Refining the weighted stochastic simulation algorithm

Dan T. Gillespie,[1,a)] Min Roh,[2] and Linda R. Petzold[2]
[1]*Dan T Gillespie Consulting, 30504 Cordoba Pl., Castaic, California 91384, USA*
[2]*Department of Computer Science, University of California Santa Barbara, Santa Barbara, California 93106, USA*

(Received 15 January 2009; accepted 21 March 2009; published online 6 May 2009)

The weighted stochastic simulation algorithm (wSSA) recently introduced by Kuwahara and Mura [J. Chem. Phys. **129**, 165101 (2008)] is an innovative variation on the stochastic simulation algorithm (SSA). It enables one to estimate, with much less computational effort than was previously thought possible using a Monte Carlo simulation procedure, the probability that a specified event will occur in a chemically reacting system within a specified time when that probability is very small. This paper presents some procedural extensions to the wSSA that enhance its effectiveness in practical applications. The paper also attempts to clarify some theoretical issues connected with the wSSA, including its connection to first passage time theory and its relation to the SSA. © *2009 American Institute of Physics*. [DOI: 10.1063/1.3116791]

I. INTRODUCTION

The *weighted stochastic simulation algorithm* (wSSA) recently introduced by Kuwahara and Mura[1] is an innovative variation on the standard stochastic simulation algorithm (SSA) which enables one to efficiently estimate the probability that a specified event will occur in a chemically reacting system within a specified time when that probability is very small, and the event is therefore "rare." The difficulty of doing this with the standard SSA has long been recognized as a limitation of the Monte Carlo simulation approach, so the wSSA is a welcomed development.

The implementation of the wSSA described in Ref. 1 does not, however, offer a convenient way to assess the accuracy of its probability estimate. In this paper we show how a simple refinement of the original wSSA procedure allows estimating a confidence interval for its estimate of the probability. This in turn, as we will also show, makes it possible to improve the efficiency of the wSSA by adjusting its parameters so as to reduce the estimated confidence interval. As yet, though, a fully automated procedure for optimizing the wSSA is not in hand.

We begin in Sec. II by giving a derivation and discussion of the wSSA that we think will help clarify why the procedure is correct. In Sec. III we present our proposed modifications to the original wSSA recipe of Ref. 1, and in Sec. IV we show how these modifications allow easy estimation of the gain in computational efficiency over the SSA. In Sec. V we give some numerical examples that illustrate the benefits of our proposed procedural refinements. In Sec. VI we discuss the relationship between the wSSA and the problem of estimating mean first passage times using as an example the problem of spontaneous transitions between the stable states of a bistable system. In Sec. VII we summarize our findings and make an observation on the relationship between the wSSA and the SSA.

a)Electronic mail: gillespiedt@mailaps.org.

II. THEORETICAL UNDERPINNINGS OF THE wSSA

We consider a well-stirred chemical system whose molecular population state at the current time t is $\mathbf{x}$. The next firing of one of the system's M reaction channels $R_1,\ldots,R_M$ will carry the system from state $\mathbf{x}$ to one of the M states $\mathbf{x}+\boldsymbol{\nu}_j$ $(j=1,\ldots,M)$, where $\boldsymbol{\nu}_j$ is (by definition) the state change caused by the firing of one R_j reaction. The fundamental premise of stochastic chemical kinetics, which underlies both the chemical master equation and the SSA, is that the probability that an R_j event will occur in the next infinitesimal time interval dt is $a_j(\mathbf{x})dt$, where a_j is called the propensity function of reaction R_j. It follows from this premise that (a) the probability that the system will jump away from state $\mathbf{x}$ between times $t+\tau$ and $t+\tau+d\tau$ is $a_0(\mathbf{x})e^{-a_0(\mathbf{x})\tau}d\tau$, where $a_0(\mathbf{x})\equiv\sum_{i=1}^{M}a_i(\mathbf{x})$, and (b) the probability that the system, upon jumping away from state $\mathbf{x}$, will jump to state $\mathbf{x}+\boldsymbol{\nu}_j$, is $a_j(\mathbf{x})/a_0(\mathbf{x})$. Applying the multiplication law of probability theory, we conclude that the probability that the next reaction will carry the system's state to $\mathbf{x}+\boldsymbol{\nu}_j$ between times $t+\tau$ and $t+\tau+d\tau$ is

$$\text{Prob}\{\mathbf{x}\to\mathbf{x}+\boldsymbol{\nu}_j \text{ in } (t+\tau,t+\tau+d\tau)\} = a_0(\mathbf{x})e^{-a_0(\mathbf{x})\tau}d\tau\times\frac{a_j(\mathbf{x})}{a_0(\mathbf{x})}. \quad (1)$$

