $\mathrm{E}(R)$ become increasingly competitive. If $R(0)$ is very close to $\mathrm{E}(R)$, then it is optimal to make a large number of very short runs. For very short runs, $\mathrm{E}(\bar{R}(T)) \approx R(0)$, meaning that we essentially take the average of the initial conditions, practically ignoring the results of the simulation. The simulation runs are then almost useless. In order to avoid such meaningless results, we have to require that the run length T/n is long enough such that the simulation can make a reasonable contribution.

3 DRIFT AND FLOW MODELS

In this section, we discuss the drift in the context of flow models (Newell, 1982). These models allow us to find good initial states. Flow models are also helpful for deciding when the number of runs is so large, and the individual runs are so short as to make the simulation meaningless.

Note that every state $i = [i_1, i_2, \ldots, i_d]$ is associated with a drift for each state variable $X_k(t)$. The drift is the derivative of $\mathrm{E}(X_k(t))$ with respect to t, given $X_k(t) = s_k$, $k = 1, 2, \ldots, d$. In a typical system, almost all states have a non-zero drift, that is, whenever one knows the state, there tends to be a drift of at least some state variables. Stochastic system in equilibrium have no drift because they can be in many different states, but we have no knowledge in which state they actually are. Of the potential states the system could be in, state variables of some states have an up-drift, which is compensated by other states where these variables have a down-drift. This compensation vanishes if we know the state we are in.

If all states visited from 0 to t either have all an up-drift for some relevant state variable, or they all have a down-drift for this variable, a simulation run extending only from 0 to t cannot be expected to provide a meaningful estimate of the equilibrium expectation. The minimum time for a simulation is therefore the time until all relevant state variables have changed the sign of the drift at least once. Hence, we must visit a pair of states (i, j) at least once where the sign of the drift changes for a particular state variable when going from i to j. Theoretically, such pairs would have to be found for every state variable, but for reasons to be discussed, there are often pairs of states where all state variables either change the sign of the drift, or where the drift goes at least to zero when going from one state of the pair to the other. To find such pairs of states, we use a flow analysis (Newell, 1982), with the flows being equated to the drifts.

To make the conversion of our stochastic model to the flow approximation, we have to make all state variables continuous. Typically, we have simple formulas for the drift, and though these formulas possibly apply only to integers, in a flow model, we disregard this restriction. For instance, in the $M/M/c$ queue with arrival rate λ and service rate μ, the drift when i elements are in the system is $\lambda - i\mu$ for $0 < i \leq c$, and this remains the formula for the flow model, except that i is now continuous.

A flow model is in a state of equilibrium if no state variable changes as the time t increases. In other words, the drifts of all state variables in the flow approximation are zero. Let $s = [s_1, s_2, \ldots, s_d]$ be this equilibrium state. If the equilibrium state s is stable, and we assume that this is the case for now, then from any state close to s there is a drift towards s. For instance, in the flow approximation of the $M/M/c$ queue, the equilibrium state is $s = \lambda/\mu$, because for the $M/M/c$ queue, we have a drift $\lambda - i\mu$, which is zero for $i = \lambda/\mu$.

If there are two states, $i = [i_1, i_2, , \ldots, i_d]$ and $j = [j_1, j_2, \ldots, j_d]$, with $i_k > s_k$ and $j_k < s_k$ for some k, then being attracted to s, i must have a down-drift, and j an up-drift. Similar results would have to be true for all k. If the model is stochastic, then it may be possible to go from i to j, and if the path goes through s, the sign of the drift changes for all state variables, or at least it falls to 0. In a stochastic system, the path may not exactly move through s, but we still can use this result as a heuristic. If a state variable is integer, we have to round it either up or down when converting the equilibrium state in the flow approximation to a state in the stochastic model. If i and j are two states obtained from s through rounding, and if one can go in one step from i to j, then the drift typically changes the sign for most state variables when going from i to j. In the $M/M/c$ model, for instance, the flow equilibrium is $r = \lambda/\mu$. If r is not integer, it is easily verified that the sign of the drift changes if one goes from $\lfloor r \rfloor$ (r rounded down) to $\lceil r \rceil$ (r rounded up), or from $\lceil r \rceil$ to $\lfloor r \rfloor$. If all state variables of s that are supposed to be integer are integer already, we only have one state rather than a pair of states. To get a pair in this case, we change one state variable by 1. This only makes the drift change to zero when moving from the modified state to s, but a sign change is very likely to occur soon. For instance, if r in the $M/M/c$ queue is integer, $\lfloor r \rfloor = \lceil r \rceil$. In this case, we have to extend the simulation to a time long enough to move beyond this no-drift state.

