*Proceedings of the 2007 Winter Simulation Conference*
*S. G. Henderson, B. Biller, M.-H. Hsieh, J. Shortle, J. D. Tew, and R. R. Barton, eds.*

# ANT-BASED APPROACH FOR DETERMINING THE CHANGE OF MEASURE IN IMPORTANCE SAMPLING

Poul E. Heegaard

Department of Telematics
Norwegian University of Science and Technology
Trondheim, N-7491, NORWAY

Werner Sandmann

Dep. Information Systems & Applied Computer Science
University of Bamberg, Feldkirchenstrasse 21
Bamberg, D-96045, GERMANY

## ABSTRACT

Importance Sampling is a potentially powerful variance reduction technique to speed up simulations where the objective depends on the occurrence of rare events. However, it is crucial to find a change of the underlying probability measure yielding estimators with significantly reduced variance compared to direct estimators. In this paper, we present a new dynamic and adaptive method for this purpose. The method is inspired by ant-based systems that are in widespread use for solving optimization problems. No intimate knowledge of the model under consideration is necessary. Instead, the method adapts to it. Different commonly used modeling paradigms such as queueing and reliability models, amongst many others, are supported by describing the new method in terms of a transition class formalism. Simulation results demonstrate the accuracy of the obtained estimates, and details of the adapted change of measure are investigated to gain insights into the inner workings of the method.

## 1 INTRODUCTION

System performance and reliability are important topics in a variety of different domains. Both performance and reliability of many systems are heavily influenced by rare events, which occur with very small probabilities but may have serious consequences. Several examples elucidate the necessity to analyze such rare events, ruins in insurance risk or finance, breakdowns of manufacturing systems, technical defects, false alarms in radar or security systems, to mention but a few. In particular in the area of computer and communication systems there are a lot of examples for the impact of rare events to system performance, such as packet losses in switched networks, bit errors in digital communications, or system failures in highly reliable systems.

Analytically, asymptotic analyses of exponentially rare events can sometimes be done using large deviations theory (Dembo and Zeitouni 1998, Shwartz and Weiss 1995) which is an elaborate, advanced and extremely technical mathematical theory but unfortunately in its applicability limited to relatively small models. Hence, analytical performance evaluation of complex systems is usually not possible at all, a fortiori in the presence of rare events. Likewise, numerical methods are typically not suitable for rare event analysis. Hence, simulation techniques are highly desirable. Unfortunately, direct simulation of rare events is not effective, since rare events occur too infrequently in simulations to compute reliable statistical estimates in reasonable time. Simulation speed-up is necessary in the sense that simulation time to get estimates with desired accuracy, for example confidence interval relative half width, must be reduced. Since the accuracy depends on the variance of the simulation estimators, such a simulation speed-up corresponds to variance reduction, and it turns out that Importance Sampling is well suited for this purpose. The basic idea of is to provoke more of the rare events of interest by changing the underlying probability measure. The systematically biased results are then appropriately weighted to provide unbiased estimates. Although theoretically an optimal zero-variance estimator always exists, it cannot be used in simulations since it explicitly depends on the unknown quantity to be estimated. Therefore, the crucial point in a successful application is to find a probability measure which leads to estimators with much smaller variance than the direct simulation estimators. Consequently, the main part of the literature is concerned with the change of measure and verifying the efficiency and the robustness of the resulting estimators. Typically, efficiency criteria and robustness properties are focused on the asymptotic behavior of estimators as the probability of the rare event approaches zero. The most prominent of such asymptotic properties are asymptotic efficiency and bounded relative error but many others are reasonable; see, e.g., Sandmann (2007), Blanchet et al. (2007) for recent investigations on estimators' asymptotics. However, in practice there are at least two drawbacks with asymptotic properties. Proving them is extremely difficult and turns out to be only possible for rather small or restricted models.

1-4244-1306-0/07/$25.00 ©2007 IEEE

Besides, in practice one is usually not really interested in asymptotically small probabilities such as $10^{-100}$ or less but in probabilities in the range of say $10^{-9}$ to $10^{-12}$. Hence, investigating asymptotic properties is of course useful and important from a theoretical point of view to get insights to the method but in practice we may be satisfied by estimators with significantly reduced variance meaning great simulation speed-up even without provable asymptotic properties. From the practitioners point of view it is more important to come up with an effective simulation speed-up for realistic models than to prove certain mathematical properties. It is required to have methods that can be used not only by the very experts in rare event simulation but by a broad range of users without intimate knowledge of the system and the analysis method.

In this paper we describe a new method for the change of measure in Importance Sampling that adapts to the given model. The method is inspired by ant-based systems that are in widespread use for optimization problems as for example network routing, resource allocation, and logistics. In Section 2 the considered type of models is described and a flexible transition class formalism is introduced which covers different model classes and domains such as queueing networks, highly reliable fault-tolerant systems, resource allocation models and many more. The Importance Sampling fundamentals are given in Section 3 which also includes the details of the new ant-based method. Simulation results are presented and discussed in Section 4. Finally, Section 5 concludes the paper and outlines further research directions.