In the usual "direct method" implementation of the SSA, the time τ to the next reaction event is chosen by sampling the exponential random variable with mean $1/a_0(\mathbf{x})$, in consonance with the first factor in Eq. (1), and the index j of the next reaction is chosen with probability $a_j(\mathbf{x})/a_0(\mathbf{x})$, in consonance with the second factor in Eq. (1). But now let us suppose, with Kuwahara and Mura,[1] that we modify the direct method SSA procedure so that, while it continues to choose the time τ to the next jump in the same way, it chooses the index j, which determines the destination $\mathbf{x}+\boldsymbol{\nu}_j$ of that jump, with probability $b_j(\mathbf{x})/b_0(\mathbf{x})$, where $\{b_1,\ldots,b_M\}$ is a possibly different set of functions from $\{a_1,\ldots,a_M\}$, and $b_0(\mathbf{x})\equiv\sum_{i=1}^{M}b_i(\mathbf{x})$. If we made that modifi-

cation, then the probability on the left hand side of Eq. (1) would be $a_0(\mathbf{x})e^{-a_0(\mathbf{x})\tau}d\tau\times(b_j(\mathbf{x})/b_0(\mathbf{x}))$. But we observe that this "incorrect" value can be converted to the "correct" value, on the right hand side of Eq. (1), simply by multiplying by the factor

$$w_j(\mathbf{x}) = \frac{a_j(\mathbf{x})/a_0(\mathbf{x})}{b_j(\mathbf{x})/b_0(\mathbf{x})}. \tag{2}$$

So in some sense, we can say that an $\mathbf{x}\to\mathbf{x}+\boldsymbol{\nu}_j$ jump generated using this modified procedure, and accorded a *statistical weight* of $w_j(\mathbf{x})$ in Eq. (2), is "equivalent" to an $\mathbf{x}\to\mathbf{x}+\boldsymbol{\nu}_j$ jump generated using the standard SSA.

This statistical weighting of a *single-reaction jump* can be extended to an *entire trajectory* of the system's state by reasoning as follows: A state trajectory is composed of a succession of single-reaction jumps. Each jump has a probability (1) that depends on the jump's starting state but *not* on the history of the trajectory that leads up to that starting state. Therefore, the probability of the trajectory as a whole is just the *product* of the probabilities of all the individual jumps (1) that make up the trajectory. Since in the modified SSA scheme the probability of each individual jump requires a correction factor of the form (2), then the correction factor for the entire trajectory—i.e., the statistical weight w of the trajectory–will be the product $w=w_{j_1}w_{j_2}w_{j_3}\cdots$, where w_{j_k} is the statistical weight (2) for the kth jump in that trajectory.

One situation where this statistical weighting logic can be applied is in the Monte Carlo averaging method of estimating the value of

$$p(\mathbf{x}_0,\mathcal{E};t) \equiv \text{the probability that the system starting at time 0 in state } \mathbf{x}_0 \text{ will first reach any state in the set } \mathcal{E} \text{ at some time } \le t. \tag{3}$$

[Note that $p(\mathbf{x}_0,\mathcal{E};t)$ is *not* the probability that the system will be in the set $\mathcal{E}$ at time t.] An obvious Monte Carlo way to estimate this probability would be to make a very large number n of regular SSA runs, with each run starting at time 0 in state $\mathbf{x}_0$ and terminating *either* when some state $\mathbf{x}'\in\mathcal{E}$ is first reached *or* when the system time reaches t. If m_n is the number of those n runs that terminate for the first reason, then the probability $p(\mathbf{x}_0,\mathcal{E};t)$ could be estimated as the fraction m_n/n, and this estimate would become exact in the limit $n\to\infty$. But m_n here could also be defined as the sum of the "weights" of the runs, where each run is given a weight of 1 if it ends because some state in the set $\mathcal{E}$ is reached before time t and a weight of 0 otherwise. This way of defining m_n is useful because it allows us to score runs in the *modified* SSA scheme, with each run that reaches some state $\mathbf{x}'\in\mathcal{E}$ before time t then being scored with its *trajectory weight* w as defined above. Kuwahara and Mura[1] recognized that this tactic could be used to advantage in the case $p(\mathbf{x}_0,\mathcal{E};t)\ll 1$, where using the standard SSA will inevitably require an impractically large number of trajectories to obtain an accurate estimate of $p(\mathbf{x}_0,\mathcal{E};t)$. As we shall elaborate in the next two sections, by using this wSSA method with the b_j functions carefully chosen so that they *increase* the likelihood of the system reaching $\mathcal{E}$, it is often possible to obtain a more accurate estimate of $p(\mathbf{x}_0,\mathcal{E};t)$ with far fewer runs.

