Simulation Modelling Practice and Theory 71 (2017) 102–113

Contents lists available at ScienceDirect

Simulation Modelling Practice and Theory

journal homepage: www.elsevier.com/locate/simpat

Output analysis for terminating simulations with partial observability

Bogumił Kamiński*, Grzegorz Koloch

Warsaw School of Economics, Al. Niepodległości 162, 02–554 Warszawa, Poland

ARTICLE INFO

Article history:
Received 19 March 2016
Revised 1 December 2016
Accepted 2 December 2016
Available online 7 December 2016

Keywords:
Simulation
Output analysis
Partial observability

ABSTRACT

In this paper we propose methods of output analysis for terminating simulations when the researcher is unable to fully specify the initial state of the simulation using observations of the real system. We call this situation “partial observability” and argue that it is common in practice, especially in the case of complex agent-based simulations. We provide classification of situations where not all input parameters or simulation state variables are observable and for each case we propose a method of terminating simulation output analysis. In particular, we focus on the situation where a rare event needs to be analyzed, since it requires careful design of a simulation experiment in order to minimize the computational budget and avoid bias in the output predictions.

1. Introduction

Complex real-world phenomena that are dynamic and include random behavior are often modeled using stochastic simulations [20,24]. Two mainstream approaches to analysis of such models concentrate on performance of the system at a given point in time or in a given state (*terminating* simulation output analysis) or over the long-run (*non-terminating* or *steady state* simulation output analysis), see. e.g. Banks et al. [6]. In this text we concentrate on the terminating case. This case is usually considered less challenging statistically than steady state analysis but, as Law [18] emphasizes, it is no less important because practitioners are often interested in the behavior of the modeled system under defined initial conditions.

In a terminating simulation output analysis scenario we want to obtain an evaluation of the simulation output given some well specified initial conditions and a terminating event condition [1,23]. The key challenge of such an analysis is proper treatment of the initial state of the simulation. Many standard references, e.g. Alexopoulos and Seila [2] do not concentrate on this issue, assuming that the initial state is fully specified. In the texts which discuss the challenges with specifying the initial state of the simulation, two standard recommendations are considered. The first of these is to collect the appropriate empirical data that allows the initial state to be determined and the second one is to use a warmup period [8,20].

In this text we analyze the situation where we have a simulator S of the real system. We have knowledge about certain characteristics of the real system at time t and want to predict a value of one (or more) of its characteristics at some future moment $t + \Delta t$, where Δt is either deterministic or determined by some random terminating event condition. If knowledge of the real system is sufficient to fully specify the initial state of the simulator S at time t then we can follow the standard methods recommended for terminating simulation output analysis [18] to evaluate the output of the simulation at time

* Corresponding author.
E-mail address: bkamins@sgh.waw.pl (B. Kamiński).

http://dx.doi.org/10.1016/j.simpat.2016.12.002
1569-190X/

$t + \Delta t$. It is worth highlighting that knowledge of the characteristics of a real system at time t can mean either that the researcher knows deterministic values defining the state or their distribution – both cases are covered by standard methods. However, very often the analyst is unable to collect all the data in the real system that is needed to specify the initial state of simulation S.

In this text we propose methods allowing formulation of simulation output predictions at time $t + \Delta t$ given that knowledge of the characteristics of the real system is insufficient to fully specify the initial state of simulation S at time t. In some cases, as we show in Section 3, it is possible to make such predictions for any type of simulation. However, if we lack information about the dynamic characteristics of the simulation, some additional information is required to make predictions.

In this text we make an additional assumption that that the simulation is non-terminating and stable. By this we mean that the simulation can be run for an arbitrarily long time and has a steady state distribution, see [5]. A basic example of such a simulation model is a M/M/1 queue with the arrival rate less than the service rate. We wish to make a prediction assuming that the system at time t has been already running for a long time and it is possible to assume that the *unconditional* distribution of its state at time t follows the steady state distribution (at least approximately). In our M/M/1 queue example this means that we choose such t, that the distribution of number of customers in the system at this time does not depend on the number of customers in the system when the simulation is started.

Using the assumption that the considered simulation at time t is unconditionally in steady state we are able to reason about the distribution of the unknown part of the state at time t *conditional* on the partial information about the specification of the state at time t which the researcher knows. Later in this section we present two detailed examples of such situations.

The above assumption, in particular, implies that the proposed methodology is not directly applicable to simulations that either do not have steady state or where it is not reasonable to assume that they can reach it in practice, e.g. simulation of one day of operations of a warehouse that only dispatches products during the day so the state of the stock only decreases.

Formally, in terminating simulation output analysis, we measure the output value y at some moment T_E given the simulation initial state C at some earlier time t. It is assumed that C and T_E are well specified random variables [18]. In consequence it is typically assumed [19] that repeated runs of the simulation generate outputs y that are independent and identically distributed. This reasoning is based on the assumption that it is possible to specify C. We argue that in practice this is often difficult. Following the terminology from control theory [14] we maintain that the state of simulation is *observable* at time t if it is possible to determine it using information collected from the real system up to time t. In a simulation modeling context we allow for the state to be identified deterministically or as a probability distribution.

If the state of simulation is not observable we specify that it is *partially observable*, that is we assume that not all state variables are observable, i.e. for some of the variables defining the state of the simulation it is not possible to identify their value using available information about the real system. Examples of such situations are as follows (in these examples we use standard notation and terminology used in queuing theory, see. e.g. [20]).

Example 1. Consider a G/G/1 queuing system with known arrival and service time distributions. Assume that we are able to measure (in the real system) the number of customers in the system N_t at time t and we seek to predict the number of customers $N_{t+\Delta t}$ in the system at some well specified future time $t + \Delta t$. If the system were memoryless (e.g. M/M/1 queue) then N_t, along with information about arrival and service rates, would constitute a full specification of the simulation state at time t. Thus the parameter N_t is observable and we have no state variables that are unobservable. However, in the case of a G/G/1 queue, full specification of the state would also include the time which elapsed since the last arrival of the customer and the time the current service process has taken so far. Given that we are able to measure N_t alone in the G/G/1 system, the state of the simulation is partially observable.

In this text we propose a method that allows the prediction of $N_{t+\Delta t}$ given knowledge of N_t alone, conditional upon the queuing system being stable and that in time t the simulation state distribution is unconditionally (without conditioning on N_t) a steady state distribution, at least approximately. In other words, when t is sufficiently large that the initial conditions, which are set when the system started its operations, do not influence the distribution of the system's state at time t.

Example 2. The system we want to investigate is an M/M/1 queue with two types of customers A and B arriving with equal probability. Assume that at a given time t we can measure how many customers of each type are in the real system but not in which sequence they have entered it. By A_t denote the number of customers of type A in the system at time t and analogously define B_t. We are interested in prediction of $A_{t+\Delta t}$ given that we know A_t and B_t. Observe, that in such a case, the state of simulation at time t can be represented as some sequence of customers of type A and B along with information about arrival and service rates. We assume that we do not know this sequence. The measured values A_t and B_t, however, as opposed to N_t in Example 1 are not part of the state of the simulation, yet they can be computed given that we knew this state (the sequence of As and Bs). Therefore only the simulation output, i.e. A_t and B_t, is measurable and the simulation state is not observable. Even though no state variables are observed we still designate this situation as partially observable since we can use knowledge of the simulation output to draw guidance about its state as they are related (though the mapping from outputs to state is not injective).

In this paper we propose a method for predicting $A_{t+\Delta t}$ given that we know that $A_t = \alpha$ and $B_t = \beta$ (again assuming that the queuing system is stable and has been already running for a long time at time t).