## 2 MODEL DESCRIPTION

Let us start with the description of the structure of models we consider in this paper. As for the state space definition, in general we allow $d$-dimensional discrete-state models, that is the state space $\mathscr{S}$ is a subset of $\mathbb{N}^d$ which may be finite or infinite. More specifically, the state space is $\mathscr{S} = \{0,\ldots,v_1\}\times\cdots\times\{0,\ldots,v_d\}$ where $v_1,\ldots,v_d \in \mathbb{N}$ may or may not be finite. Any transition may affect at most two of the components $i, j \in \{1,\ldots,d\}$ of the state $x = (x_1,\ldots,x_d)$ in such a way that the according component values $x_i$ and $x_j$ are either decreased or increased by one. Although this may appear quite limited at a first glance, in fact a wide range of models is covered such as queueing networks, reliability models, resource allocation models, amongst many others. As we are concerned with Markovian models it would be possible to describe each specific model via an initial probability distribution and a generator matrix. However, in addition to the problem of rare events, models usually tend to be very large. The size of the state space typically increases exponentially with the number of system components or, in other words, the model dimensionality. This effect is known as state space explosion and makes the model intractable by means of numerical solution approaches. One major advantage of simulation is that it does not suffer from state space explosion since the state space need not be explicitly enumerated. Thus, a model description that better reflects the event system character of the model is well suited, in particular for simulation purposes. Diverse specifications of such event systems can be found in the literature. Here, we adopt the transition class formalism as it previously appeared in Sandmann (2004).

### 2.1 Transition Class Formalism

In order to construct an appropriate Markovian event system we have to define its state space and to specify all events that are possible where events correspond to transitions from one to another state. To provide a formal description, it is necessary to define under which conditions a certain transition may occur, how it affects the system state and at which rate it occurs. We do this by classifying the possible transitions according to their effects which yields what we call transition classes. A transition class is a triplet $\tau = (\mathscr{U}, u, \alpha)$ where $\mathscr{U} \subseteq \mathbb{N}^d$ is the source state space containing all states in which the according transition is possible, $u : \mathscr{U} \to \mathbb{N}^d$ is the destination state function giving the state $u(x) \in \mathbb{N}^d$ that is entered when the according transition occurs in state $x \in \mathscr{U}$, and $\alpha : \mathscr{U} \to \mathbb{R}$ is the transition rate function giving the rate $\alpha(x) \in \mathbb{R}$ at which the according transition occurs in state $x \in \mathscr{U}$. A set of such transition classes then completely describes a model.

To define the general model class under consideration in terms of transition classes in a unified way without requiring excessive case differentiations we additionally introduce a pseudo-component 0 with according component entry $x_0$ that does not belong to the system state $x = (x_1,\ldots,x_d)$ but can be interpreted as the outside of a system. As will soon become clear, for ease of the unified definition it is convenient to set $x_0 < \infty$ and $v_0 = \infty$. Now, we are prepared to express our previously described general model class in a very concise formal manner using transition classes.

Obviously, since we allow two state components to be changed by one transition, given a $d$-dimensional state space and our additional pseudo-component, we need $d \cdot (d+1)$ transition classes. Let $i, j \in \{0,\ldots,d\}, i \neq j$. That means all possible combinations of two different state components including the pseudo-component are covered by according choices of $i$ and $j$ and uniquely define the transition classes $\tau_{(i,j)} = (\mathscr{U}_{(i,j)}, u_{(i,j)}, \alpha_{(i,j)})$. The source state spaces are $\mathscr{U}_{(i,j)} = \{(x_1,\ldots,x_d) : x_i > 0 \ \wedge \ x_j < v_j\}$ and the destination state functions are defined by $u_{(i,j)}(x_1,\ldots,x_d) = (x_1,\ldots,x_d) - \mathbf{1}_i \cdot \chi(x_i > 0) + \mathbf{1}_j \cdot \chi(x_j < v_j)$ where for any $n \in \mathbb{N}$, $\mathbf{1}_n$ denotes the $d$-dimensional vector with entry 1 at component $n$ and zero-entries at all other components, and $\chi(\cdot)$ is the characteristic function, i.e. it equals one if the logical expression in its argument is true and it equals zero otherwise. Finally, any function $\alpha_{(i,j)} : \mathscr{U}_{(i,j)} \to \mathbb{R}$ is allowed as transition rate function.

### 2.2 Covered Model Classes

To give an impression of the generality of our model description we briefly outline how to model queueing networks and highly reliable systems. Many more model classes can be expressed in a similar fashion.

First consider a $d$-node queueing network. The state component entries then correspond to the number of customers in the according network node, and within the introduced formalism customers may arrive at any network node, may move from any node to any other node and may leave the system from any node. More precisely, if a customer moves from node $i$ to node $j$ then transition class $\tau_{(i,j)}$ applies. External arrivals as well as departures from the system are modeled via the pseudo-component where transition class $\tau_{(0,j)}$ models external arrivals to network node $j$ and transition class $\tau_{(i,0)}$ models departures from node $i$. In case of bounded buffer capacities these are covered by finite values of the $\nu_i$. Since no restrictions are made for the transition rate functions, all arrival and service rates may be state dependent. In the special case of state independent arrival or service rates the according state transition functions are constant.

Highly reliable systems with $d$ types of system components that are subject to failures and repairs can be also conveniently modeled. In this case any state component entry corresponds to the number of failed components of the according type, a failure of component $i$ is modeled by transition class $\tau_{(0,i)}$ and a repair of component $i$ by transition class $\tau_{(i,0)}$. One may equivalently consider the number of operating (non-failed) system components as the state component entries but this is just a matter of taste. In any case, finite values of $\nu_i$ mean that exactly $\nu_i$ components (operating or failed) of type $i$ are present in the system. Again, both failure rates and repair rates may depend on the system state and state independent rates are covered by constant transition rate functions.