The wSSA procedure given in Ref. 1 for computing $p(\mathbf{x}_0,\mathcal{E};T)$ in this way goes as follows:

1° $m_n\leftarrow 0$.
2° **for** $k=1$ to n, **do**
3° $s\leftarrow 0$, $\mathbf{x}\leftarrow\mathbf{x}_0$, $w\leftarrow 1$.
4° evaluate all $a_i(\mathbf{x})$ and $b_i(\mathbf{x})$; calculate $a_0(\mathbf{x})$ and $b_0(\mathbf{x})$.
5° **while** $s\le t$, **do**
6° **if** $x\in\mathcal{E}$, **then**
7° $m_n\leftarrow m_n+w$.
8° break out of the while loop.
9° **end if**
10° generate two unit-interval uniform random numbers r_1 and r_2.
11° $\tau\leftarrow a_0^{-1}(\mathbf{x})\ln(1/r_1)$.
12° $j\leftarrow$ smallest integer satisfying $\Sigma_{i=1}^{j}b_i(\mathbf{x})\ge r_2 b_0(\mathbf{x})$.
13° $w\leftarrow w\times(a_j(\mathbf{x})/b_j(\mathbf{x}))\times(b_0(\mathbf{x})/a_0(\mathbf{x}))$.
14° $s\leftarrow s+\tau$, $\mathbf{x}\leftarrow\mathbf{x}+\boldsymbol{\nu}_j$.
15° update $a_i(\mathbf{x})$ and $b_i(\mathbf{x})$; recalculate $a_0(\mathbf{x})$ and $b_0(\mathbf{x})$.
16° **end while**
17° **end for**
18° report $p(\mathbf{x}_0,\mathcal{E};t)=m_n/n$.

Assumed given for the above procedure are the reaction propensity functions a_j and the associated state-change vectors $\boldsymbol{\nu}_j$, the target set of states $\mathcal{E}$ and the time t by which the system should reach that set, the total number of runs n that will be made to obtain the estimate, and the step-biasing functions b_j (which Kuwahara and Mura called predilection functions). The variable m_n in the above procedure is the *sum* of the statistical weights w of the n run trajectories. The value of w for each trajectory is constructed in step 13°, as the product of the weights w_j in Eq. (2) of all the reaction jumps making up that trajectory; however, if a trajectory ends because in the given time t the set $\mathcal{E}$ has not been reached, the weight of that trajectory is summarily set to zero. Note that the use of a_0 instead of b_0 to compute the jump time τ in step 11° follows from the analysis leading from Eqs. (1) and (2): the wSSA introduces an artificial bias in choosing j, but it always chooses τ "properly" according to the true propensity functions. This strategy of using the correct τ is vital for allotting to each trajectory the proper amount of time t to reach the target set of states $\mathcal{E}$.

If the b_j functions are chosen to be the *same* as the a_j functions, then the above procedure evidently reduces to the standard SSA. Thus, the key to making the wSSA more efficient than the SSA is to choose the b_j functions "appropriately." It is seen from step 13°, though, that b_j must not have a harder zero at any accessible state point than a_j, for otherwise the weight at that state point would be infinite. To keep that from happening, Kuwahara and Mura proposed the simple procedure of setting

$$b_j(\mathbf{x}) = \gamma_j a_j(\mathbf{x}) \quad (j=1,\ldots,M), \tag{4}$$

where each proportionality constant $\gamma_j>0$, which we shall

call the *importance sampling factor* for reaction R_j, is chosen to be ≥ 1 if the occurrence of reaction R_j *increases* the chances of the system reaching the set $\mathcal{E}$ and ≤ 1 otherwise. This way of choosing the b functions seems quite reasonable, although a minor subtlety not mentioned in Ref. 1 is that, since the wSSA works by altering the *relative* sizes of the propensity functions for state selection, only $M-1$ of the γ_j matter; in particular, in a system with only one reaction, weighting that reaction by any factor γ will produce a single step weight (2) that is always unity, and the wSSA therefore reduces to the SSA. But of course, single-reaction systems are not very interesting in this context. A more important question in connection with Eq. (4) is: Are there optimal values for the γ_j? And if so, how might we identify them?

III. THE VARIANCE AND ITS BENEFITS

The statistical weighting strategy described in connection with Eq. (4) evidently has the effect of increasing the firing rates of those "important reactions" that move the system toward the target states $\mathcal{E}$, thus producing more "important trajectories" that reach that target. Equation (2) shows that boosting the likelihoods of those successful trajectories in this way will cause them to have statistical weights $w<1$. As was noted and discussed at some length in Ref. 1, this procedure is an example of a general Monte Carlo technique called *importance sampling*. However, the description of the importance sampling strategy given in Ref. 1 is incomplete because it makes no mention of something called the "sample variance."