In some flow models, there is no state with a flow of zero, and in this case, there is either no equilibrium, and consequently no meaningful equilibrium expectation $\mathrm{E}(R)$ to estimate, or the state variables hit some boundary. For instance, in the $M/M/1$ queue, the drift for all non-zero states is $\lambda - \mu$, and if $\lambda < \mu$, the flow model hits a boundary at zero. It is easily verified that when hitting the boundary in an $M/M/1$ queue, or even in an open network of $M/M/1$ queues, the sign of the drift changes at the boundary points. The same can be expected for most other systems when they hit a boundary.

In summary, the minimum length of the simulation should be such that we either go beyond a sign change of the drift of all relevant state variables, or at least a state where the drift is zero. Hence, we have to find pairs (i, j) where the drift of all relevant state

variables changes as we go from i to j, or where the drift of all state variables goes to zero. Before j is reached, the drift is always in the same direction, and the results of the simulation are close to meaningless.

Drift-change pairs (i, j) can be unstable, that is, though the drift changes sign when going from i to j, the drift can move us away from j. For example, consider a birth-death process with a maximum population of N, and with birth rate λ_i, $0 \leq i < N$ and death rate μ_i, $0 < i \leq N$, where i is the size of the population. If there is a value k such that $\lambda_i - \mu_i > 0$ for $k \leq i < N$ and $\lambda_i - \mu_i < 0$ for $0 < i < k$, then there is a drift toward N if $i \geq k$, and toward 0 otherwise. In situations where such unstable drift-change pairs are possible, multiple runs are suggested, starting with a drift-change pair, with each run closely monitored. However, we will not consider this case further.

Stable drift-change pairs will be called *pairs of attraction*, or, if there is only one point *states of attraction*. For many models, starting a simulation with a member of a pair of attraction can be recommended because one can expect that a sign change in the drift will occur in the near future, giving the simulation some validity. Often, such starting states are also states close to the state with the highest probability or the highest frequency. For instance, in the $M/M/1$ queue, the pair of attraction is $(0,1)$, with 0 being the state with the highest probability, and 1 the state with the highest frequency.

Since it is usually easy to formulate flow models and find pairs of attraction, starting states from such pairs can be recommended. The fact that they are often close to the maximal equilibrium probability is a further advantage. The state variables of pairs of attraction normally have values that lie below their expected values. Consequently, choosing states with the state variable having slightly higher values than the ones in pairs of attraction may be advantageous in some cases, especially if multiple runs are planed.

4 EXAMPLES

In this section, we present a number of examples to highlight the different issues discussed earlier. In our tables and discussions, the derivative of MSE is always multiplied by T because otherwise, it is very small number, and representing small numbers in decimal form requires a lot of space.

First, we show, using the $M/M/1$ queue with $X(t) = R(t)$ representing the number of elements in the system as an example, that multiple runs can reduce $\mathrm{MSE}(\bar{R}(T))$. The arrival rate is $\lambda = 0.8$ and the service rate is $\mu = 1$. Since the pair of attraction is $(0,1)$, we start with $X(0) = 0$, and we use $T = 1000$. Since our programs cannot handle infinite-state queues, we restrict the number in the system to 50. For this problem, $T\mathrm{MSE}'(\bar{R}(T)) = 2.340$, indicating that the MSE is increasing when introducing a warm-up period. As one can see from Table 1, the MSE is smallest when 6 runs are made.

Table 1: MSE for $M/M/1$ queue, number of runs varying from 1 to 8, $\lambda = 0.8$, $\mu = 1$, $T = 1000$, $X(0) = 0$.

Runs	1	2	3	4
MSE	1.5691	1.3748	1.2154	1.1045
Runs	5	6	7	8
MSE	1.0412	1.0178	1.0259	1.0583

In the case of the $M/M/1/N$ queue, one may also be interested how the buffer size N and $\rho = \lambda/\mu$ affect the optimal number of runs. Hence, in Table 2, we do the calculation for a buffer size of $N = 10$ and $N = 50$, and for $\rho = 0.8$ and 0.9. Table 2 shows in column "Opt. runs" that more runs are optimal if the buffer size has the higher value, and that fewer runs are optimal if ρ has the higher value.

Table 2: Optimal number of runs for $M/M/1/N$ queue with varying N and λ when $\mu = 1$, $T = 1000$ and $X(0) = 0$.