### 2.3 Example Model

As mentioned in the previous section the model description is quite general. Consider as an example a single queue fed with $d$ different user types (customer classes) as illustrated in Figure 1. There may be either finite or infinite populations of each user type $i$, that is the sources $M_i$ feeding the system with users of type $i$ may be finite or infinite. Arrival rates $\lambda_i(x)$ as well as service times $1/\mu_i(x)$ can depend on the system state $x$, and there might be multiple servers. The system capacity $N$ is the sum of the number of servers and queueing positions. Formally, this is embedded in the general model description in terms of the transition class formalism where the transition rate functions are $\alpha_{(i,j)}(x) = 0$ for $i > 0 \wedge j > 0$. With regard to the notation for the queueing system example, we then get for the remaining transition rate functions $\alpha_{(0,i)}(x) = \lambda_i(x)$ and $\alpha_{(i,0)}(x) = \mu_i(x)$.

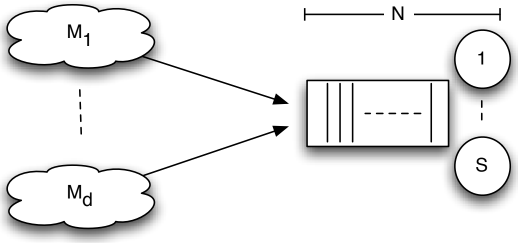

Figure 1: Example system

In this paper we will focus on the rare event analysis of two-dimensional models where transitions may occur only in horizontal and vertical direction as depicted in Figure 2. Obviously, the previously described queueing example (now with $d = 2$ customer classes) fits to this type of models. For such a model we then only need four transition classes, namely exactly those where either $i$ or $j$ equals zero. That is the source state spaces and the according destination state functions are

$$\begin{aligned}
\mathscr{U}_{(0,1)} &= \{(x_1,x_2) : x_1 < \nu_1\}, u_{(0,1)}(x_1,x_2) = (x_1+1,x_2),\\
\mathscr{U}_{(0,2)} &= \{(x_1,x_2) : x_2 < \nu_2\}, u_{(0,2)}(x_1,x_2) = (x_1,x_2+1),\\
\mathscr{U}_{(1,0)} &= \{(x_1,x_2) : x_1 > 0\}, u_{(1,0)}(x_1,x_2) = (x_1-1,x_2),\\
\mathscr{U}_{(2,0)} &= \{(x_1,x_2) : x_2 > 0\}, u_{(2,0)}(x_1,x_2) = (x_1,x_2-1).
\end{aligned}$$

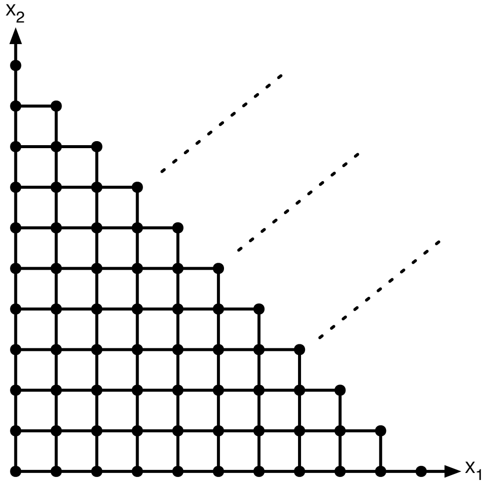

Figure 2: Example model structure

Of course, neither the transition class formalism nor our new ant-based method for determining the change of measure in Importance Sampling are limited to two dimensions. Rather, the choice of two-dimensional models renders possible to illustratively visualize the way our method works.

## 3 ANT-BASED IMPORTANCE SAMPLING

In this section a dynamic adaptive Importance Sampling method is introduced. A major novelty is the distributed search algorithm that is applied to obtain the weighting of the transition rate functions and thus the change of measure for the Importance Sampling simulation. The weights are obtained by an ant-based system that provides a gradual change in the rate functions of the transitions that are in the sampled paths. This change in the rate functions yields the change of measure required for the Importance Sampling method used for speeding up the simulation of rare events.

In the following we describe the Importance Sampling principles and briefly review other adaptive and dynamic approaches for obtaining the change of measure, before the ant-based approach is described.

### 3.1 Importance Sampling

Importance Sampling is a variance reduction technique that makes use of a change of measure. The original system is simulated under a different probability measure, and the systematically biased results are weighted by a correcting factor, the likelihood ratio, to yield unbiased estimates.

Importance Sampling is already known for a long time, originally developed in the early 1940s in the context of static multi-dimensional Monte Carlo integration arising in problems in nuclear physics, see Hammersley and Handscomb (1964) for early applications. Since more than two decades it has been recognized that it is also a potentially powerful technique for rare event simulation in stochastic models of dynamic systems. The link between Importance Sampling and large deviations theory and thus the connection between Importance Sampling and rare events was first established in Cottrell, Fort, and Malgouyres (1983). In Glynn and Iglehart (1989) the framework for stochastic processes including generalized semi-Markov processes and Markov processes has been given which thus constitutes the particular formal basis for the types of models we consider in this paper. Since then a variety of applications in systems performance evaluation such as reliability models and queueing networks have appeared; see, e.g., Heidelberger (1995), Juneja and Shahabuddin (2006) for surveys.