In the Appendix, we give a brief review of the general theory underlying Monte Carlo averaging and the allied technique of importance sampling which explains the vital connecting role played by the sample variance. The bottom line for the wSSA procedure described in Sec. II is this: The computation of the *sample mean* m_n/n of the weights of the n wSSA trajectories should be accompanied by a computation of the *sample variance* of those trajectory weights. Doing that not only provides us with a quantitative estimate of the *uncertainty* in the approximation $p(\mathbf{x}_0,\mathcal{E};t)\approx m_n/n$ but also helps us find the values of the parameters γ_j in Eq. (4) that *minimize* that uncertainty. More specifically (see the Appendix for details), in addition to computing the sample first moment (or sample mean) of the weights of the wSSA-generated trajectories,

$$\frac{m_n}{n} \equiv \frac{m_n^{(1)}}{n} \equiv \frac{1}{n}\sum_{k=1}^{n} w_k, \tag{5}$$

where w_k is the statistical weight of run k [equal to the product of the weights (2) of each reaction that occurs in run k if that run reaches $\mathcal{E}$ before t and zero otherwise], we should also compute the sample second moment of those weights,

$$\frac{m_n^{(2)}}{n} \equiv \frac{1}{n}\sum_{k=1}^{n} w_k^2. \tag{6}$$

The sample variance of the weights is then given by the difference between the sample second moment and the square of the sample first moment:[2]

$$\sigma^2 = (m_n^{(2)}/n) - (m_n^{(1)}/n)^2. \tag{7}$$

The final estimate $p(\mathbf{x}_0,\mathcal{E};t)\approx m_n^{(1)}/n$ can then be assigned a "one-standard-deviation normal confidence interval" of

$$\text{uncertainty} = \pm \frac{\sigma}{\sqrt{n}}. \tag{8}$$

This means that the probability that the true value of $p(\mathbf{x}_0,\mathcal{E};t)$ will lie within $\sigma/\sqrt{n}$ of the estimate $m_n^{(1)}/n$ is 68%. Doubling the uncertainty interval (8) raises the confidence level to 95%, and tripling it gives us a confidence level of 99.7%. Furthermore, by performing multiple runs that vary the b_j functions, which in practice means systematically varying the parameters γ_j in Eq. (4), we can, at least in principle, find the values of γ_j that give the smallest σ^2, and hence according to Eq. (8) the most accurate estimate of $p(\mathbf{x}_0,\mathcal{E};t)$ for a given value of n.

All of the foregoing is premised on the assumption that n has been taken "sufficiently large." That is because there is some "bootstrapping logic" used in the classical Monte Carlo averaging method (independently of importance sampling): The values for $m_n^{(1)}$ and $m_n^{(2)}$ computed in Eqs. (5) and (6) will vary from one set of n runs to the next, so the computed value of σ^2 in Eqs. (7) and (8) will also vary. Therefore, as discussed more fully in the Appendix at Eqs. (A9) and (A10), the computed uncertainty in the estimate of the mean is itself only an estimate. And, like the estimate of the mean, the estimate of the uncertainty will be reasonably accurate only if a sufficiently large number n of runs have been used. In practice, this means that only when several repetitions of an n-run calculation are found to produce approximately the same estimates for $m_n^{(1)}$ and $m_n^{(2)}$ can we be sure that n has been taken large enough to draw reliable conclusions.

When the original wSSA recipe in Sec. II is modified to include the changes described above, we obtain the recipe given below:

1° $m_n^{(1)} \leftarrow 0$, $m_n^{(2)} \leftarrow 0$
2° **for** $k=1$ to n, **do**
3° $s \leftarrow 0$, $\mathbf{x} \leftarrow \mathbf{x}_0$, $w \leftarrow 1$
4° evaluate all $a_i(\mathbf{x})$ and $b_i(\mathbf{x})$; calculate $a_0(\mathbf{x})$ and $b_0(\mathbf{x})$.
5° **while** $s \leq t$, **do**
6° **if** $\mathbf{x} \in \mathcal{E}$, **then**
7° $m_n^{(1)} \leftarrow m_n^{(1)} + w$, $m_n^{(2)} \leftarrow m_n^{(2)} + w^2$.
8° break out of the while loop.
9° **end if**
10° generate two unit-interval uniform random numbers r_1 and r_2.
11° $\tau \leftarrow a_0^{-1}(\mathbf{x})\ln(1/r_1)$
12° $j \leftarrow$ smallest integer satisfying $\sum_{i=1}^{j} b_i(\mathbf{x}) \geq r_2 b_0(\mathbf{x})$.
13° $w \leftarrow w \times (a_j(\mathbf{x})/b_j(\mathbf{x})) \times (b_0(\mathbf{x})/a_0(\mathbf{x}))$.
14° $s \leftarrow s+\tau$, $\mathbf{x} \leftarrow \mathbf{x} + \boldsymbol{\nu}_j$.
15° update $a_i(\mathbf{x})$ and $b_i(\mathbf{x})$; recalculate $a_0(\mathbf{x})$ and $b_0(\mathbf{x})$.
16° **end while**