		1 Run		Multiple Runs	
N	λ	MSE	$T\mathrm{MSE}'$	Opt. runs	MSE
10	0.8	0.171	0.180	3	0.168
50	0.8	1.569	2.340	6	1.018
10	0.9	0.228	0.124	2	0.227
50	0.9	14.856	16.770	3	9.555

To demonstrate that the derivative of MSE with respect to w is a good indicator for using multiple runs, consider the $M/M/20/50$ queue with $\lambda = 0.8$, $\mu = 1/c = 0.05$, and $T = 1000$. The derivatives of MSE in this case are negative for $X(0) < 9$ and $X(0) > 25$, and they are positive for $X(0)$ ranging from 9 to 25. Table 3 gives the MSE for the number of runs from 1 to 3, and the optimal number of runs (Runs^*) for $X(0)$ ranging from 6 to 10. As one can see, for both $X(0) = 7$ and $X(0) = 8$, the results of Table 3 suggest multiple runs, even though the derivative in question is negative. In such cases, multiple runs may be combined with warm-up periods. This, however, seems to be an exception. For $X(0) = 10$, three runs are optimal because the MSE for 4 runs is 2.077.

The next question we need to address is whether or not states obtained from pairs of attraction will lead to low values of the derivative of the MSE. To resolve this question, consider first the $M/M/20/50$ queue discussed earlier. In this case, $\lambda/\mu = 16$, that is, for $X(0) = 16$, there is no drift. Hence, this should be a good starting state. However, the starting state with

Table 3: MSE($\bar{\bar{R}}(T)$) for the $M/M/20/50$ queue with $\lambda = 0.8$, $\mu = 0.05$, and $T = 1000$.

$X(0)$	TMSE$'$	MSE			Runs*
		1 run	2 runs	3 runs	
6	-2.207	2.209	2.191	2.364	1
7	-1.184	2.200	2.147	2.260	2
8	-0.243	2.192	2.107	2.165	2
9	0.615	2.186	2.072	2.079	2
10	1.389	2.180	2.040	2.002	3

the lowest MSE when using only 1 run is $X(0) = 13$, with MSE $= 2.177$. For $X(0) = 16$, the MSE is 2.181, which is not much higher. However, when starting with $X(0) = 16$, and doing 12 runs, one finds an MSE of only 1.069, which is significantly lower. The best one can obtain with $X(0) = 13$ is 4 runs with an MSE of 1.746.

We now increase T by a factor of 10 form 1000 to 10000. In this case, the derivatives of of the MSE with respect to the warm-up period is positive for $8 \leq X(0) \leq 24$, and negative otherwise, which is very close to the result obtained for $T = 1000$. Also, the optimal number of runs does not change significantly for $X(0)$ between 6 to 10, as seen in Table 4. Also, the starting state with the lowest MSE is state number 13, as before. Hence, in this model, the best starting state, and the best number of runs is insensitive to T.

Table 4: MSE($\bar{\bar{R}}(T)$) for the $M/M/20/50$ queue with $\lambda = 0.8$, $\mu = 0.05$, and $T = 1000$.

$X(0)$	TMSE$'$	MSE			Runs*
		1 run	2 runs	3 runs	
6	-0.185	0.239	0.239	0.240	1
7	-0.081	0.239	0.238	0.239	2
8	0.015	0.239	0.238	0.238	2
9	0.102	0.239	0.238	0.238	3
10	0.180	0.239	0.237	0.237	3

As our final example, consider a three station tandem queue with stations 1, 2 and 3. Each station has a buffer size of 5, including the place for the part being served. All jobs arrive at station 1, and must proceed to station 2, then 3, in that order, after which they depart. If the first buffer is full, arrivals are lost. If the buffer of station 2 or 3 is full, the previous station is blocked, that is, departures are delayed until there is a place in the buffer they go to. Arrivals are Poisson with a rate $\lambda = 0.75$, and service is exponential, with a rate of $\mu_i = 1$ for station i, $i = 1,2,3$. We use $T = 1000$. Our $R(t)$ is given by the total number in the system. The number in station i will be denoted by $X_i(t)$, $i = 1,2,3$, and the state where $X_i(t) = x_i$, $i = 1,2,3$ will be denoted by $[x_1,x_2,x_3]$. With this notation, one pair of attraction is the state $([0,0,0],[1,0,0])$. Unfortunately, both states of this pair show that increasing the warm-up period from 0 to some positive value would decrease the MSE. Accordingly, using two runs for state $[0,0,0]$ would increase the MSE form originally 0.1710 to 0.1736. For state $[1,0,0]$, the corresponding values are 0.1705 and 0.1715. The state with the lowest MSE is state $[5,0,0]$, with an MSE of 0.1694. Note that in our model, $\mathrm{E}(R) = 5.44$, and for state $[5,0,0]$, $R(t) = 5$, which is close to $\mathrm{E}(R)$. When starting in this state, the MSE is reduced from 0.1715 to 0.1661 if two runs are made. Also, the derivative of the MSE is positive. In fact, out of the 216 states that can be used for starting the simulation, 153 states have a positive derivative of the MSE, and in 112 starting states, 2 runs are better than 1. There are thus better states to start the simulation in than $[0,0,0]$ or $[1,0,0]$. On the other hand, the potential improvement in the MSE by using states other than $[0,0,0]$ or $[1,0,0]$ is limited. However, since pairs of attraction are so easy to find, and since the possible improvements obtainable by using other starting states is small, using pairs of attraction is still no mistake. Improvements can also be made by increasing the state variables slightly. For instance, for state $[1,1,1]$, the MSE is 0.1702, its derivative is positive, that is, no warm-up period should be used, and if two runs are made, the MSE decreases to 0.1696.