#### 3.1.1 Formal Basis

In a general measure theoretic setting, Importance Sampling is based on the Radon-Nikodym theorem, and all applications of Importance Sampling to specific model types and domains can be derived from this setting. Consider two probability measures $P$ and $P^*$ on a measurable space $(\Omega, \mathscr{A})$, where $P$ is absolutely continuous with respect to $P^*$, which means that for all $A \in \mathscr{A}$, $P^*(A) = 0 \ \Rightarrow\ P(A) = 0$. Then, the Radon-Nikodym theorem guarantees that the Radon-Nikodym derivative $L = dP/dP^*$ exists and that

$$\forall A \in \mathscr{A} :\ P(A) = \int_A L(\omega) dP^*.$$

In the context of Importance Sampling the probability measure $P^*$ is called the *Importance Sampling measure*, and $L$ is referred to as the *likelihood ratio*. The basic property exploited by Importance Sampling is that expectations with respect to $P$ are identical to expectations with respect to $P^*$ when weighting by the likelihood ratio. Let $L$ be a version of the likelihood ratio and $Y$ a random variable on $(\Omega, \mathscr{A})$. Then

$$E_P[Y] = \int Y(\omega) dP = \int Y(\omega) L(\omega) dP^* = E_{P^*}[YL]$$

where $E_P$ and $E_{P^*}$ denote expectations with respect to the probability measures $P$ and $P^*$, respectively. Hence, when a simulation is performed under $P^*$ and then weighted by $L$ the result is unbiased. Using a different density or probability distribution/measure is called a *change of measure*, and it is the essential part and the art of Importance Sampling to perform this change of measure such that estimators can be achieved that have significantly reduced variances compared to direct simulation estimators. In fact, it is by no means guaranteed that Importance Sampling always results in reduced variances but it may even yield an infinite variance increase. Hence, to speak of on art in choosing the change of measure seems reasonable. However, to be useful for a broad range of applications and to be used by non-expert users it is necessary to come up with methods that yield Importance Sampling simulations with significantly reduced variance.

In Markovian models the probability measures $P$ and $P^*$ correspond to path probabilities or densities. These are given by the product of transition probabilities in the discrete-time case or transition rates and according exponential holding times in the continuous-time case; see, e.g., Glynn and Iglehart (1989), Heidelberger (1995), Sandmann (2005), Juneja and Shahabuddin (2006) for full formal descriptions in these settings and Sandmann (2004) for according descriptions in terms of the transition class formalism. In the latter case, rates are expressed in terms of rate functions (similarly transition probability functions can be introduced in the discrete-time case) which thus determine the path densities. The condition of absolute continuity corresponds to the condition that all paths that are possible in the original model must remain possible under Importance Sampling (which can be restricted to paths that include the rare event of interest). This can be obviously achieved by the requirement that all originally positive rates remain positive after the change of measure. To state it concisely, the essential point is to assign new rates to any transition which in terms of the transition class formalism means to determine appropriate transition rate functions $\alpha^*_{(i,j)}$ for Importance Sampling. Of course, these may (and usually do) depend on the system state meaning that they are not constant.

### 3.1.2 Adaptive Approaches

Adaptive approaches aim at learning a good change of measure and are thus potentially well suited for use without intimate knowledge of the model at hand. The general strategy of adaptive Importance Sampling is to start with some initial change of measure, perform a couple of independent simulation runs using this change of measure, and update the change of measure according to some rules that depend on and hence characterize the specific adaptive method. Then the simulation is continued by making multiple independent simulation runs with the updated Importance Sampling measure and so on until finally the method converges to an Importance Sampling measure with which the actual simulation is performed.

Several adaptive schemes have been proposed in the context of estimating small bit error rates in digital communications; see Stadler and Roy (1993), Srinivasan (2002). Queueing models were addressed in a series of papers where the common idea was to utilize stochastic optimization techniques to adaptively minimize the sample variance of the estimator. For simulating buffer overflows in single server queues a random search method and mean field annealing in combination with a dynamic regenerative simulation were applied in Devetsikiotis and Townsend (1993a) and Devetsikiotis and Townsend (1993b), respectively. Another approach uses stochastic gradient methods and was applied to single server queues in Devetsikiotis et al. (1993) and extended to tandem queues in Freebersyser et al. (1996). However, all these methods require quite large computational efforts and therefore only apply to rather limited models. While the aforementioned approaches aim at directly minimizing the (estimated) variance, the cross entropy method (Rubinstein 1997, Rubinstein and Kroese 2004) aims at minimizing the cross entropy between the currently used measure and the (unknown) optimal measure. Unfortunately, applied to Markovian models as first done in de Boer (2000) in the context of queueing networks, due to huge storage requirements the method is limited to quite small models when considering general change of measure strategies. Larger models can be simulated by restriction to state independent change of measure but it appears that this works quite well for some systems but not for others. It is shown in de Boer (2006) that state independent rates cannot yield asymptotically optimal estimators even in the simple case of the two-node tandem queue. For more general models, it is not clear if such an application of Importance Sampling can provide variance reduction at all.

Other promising approaches to adaptive Importance Sampling are dynamic. In the approaches described above the change of measure is improved from one iteration to the next where each iteration step consists of possibly multiple simulation runs. In dynamic approaches the Importance Sampling measure can also be changed within any simulation run. In the mentioned dynamic regenerative simulations this change only occurs when the set of rare states of interest has been visited some times whereas other approaches allow for such a change after any simulated transition. In Carrasco (1992) the notion of failure distances was introduced in the context of Markovian fault-tolerant systems where the failure distance is a metric for the distance from any state to a target set of rare states. All transition rates or probabilities are then changed after each simulated transition. The major drawback of this method is that the simulation speed-up strongly depends on an accurate computation of the failure distances which involves the computation of minimum cut sets, in general an NP-hard problem. Hence, the applicability of this approach is quite limited. The basic idea of the dynamic change of measure introduced in Heegaard (1998a), Heegaard (1998b) is to change the transition rate functions in accordance to the importance of a path with this transition as the first step. The target importance is determined by a “lookahead” approach inspired by the failure distance measures in Carrasco (1992). The path likelihood is estimated by determining the most likely path from the current state to any state in the set of rare states of interest.