17° **end for**
18° $\sigma^2=(m_n^{(2)}/n)-(m_n^{(1)}/n)^2$
19° repeat from 1° using different b functions to minimize σ^2.
20° estimate $p(\mathbf{x}_0,\mathcal{E};t)=m_n^{(1)}/n$, with a 68% uncertainty of $\pm\sigma/\sqrt{n}$.

Steps 1°–17° are identical to those in the earlier procedure in Sec. II, except for the additional computations involving the new variable $m_n^{(2)}$ in steps 1° and 7°. The new step 18° computes the variance. Step 19° tunes the importance sampling parameters γ_j in Eq. (4) to minimize that variance. And step 20° uses the optimal set of γ_j values thus found to compute the best estimate of $p(\mathbf{x}_0,\mathcal{E};t)$, along with its associated confidence interval. In practice, step 19° usually has to be done manually, external to the computer program, since the search over γ_j space requires some intuitive guessing; this is typical in most applications of importance sampling.[3] An overall check on the validity of the computation can be made by repeating it a few times with different random number seeds to verify that the estimates obtained for $p(\mathbf{x}_0,\mathcal{E};t)$ and its confidence interval are reproducible and consistent. If they are not, then n has probably not been chosen large enough.

IV. GAIN IN COMPUTATIONAL EFFICIENCY

The problem with using unweighted SSA trajectories to estimate $p(\mathbf{x}_0,\mathcal{E};t)$ when that probability is $\ll 1$ is that we are then trying to estimate the average of a set of numbers (the trajectory weights) which are all either 0 or 1 when that average is much closer to 0 than to 1. The sporadic occurrence of a few 1's among a multitude of 0's makes this estimate subject to very large statistical fluctuations for any reasonable number of trajectories n. How does importance sampling overcome this problem? If the reaction biasing is done properly, most of the "successful" trajectories that reach the target set $\mathcal{E}$ within the allotted time t will have weights that are much less than 1, and hence closer to the average. Most of the "unsuccessful" trajectories will rack up weights in step 13° that are much greater than 1, but when the simulated time reaches the limit t without the set $\mathcal{E}$ having been reached, those large weights are summarily reset to zero (they never get accumulated in $m_n^{(1)}$ and $m_n^{(2)}$ in step 7°). The result is that the bulk of the contribution to the sample average comes from weights that are much closer to the average than are the unit weights of the successful SSA trajectories. This produces a smaller scatter in the weights of wSSA trajectories about their average, as measured by their standard deviation σ, and hence a more accurate estimate of that average. Note, however, that if the event in question is *not* rare, i.e., if $p(\mathbf{x}_0,\mathcal{E};t)$ is not $\ll 1$, then the unit trajectory weights of the SSA do not pose a statistical problem. In that case there is little to be gained by importance sampling, and the ordinary SSA should be adequate. Note also that the rarity of the event is always connected to the size of t. Since $p(\mathbf{x}_0,\mathcal{E};t)\to 1$ as $t\to\infty$, it is always possible to convert a rare event into a likely event simply by taking t sufficiently large.

To better understand how variance reduction through importance sampling helps when $p(\mathbf{x}_0,\mathcal{E};t)\ll 1$, let us consider what happens when *no* importance sampling is done, i.e., when $b_j=a_j$ for all j and every successful trajectory gets assigned a weight $w=1$. Letting m_n denote the number of successful runs obtained out of n total, it follows from definitions (5) and (6) that

$$m_n^{(1)}=m_n\times 1=m_n,\quad m_n^{(2)}=m_n\times 1^2=m_n.$$

Equation (7) then gives for the sample variance

$$\sigma^2=(m_n/n)-(m_n/n)^2=(m_n/n)(1-(m_n/n)).$$

The uncertainty (8) is therefore[4]

$$\text{uncertainty}=\pm\sqrt{\frac{(m_n/n)(1-(m_n/n))}{n}}, \tag{9a}$$

and this implies a *relative* uncertainty of

$$\text{relative uncertainty}\equiv\frac{\text{uncertainty}}{m_n/n}=\pm\sqrt{\frac{1-(m_n/n)}{m_n}}. \tag{9b}$$

When $p(\mathbf{x}_0,\mathcal{E};t)\approx m_n/n\ll 1$, Eq. (9b) simplifies to

$$\text{relative uncertainty}\approx\pm\sqrt{\frac{1}{m_n}}\quad(\text{if } m_n/n\ll 1). \tag{10}$$

This shows that if only *one* successful run is encountered in the n SSA runs, then the relative uncertainty in the estimate of $p(\mathbf{x}_0,\mathcal{E};t)$ will be 100%, and if *four* successful runs are encountered, the relative uncertainty will be 50%. To reduce the relative uncertainty to a respectably accurate 1% would, according to Eq. (10), require 10 000 successful SSA runs, and that would be practically impossible for a truly rare event.