5 CONCLUSIONS

In this paper, we demonstrated that multiple runs may be optimal if the simulation starts in a well-chosen state. One method we presented for finding well chosen starting states uses flow analysis to find pairs of attraction: The elements of these pairs typically yield good starting states. Sometimes, it is best not to use these starting states directly, but to change the values of the state variable to bring them closer to their equilibrium expectations. In particular, state variables representing queues and obtained from pairs of attraction typically have values below their equilibrium expectation. Once a good starting state is at hand, its effect would be watered down by any warm-up period. The beneficial effect of a well chosen starting state can be increased by making multiple runs. In this case, the effect of the bias is strengthened, which implies that starting states with $R(0)$ close to $\mathrm{E}(R)$ become increasingly advantageous.

We also presented a number of numerical experiments. Though in the models used, pairs of attraction did not necessarily provide the minimal mean squared error, they led to a reasonable performance at all times, especially after a slight increase of the state variables.

ACKNOWLEDGEMENTS

This research was partially supported by NSERC of Canada, Grant 8112.

REFERENCES

Alexopoulos, C. and Goldman, D. (2004). To batch or not to batch? *ACM Transaction on Modeling and Computer Simulation*, 14(4):76–214.

Asmussen, S. and Glynn, P. W. (2007). *Stochastic Simulation: Algorithms and Analysis*, volume 57 of *Stochastic Modelling and Applied Probability*. Springer Verlag, New York.

Banks, J., Carson, J. S., Nelson, B. L., and Nicol, D. (2005). *Discrete Event Simulation*. Prentice Hall, Englewood Cliffs, NJ.

Grassmann, W. K. (1982). Initial bias and estimation error in discrete event simulation. In Highland, H. G., Chao, Y. W., and Madrigal, O., editors, *Proceedings of the 1982 Winter Simulation Conference*, pages 377–384, Piscataway, NY. The Institute of Electrical and Electronics Engineers, Inc.

Grassmann, W. K. (1987). Means and variances of time averages in Markovian environments. *European Journal of Operationonal Research*, 31:132–139.

Grassmann, W. K. (2008). Warm-up periods in simulation can be detrimental. *Probabilitiy in Engineering and Informational Sciences*, 22:415–429.

Grassmann, W. K. (2011). Rethinking the initialization bias problem in steady-state discrete event simulation. In Jain, S., Creasey, R. R., Himmelspach, J., White, K. P., and Fu, M., editors, *Proceedings of the 2011 Winter Simulation Conference*, pages 593–599, Piscataway, NY. The Institute of Electrical and Electronics Engineers, Inc.

Grassmann, W. K. (2014). Factors affecting warm-up periods in discrete event simulation. *Simulation*, 90(1):11–23.

Kelton, W. D. (1989). Random initialization methods in simulation. *IIE Transactions*, 21(4):355–367.

Kelton, W. D. and Law, A. M. (1985). The transient behavior of the $M/M/c$ queue, with implication for steady-state simulation. *Operations Research*, 33:378–396.

Law, A. M. and Kelton, W. D. (2000). *Simulation Modelling and Analysis*. McGraw Hill, New York, third edition.

Madansky, A. (1976). Optimal conditions for a simulation problem. *Operations Research*, 24:572–577.

Newell, G. F. (1982). *Applications of Queueing Theory*. Chapman and Hall, London, second edition.

Pasupathy, R. and Schmeiser, B. (2010). The initial transient in steady state point estimation: Context, a biography, the MSE criterium, and the MSER statistic. In Johansson, B., Jain, S., Montaya-Torres, J., Hugan, J., and Yücesan, E., editors, *Proceedings of the 2010 Winter Simulation Conference*, pages 184–197, Piscataway, NY. The Institute of Electrical and Electronics Engineers, Inc.

Pawlikowski, K. (1990). Steady-state simulation of queueing processes: a survey of problems and solutions. *ACM Computing Surveys*, 22(2):123–170.

Tocher, K. D. (1963). *The Art of Simulation*. English University Press, London.

Whitt, W. (1991). The efficiency of one long run versus independent replications in steady-state simulation. *Management Science*, 37(6):645–666.

Wilson, J. R. and Prisker, A. A. B. (1978). Evaluation of startup policies in simulation experiments. *Simulation*, 31:79–88.