## 3.2 Ant-Based Adaptive Approach

Although a purely analytical approach to rare event analysis is usually impossible for realistic models of complex systems, large deviations theory gives valuable insights and guidelines. Rare events typically occur on certain most likely paths and for Importance Sampling it is known that the change of measure should mainly emphasize these paths. While it is difficult to analytically determine the most likely paths via large deviations theory they can be estimated by e.g. the lookahead approach applied in Heegaard (1998b). However, the computational demand of the lookahead approach increases as the dimensionality of the state space increases.

In our new approach the target importance is instead determined by an ant-based search and update procedure. In Dorigo, Maniezzo, and Colorni (1996) a multi-agent system (Ant Colony Optimization) inspired by the behavior of ants was introduced to solve combinatorial optimization problems. In Schoonderwoerd et al. (1997) a similar system is designed to solve problems in telecommunication networks. The ants search iteratively for paths in a connected graph (a network) between source nodes and destination nodes.

The path quality is evaluated on arrival to a destination node and then each ant backtracks over the links along the reverse path back to the source node leaving *pheromones* to guide future ants in their search for the same destination. The better the path, the stronger the pheromone updates. We apply a similar approach here to gradually let the Importance Sampling change of measure adapt to the current model. The nodes are now states, and the links are state transitions, the source nodes are the origin or regenerative states, and the destination nodes are given as the set $\mathscr{R}$ of rare events of interest.

Let $x = (x_1, \cdots, x_d)$ be the state vector and $\pi(x,y)$ a path between state $x$ and $y$. The path between state $x$ and a state in the rare event set is denoted $\pi(x,\mathscr{R})$ and the $r$-th sampled path is $\pi^{(r)}$. The probability of a path $\pi$ is $p(\pi)$. The maximum normalised probability of a path from state $x$ to a state in $\mathscr{R}$, given that a transition according to $\tau_{(i,j)}$ occurred in state $x$ is

$$p_{\max}(\pi(x,\mathscr{R}) \mid \tau_{(i,j)}) = \frac{\max_r\{p(\pi^{(r)}(x,\mathscr{R}) \mid \tau_{(i,j)})\}}{\sum_{k=1}^{d} \max_r\{p(\pi^{(r)}(x,\mathscr{R}) \mid \tau_{(i,k)})\}} \tag{1}$$

where for any path $\pi(x,\mathscr{R})$ from $x$ to the set of rare states, $p(\pi(x,\mathscr{R}) \mid \tau_{(i,j)})$ denotes the probability of that path given that a transition according to $\tau_{(i,j)}$ occurred in state $x$. Note that we only sample over and update the *visited paths*, and not over all possible paths, which is a significant saving both in terms of computation and storage demands. An alternative to the max-function in (1) is the sum of all visited paths that end up in $\mathscr{R}$. The max-function is expected to give a quicker convergence but not necessary to the most likely of the rare events. The sum is a viable alternative that will be explored in further work when investigating the method on a broader set of examples. The ant-based procedure is described in Algorithm 1.

**Algorithm 1** Simulation procedure

```
repeat
    Sample a path $\pi^{(r)}$ from $x$ towards a target state in $\mathscr{R}$
    if hitting the origin states; then
        STOP
    else if $\pi^{(r)}$ contains states of the set $\mathscr{R}$ then
        for each state $y \in \pi^{(r)}$ update the maximum path probability $p_{\max}(\pi(y,\mathscr{R}) \mid \tau_{(i,j)})$ according to (1);
    end if
until end of simulation condition
```

The random search of the ants in every state is governed by a *random proportional rule* that is incrementally updated for every new path found by the ants. This proportional rule is determined by the normalized Importance Sampling transition rate functions $\alpha^*_{(i,j)}(x)$ in each state. Each sampled path that includes visits to $\mathscr{R}$ invokes a recalculation of the maximum path likelihood for all states in the path. Then the transition rate functions are changed for all transitions in the path according to the following updating rule for the change of measure:

$$\begin{aligned}
\alpha^*_{(0,j)}(x) &= \alpha_{(0,j)}(x) + p_{\max}(\pi(x,\mathscr{R}) \mid \tau_{(0,j)}) \cdot \Delta_{(0,j)}(x) \\
\alpha^*_{(j,0)}(u_{(0,j)}(x)) &= \alpha_{(j,0)}(u_{(0,j)}(x)) \\
&\qquad - p_{\max}(\pi(x,\mathscr{R}) \mid \tau_{(0,j)}) \cdot \Delta_{(0,j)}(x)
\end{aligned}$$

where

$$\Delta_{(0,j)}(x) = \alpha_{(j,0)}(u_{(0,j)}(x)) - \alpha_{(0,j)}(x)$$

provided that $\alpha_{(j,0)}(u_{(0,j)}(x)) > \alpha_{(0,j)}(x)$. Otherwise $\Delta_{(0,j)}(x) = 0$, which means that the transition rate functions are not changed. Note from Section 2.3 that in this paper we consider $\alpha_{(i,j)}(x) = 0$ and hence $\alpha^*_{(i,j)}(x) = 0$ for $i > 0 \ \wedge \ j > 0$. This will be generalised in further work.