These considerations allow us to estimate the number of unweighted SSA runs, n^{SSA}, that would be needed to yield an estimate of $p(\mathbf{x}_0,\mathcal{E};t)$ that has *the same relative accuracy* as the estimate obtained in a wSSA calculation. Thus, suppose a wSSA calculation with n^{wSSA} runs has produced the estimate $\hat{p}$ $(=m_n^{(1)}/n^{\mathrm{wSSA}})$ with a one-standard-deviation uncertainty u^{wSSA} $(=\sigma^{\mathrm{wSSA}}/\sqrt{n^{\mathrm{wSSA}}})$. The relative uncertainty is $u^{\mathrm{wSSA}}/\hat{p}$. According to Eq. (10), to get that same relative uncertainty using the unweighted SSA, we would need m^{SSA} successful SSA runs such that

$$\sqrt{\frac{1}{m^{\mathrm{SSA}}}}=\frac{u^{\mathrm{wSSA}}}{\hat{p}}.$$

But to get m^{SSA} successful runs with the SSA, we would need to make n^{SSA} total runs, where

$$m^{\mathrm{SSA}}/n^{\mathrm{SSA}}=\hat{p}.$$

Solving this last equation for m^{SSA}, substituting the result into the preceding equation, and then solving it for n^{SSA}, we obtain

$$n^{\mathrm{SSA}}=\frac{\hat{p}}{(u^{\mathrm{wSSA}})^2}\quad(\text{if } \hat{p}\ll 1). \tag{11}$$

A rough measure of the *gain in computational efficiency* of

the wSSA over the SSA is provided by the ratio of n^{SSA} to n^{wSSA}:

$$g \equiv \frac{n^{\mathrm{SSA}}}{n^{\mathrm{wSSA}}} = \frac{\hat{p}}{n^{\mathrm{wSSA}}(u^{\mathrm{wSSA}})^2}.$$

Since $u^{\mathrm{wSSA}} = \sigma^{\mathrm{wSSA}}/\sqrt{n^{\mathrm{wSSA}}}$, this simplifies to

$$g = \frac{\hat{p}}{(\sigma^{\mathrm{wSSA}})^2} \quad (\text{if } \hat{p} \ll 1). \tag{12}$$

The result (12) shows why the wSSA's strategy of minimizing the variance when $p(\mathbf{x}_0, \mathcal{E}; t) \ll 1$ is the key to obtaining a large gain in computational efficiency over the unweighted SSA: If we can contrive to *halve* the variance, we will *double* the efficiency.

V. NUMERICAL EXAMPLES

Reference 1 illustrated the wSSA by applying it to two simple systems. In this section we repeat those applications in order to illustrate the benefits of the refinements introduced in Sec. III.

The first example in Ref. 1 concerns the simple system

$$S_1 \xrightarrow{k_1} S_1 + S_2, \quad S_2 \xrightarrow{k_2} \varnothing, \tag{13}$$

with $k_1 = 1$ and $k_2 = 0.025$. Since the S_1 population x_1 remains constant in these reactions, Eq. (13) is mathematically the same as the reaction set $\varnothing \underset{k_2}{\overset{k_1 x_1}{\rightleftarrows}} S_2$. This reaction set has been well studied,[5] and the steady-state (equilibrium) population of species S_2 is known to be the Poisson random variable with mean and variance $k_1 x_1 / k_2$. Reference 1 takes $x_1 = 1$, so at equilibrium the S_2 population in Eq. (13) will be fluctuating about a mean of $k_1/k_2 = 40$ with a standard deviation of $\sqrt{40} = 6.3$. For this system, Ref. 1 sought to estimate, for several values of ε_2 between 65 and 80, the probability $p(40, \varepsilon_2; 100)$ that with $x_1 = 1$, the S_2 population, starting at the value 40, will reach the value ε_2 before time $t = 100$. Since the S_2 populations 65 and 80 are, respectively, about four and six standard deviations *above* the equilibrium value 40, then the biasing strategy for the wSSA must be to encourage reaction R_1, which increases the S_2 population, and/or discourage reaction R_2, which decreases the S_2 population. Of the several ways in which that might be done, Ref. 1 adopted scheme (4), taking $\gamma_1 = \alpha$ and $\gamma_2 = 1/\alpha$ with $\alpha = 1.2$.

Addressing first the case $\varepsilon_2 = 65$, we show in Fig. 1(a) a plot of σ^2 versus α for a range of α values near 1.2. In this plot, the center dot on each vertical bar is the average of the σ^2 results found in four runs of the wSSA procedure in Sec. III (or more specifically, steps 1°–18° of that procedure), with each run containing $n = 10^6$ trajectories. The span of each vertical bar indicates the one-standard-deviation envelope of the four σ^2 values. It is seen from this plot that the value of α that minimizes σ^2 for $\varepsilon_2 = 65$ is approximately 1.20, which is just the value used in Ref. 1. But Fig. 1(a) assures us that this value in fact gives the *optimal* importance sampling, at least for this value of ε_2 and this way of parametrizing γ_1 and γ_2. Using this optimal α value in a longer run

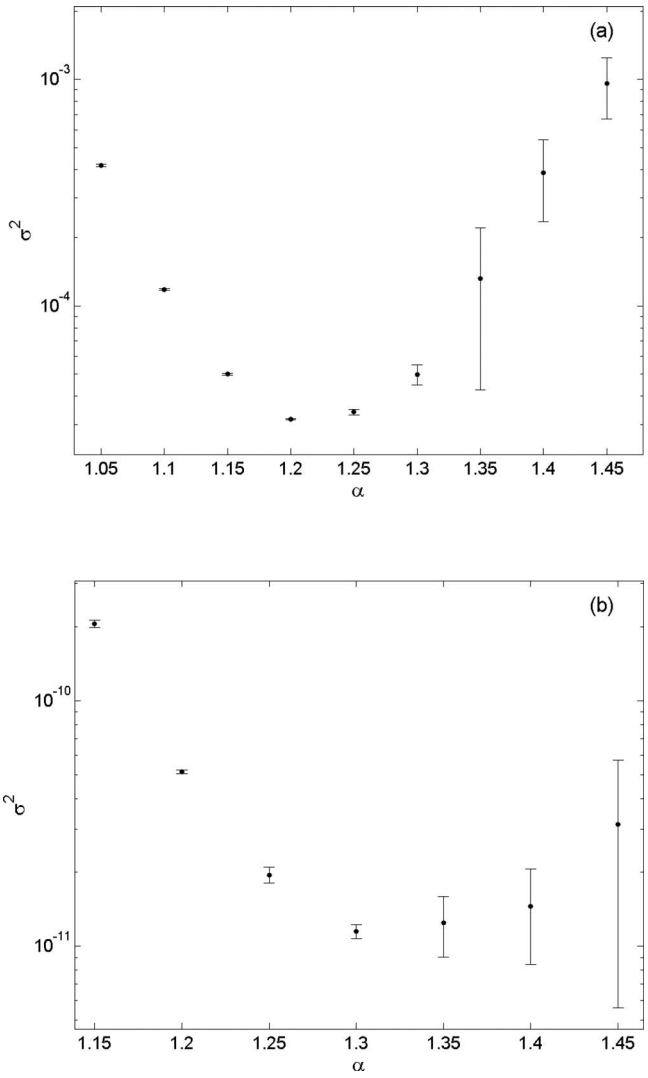

FIG. 1. (a) A plot of σ^2 vs α obtained in wSSA runs of reactions (13) that were designed to determine $p(40, \varepsilon_2; 100)$ for $\varepsilon_2 = 65$ using the biasing scheme $\gamma_1 = \alpha$ and $\gamma_2 = 1/\alpha$. Each vertical bar shows the estimated mean and one standard deviation of σ^2 at that α value as found in four $n = 10^6$ runs of the modified wSSA procedure in Sec. III. The optimal α value, defined as that which produces the smallest σ^2, is seen to be 1.20. (b) A similar plot for $\varepsilon_2 = 80$, except that here each σ^2 estimate was computed from four $n = 10^7$ runs. The optimal α value here is evidently 1.30, which gives a stronger bias than was optimal for the case in (a).

of the wSSA, now taking $n = 10^7$ as was done in Ref. 1, we obtained

$$p(40,65;100) = 2.307 \times 10^{-3} \pm 0.003 \times 10^{-3} \quad (95\% \text{ confidence}). \tag{14}$$

In this final result, we have been conservative and given the *two*-standard-deviation uncertainty interval. To estimate the gain in efficiency provided by the wSSA over the SSA, we substitute $\hat{p} = 2.3 \times 10^{-3}$ and $u^{\mathrm{wSSA}} = 0.0015 \times 10^{-3}$ into Eq. (11), and we get $n^{\mathrm{SSA}} = 1.025 \times 10^9$. Since result (14) was obtained with $n^{\mathrm{wSSA}} = 10^7$ wSSA runs, then the efficiency gain here over the SSA is $g = 103$, i.e., the computer running time to get result (14) using the unweighted SSA would be about a hundred times longer.