Furthermore, note that only transition rate functions *along the sampled path* are updated. This strongly reduces the computational and storage demands compared to other approaches. All transitions in the path need only to store the (changed) transition rate, $\alpha^*$ (in direct simulation we need to store or retrieve the original rate for each transition), and one $p_{\max}$ for each target state space $\mathscr{R}$, typically one in a rare event simulation setup. The values accumulated and stored in $\pi^{(r)}$ for one sample $r$ are deleted when the transition rates have been updated. Initially, when no information of the target likelihood exists, the ants search the state space by a *guided random walk*, i.e. all the $\alpha^*$ in one node are equal and the next transition is chosen according to a uniform distribution. We call it *guided* because we will not allow transitions back to the state where we came from and use the same transition twice. After the initial phase the ants switch to the updating rule in (2).

## 4 NUMERICAL EXAMPLES

In this section we present numerical results for estimated means and their accuracy, as obtained by the ant-based Importance Sampling simulation applied to different variants of the system example from Section 2.3. The (statistical) accuracy will be given in terms of relative errors of the estimates and checked by comparison to exact values that are available for our example model. Furthermore, we illustrate the change of measure adapted by our method. As mentioned before, one major reason that we have chosen a two-dimensional model is that in this case the change of measure can be nicely visualized as done in Section 4.2.

### 4.1 Estimated Mean and Accuracy

We performed excessive simulation studies for many different models. To demonstrate and illustrate the behavior and efficiency of the method we present results for several parameters settings of the two-dimensional example model as introduced in Section 2.3. Each simulation experiment

Table 1: Parameter and rare event state space

| Case | $\alpha_{(0,1)}(x)$ | $\alpha_{(0,2)}(x)$ | $\alpha_{(1,0)}(x)$ | $\alpha_{(2,0)}(x)$ | $\mathscr{R}$ |
|---|---|---|---|---|---|
| I | 0.1 | 0.1 | 0.9 | 0.9 | $x_1 + x_2 = 10$ |
| II | 0.1 | 0.08 | 0.9 | 0.92 | $x_1 + x_2 = 10$ |
| III | 0.1 | 0.01 | 0.9 | 0.99 | $x_1 + x_2 = 10$ |
| IV | 0.1 | 0.1 | 0.9 | 0.9 | $x_1 = 5, x_2 = 5$ |
| V | 0.1 | 0.1 | 0.9 | 0.9 | $x_1 = 7, x_2 = 3$ |
| VI | 0.1 | 0.08 | 0.9 | 0.92 | $x_1 = 7, x_2 = 3$ |
| VII | 0.1 | $0.01(10 - x_2)$ | 0.9 | 0.9 | $x_1 = 7, x_2 = 3$ |
| VIII | 0.1 | $0.01(10 - x_2)$ | 0.9 | 0.9 | $x_1 + x_2 = 10$ |
| IX | $0.01(10 - x_1)$ | $0.01(10 - x_2)$ | 0.9 | 0.9 | $x_1 + x_2 = 10$ |
| X | $0.05(10 - x_1)$ | $0.05(10 - x_2)$ | $0.99\min(10, x_1)$ | $0.99\min(10, x_2)$ | $x_1 + x_2 = 10$ |
| XI | $0.02(20 - x_1)$ | $0.02(20 - x_2)$ | 0.8 | 0.8 | $x_1 + x_2 = 20$ |

consists of 50000 samples from independent regenerative cycles. The transition rate functions $\alpha_{(i,j)}(x)$ and the rare event state space, $\mathscr{R}$, for each simulation case are listed in Table 1. The simulated cases include models with state independent transition rates, with mixtures of state dependent and state independent rates, and cases with equal and different rates in the two dimensions. The objective in all cases is to estimate the steady state probabilities of the states in the rare event state set $\mathscr{R}$ which was chosen to include either all combinations where $x_1 + x_2 = N$ for $N = 10$ and 20, or a specific single state chosen as $(x_1, x_2) = (5,5)$ or $(x_1, x_2) = (7,3)$, respectively.

The results of all simulation cases are shown in Table 2. The table includes exact values obtained by a numerical method described in Iversen (1987) as well as the estimated means and the relative errors of the sample means from the simulation experiments. The results show good performance. All estimates are extremely accurate with very small relative errors.

Table 2: Simulation results

| Case | exact | $\bar{X}$ | $S_{\bar{X}}/\bar{X}$ |
|---|---|---|---|
| I | $2.49 \times 10^{-09}$ | $2.52 \times 10^{-09}$ | 0.025 |
| II | $9.98 \times 10^{-10}$ | $9.97 \times 10^{-10}$ | 0.024 |
| III | $2.77 \times 10^{-10}$ | $2.81 \times 10^{-10}$ | 0.008 |
| IV | $2.27 \times 10^{-10}$ | $2.13 \times 10^{-10}$ | 0.052 |
| V | $2.27 \times 10^{-10}$ | $2.24 \times 10^{-10}$ | 0.045 |
| VI | $1.12 \times 10^{-10}$ | $1.10 \times 10^{-10}$ | 0.050 |
| VII | $1.24 \times 10^{-10}$ | $1.17 \times 10^{-10}$ | 0.045 |
| VIII | $8.84 \times 10^{-10}$ | $8.76 \times 10^{-10}$ | 0.044 |
| IX | $8.45 \times 10^{-11}$ | $8.80 \times 10^{-11}$ | 0.054 |
| X | $7.45 \times 10^{-09}$ | $7.41 \times 10^{-09}$ | 0.067 |
| XI | $6.31 \times 10^{-09}$ | $6.79 \times 10^{-09}$ | 0.062 |

## 4.2 Inner Workings of Ant-Based Change of Measure

In order to enhance the understanding of how the transition rate functions are updated, we have studied the change of measure obtained by our method in more detail. In Figure 3 the change of the rates for the transitions that increase the number of items in the system, i.e. either $x_1$ or $x_2$, is visualized. The thicker the lines, the more the rates are increased (five levels). In Figure 4 the change of the rates for the transitions that decrease the number of items in the system, i.e. either $x_1$ or $x_2$ is visualized similarly. The thicker the lines, the more the rates are decreased. The figures show the change of measure for simulation case IX in Table 1. This case has state dependent rate functions for transitions that increase the population, and state independent rate functions otherwise. This implies that the most likely of the rare events are in the center of the rare event state space as indicated by the marginal distribution $P(\{(x_1, x_2)\} : x_1 + x_2 = N)$ in Figure 3 and 4.

It can be observed from Figure 3 that the change of the rates for the transition class that increases the population in the system tends to be stronger for states closer to the resource boundary than in states further away from the boundary for the same transition class. At the same time we observe from Figure 4 that the change of the rates for the transition class that decreases the population in the system

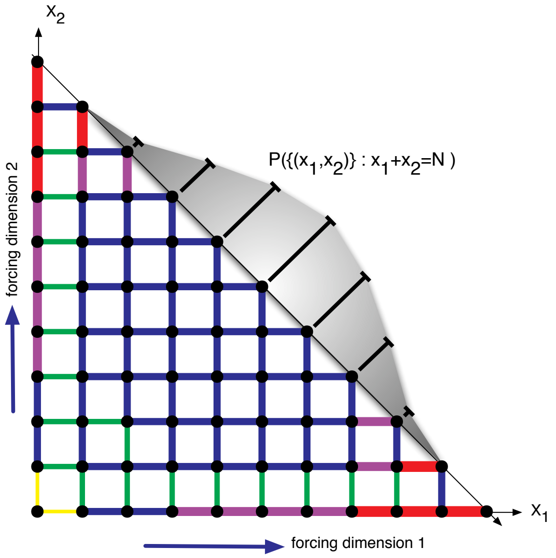

Figure 3: Change of measure with state dependent rate function - the relative increase in the rate functions for the transitions that increase the system population

tends to be slightly stronger in states far away from the resource boundary than in states close to the boundary for the same transition class. For the interior states we observe that the change of measure is less dependent on the state.

## 5 CONCLUSION AND FURTHER RESEARCH

We have presented a new ant-based method for adaptively obtaining the change of measure in Importance Sampling for rare event simulation. Numerical results for several variants of a two-dimensional example model show that the Importance Sampling simulations yield very accurate estimates. The inner workings of the method have been illustrated by means of visualizing the state dependent relative increases and decreases in rate functions. Preliminary tests indicate that the method also works well for models of higher dimensionality. Of course, visualization of the change of measure in a similar fashion as for the two-dimensional case is hardly possible.

It should be noted that the applicability of the method is not restricted to exponentially distributed times. It is easy to incorporate phase-type distributions into the transition class formalism and then to apply the ant-based method similarly as it is done in the present paper. This will be one topic of future investigations. Further research also includes systematic studies of the properties of the Importance Sampling estimators obtained via the ant-based approach. Although a main motivation is the applicability to rare events with probabilities in orders of magnitudes of practical interest, it is of course also interesting and useful to examine asymptotic properties.

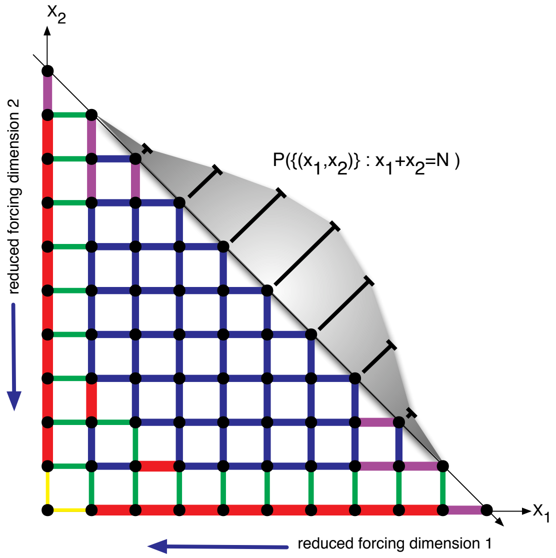
$x_2$
$x_1$
reduced forcing dimension 2
reduced forcing dimension 1
$P(\{(x_1,x_2)\} : x_1+x_2=N)$

Figure 4: Change of measure with state dependent rate function - the relative decrease in the rate functions for the transitions that decrease the system population

## REFERENCES

Blanchet, J. H., P. W. Glynn, P. L'Ecuyer, W. Sandmann, and B. Tuffin. 2007. Asymptotic robustness of estimators in rare-event simulation. In *Proceedings of the 2007 INFORMS Simulation Society Research Workshop*. To appear.

Carrasco, J. A. 1992. Failure distance based simulation of repairable fault-tolerant systems. In *Proceedings of the 5th International Conference on Modeling Techniques and Tools for Computer Performance Evaluation*, 337–351.

Cottrell, M., J. C. Fort, and G. Malgouyres. 1983. Large deviations and rare events in the study of stochastic algorithms. *IEEE Transactions on Automatic Control* 28:907–920.

de Boer, P.-T. 2000. *Analysis and efficient simulation of queueing models of telecommunications systems*. Ph. D. thesis, University of Twente, The Netherlands.

de Boer, P.-T. 2006. Analysis of state-independent importance sampling measures for the two-node tandem queue. *ACM Transactions on Modeling and Computer Simulation* 16 (3): 225–250.

Dembo, A., and O. Zeitouni. 1998. *Large deviations techniques and applications*. 2nd ed. Springer.

Devetsikiotis, M., W. A. Al-Qaq, J. A. Freebersyser, and J. K. Townsend. 1993. Stochastic gradient techniques for the efficient simulation of high-speed networks using importance sampling. In *Proceedings of the IEEE Global Telecommunications Conference, GLOBECOM'93*, 1718–1722.

Devetsikiotis, M., and J. K. Townsend. 1993a. An algorithmic approach to the optimization of importance sampling parameters in digital communication system simulation. *IEEE Transactions on Communications* 41 (10): 1464–1473.

Devetsikiotis, M., and J. K. Townsend. 1993b. Statistical optimization of dynamic importance sampling parameters for efficient simulation of communication networks. *IEEE/ACM Transactions on Networking* 1 (3): 293–305.

Dorigo, M., V. Maniezzo, and A. Colorni. 1996. The ant system: Optimization by a colony of cooperating agents. *IEEE Transactions on Systems, Man, Cybernetics, Part B: Cybernetics* 26 (1): 29–41.

Freebersyser, J. A., M. Devetsikiotis, W. A. Al-Qaq, and J. K. Townsend. 1996. Fast simulation of tandem networks using importance sampling and stochastic gradient techniques. In *Proceedings of the International Conference on Communications, ICC'96*.

Glynn, P. W., and D. L. Iglehart. 1989. Importance sampling for stochastic simulations. *Management Science* 35 (11): 1367–1392.

Hammersley, J. M., and D. C. Handscomb. 1964. *Monte carlo methods*. Methuen.

Heegaard, P. E. 1998a. *Efficient simulation of network performance by importance sampling*. Ph. D. thesis, Norwegian University of Science and Technology.

Heegaard, P. E. 1998b. A scheme for adaptive biasing in importance sampling. *AEÜ International Journal of Electronics and Communications, Special Issue on Rare Event Simulation* 52 (3): 172–182.

Heidelberger, P. 1995. Fast simulation of rare events in queueing and reliability models. *ACM Transactions on Modeling and Computer Simulation* 5 (1): 43–85.

Iversen, V. B. 1987. A simple convolution algorithm for the exact evaluation of multi-service loss system with heterogeneous traffic flows and access control. In *The 7th Nordic Teletraffic Seminar (NTS-7)*, IX.3–1–IX.3–22.

Juneja, S., and P. Shahabuddin. 2006. Rare event simulation techniques: An introduction and recent advances. In *Simulation*, ed. S. G. Henderson and B. L. Nelson, Handbooks in Operations Research and Management Science, 291–350. Amsterdam, The Netherlands: Elsevier. Chapter 11.

Rubinstein, R. Y. 1997. Optimization of computer simulation with rare events. *European Journal of Operations Research* 99:89–112.

Rubinstein, R. Y., and D. P. Kroese. 2004. *The cross entropy method: A unified approach to combinatorial optimization, Monte-Carlo simulation and machine learning*. Springer.

Sandmann, W. 2004. Structured description of Markovian network models and its potentials for efficient rare event simulation. In *Proceedings of the 2nd International Conference on Performance Modelling and Evaluation of Heterogeneous Networks, HetNets'04*, P39/1–10.

Sandmann, W. 2005. Importance sampling in Markovian settings. In *Proceedings of the 2005 Winter Simulation Conference, WSC'05*, 499–508.

Sandmann, W. 2007. Efficiency of importance sampling estimators. *Journal of Simulation* 1 (2): 137–145.

Schoonderwoerd, R., O. Holland, J. Bruten, and L. Rothkrantz. 1997. Ant-based Load Balancing in Telecommunications Networks. *Adaptive Behavior* 5 (2): 169–207.

Shwartz, A., and A. Weiss. 1995. *Large deviations for performance analysis*. Chapman & Hall.

Srinivasan, R. 2002. *Importance sampling: Applications in communications and detection*. Springer.

Stadler, J. S., and S. Roy. 1993. Adaptive importance sampling. *IEEE Journal on Selected Areas in Communications* 11 (3): 309–316.

## AUTHOR BIOGRAPHIES

**POUL E. HEEGAARD** is an associate professor at Department of Telematics at Norwegian University of Science and Technology (NTNU) and a senior scientist at Telenor R&I. He received his MSc (Siv. Ing.) in 1988 and his PhD (Dr. Ing.) in 1998 from NTNU. His research interest is within the areas of performance and dependability evaluation of communication systems. He has special interests in speedup simulation techniques, and adaptive, distributed monitoring and management techniques in dynamic networks. His e-mail address is ⟨poul.heegaard@item.ntnu.no⟩, and his web page can be found at ⟨http://www.item.ntnu.no/~poulh/⟩.

**WERNER SANDMANN** studied Computer Science and Mathematics at the University of Bonn (Germany) where he received his diploma degree and his PhD (Dr. rer. nat.) both in Computer Science in 1998 and 2004, respectively. From 1998 to 2003 he was a scientific member of the Computer Science Department at the University of Bonn. Since 2004 he is an assistant professor at the Department of Information Systems and Applied Computer Science of the University of Bamberg (Germany). His research interests are in applied probability and stochastic modeling, including computer systems performance evaluation, reliability, quality of services, computational biology, analytical and numerical solution techniques, and simulation. His e-mail address is werner.sandmann@wiai.uni-bamberg.de, and his web page can be found at ⟨http://www.uni-bamberg.de/index.php?id=8512⟩.