For the case $\varepsilon_2 = 80$, the plot of σ^2 versus α is shown in Fig. 1(b). In this case, obtaining a reasonably accurate esti-

mate of σ^2 at each α value required using four runs with $n=10^7$. But even then, as we move farther above $\alpha=1.3$, it evidently becomes very difficult to estimate σ^2 accurately in a run with only $n=10^7$ trajectories, as is indicated by the vertical bars showing the scatter (standard deviation) observed in four such runs. But each dot represents the combined estimate of σ^2 for $n=4\times10^7$ runs, and they allow us to see that the minimum σ^2 is obtained at about $\alpha=1.3$. That value, being further from 1 than the α value 1.20 which Ref. 1 used for $\varepsilon_2=80$ as well as for $\varepsilon_2=65$, represents a stronger bias than $\alpha=1.2$, which is reasonable. The four runs for $\alpha=1.3$ were finally combined into one run, an operation made easy by outputting at the end of each run the values of the cumulative sums $m_n^{(1)}$ and $m_n^{(2)}$: The four sums for $m_n^{(1)}$ were added together to get $m_{4n}^{(1)}$, and the four sums for $m_n^{(2)}$ similarly gave $m_{4n}^{(2)}$. This yielded the $n=4\times10^{-7}$ estimate

$$p(40,80;100) = 3.014\times10^{-7} \pm 0.011\times10^{-7} \quad (95\%\ \text{confidence}), \tag{15}$$

where again we have given a conservative two-standard-deviation uncertainty interval. To estimate the gain in efficiency provided by the wSSA over the SSA, we substitute $\hat{p}=3\times10^{-7}$ and $u^{\text{wSSA}}=0.0055\times10^{-7}$ into Eq. (11), and we find $n^{\text{SSA}}=9.96\times10^{11}$. Since result (13) was obtained with $n^{\text{wSSA}}=4\times10^7$ wSSA runs, the efficiency gain over the SSA is $g=2.5\times10^4$, which is truly substantial.

The second system considered in Ref. 1 is the six-reaction set

$$S_1 + S_2 \underset{k_2}{\overset{k_1}{\rightleftarrows}} S_3 \overset{k_3}{\rightarrow} S_1 + S_5,$$
$$S_4 + S_5 \underset{k_5}{\overset{k_4}{\rightleftarrows}} S_6 \overset{k_6}{\rightarrow} S_4 + S_2, \tag{16}$$

with the rate constants $k_1=k_2=k_4=k_5=1$ and $k_3=k_6=0.1$. These reactions are essentially a forward-reverse pair of enzyme-substrate reactions, with the first three reactions describing the S_1-catalyzed conversion of S_2 to S_5 and the last three reactions describing the S_4-catalyzed conversion of S_5 back to S_2. As was noted in Ref. 1, for the initial condition $\mathbf{x}_0=(1,50,0,1,50,0)$, each of the S_2 and S_5 populations tends to equilibrate about its initial value 50. Reference 1 sought to estimate, for several values of ε_5 between 40 and 25, the probability $p(\mathbf{x}_0,\varepsilon_5;100)$ that the S_5 population, initially at 50 molecules, will reach the value ε_5 before time $t=100$. Since those target S_5 populations are smaller than the $\mathbf{x}_0$ value 50, the wSSA biasing strategy should suppress the creation of S_5 molecules. One way to do that would be to discourage reaction R_3, which creates S_5 molecules, and encourage reaction R_6, which by creating S_4 molecules encourages the consumption of S_5 molecules via reaction R_4. The specific procedure adopted in Ref. 1 for doing that was to implement biasing scheme (4) with all the biasing parameters γ_j set to 1, except $\gamma_3=\alpha$ and $\gamma_6=1/\alpha$ with $\alpha=0.5$.

For the case $\varepsilon_5=40$, we first made some preliminary wSSA runs in order to estimate σ^2 for several values of α in the neighborhood of 0.5. The results are shown in Fig. 2(a). Here the center dot on each vertical bar shows the average of the σ^2 values found in four wSSA runs at that α, with each run containing $n=10^5$ trajectories. As before, the span of each vertical bar indicates the associated one-standard-deviation envelope. It is seen from this plot that the value of α that minimizes σ^2 for $\varepsilon_5=40$ is approximately 0.60, which is less biased (closer to 1) than the value 0.5 used in Ref. 1. Taking 0.60 as the optimal α value, we then made a longer $n=10^7$ run and got

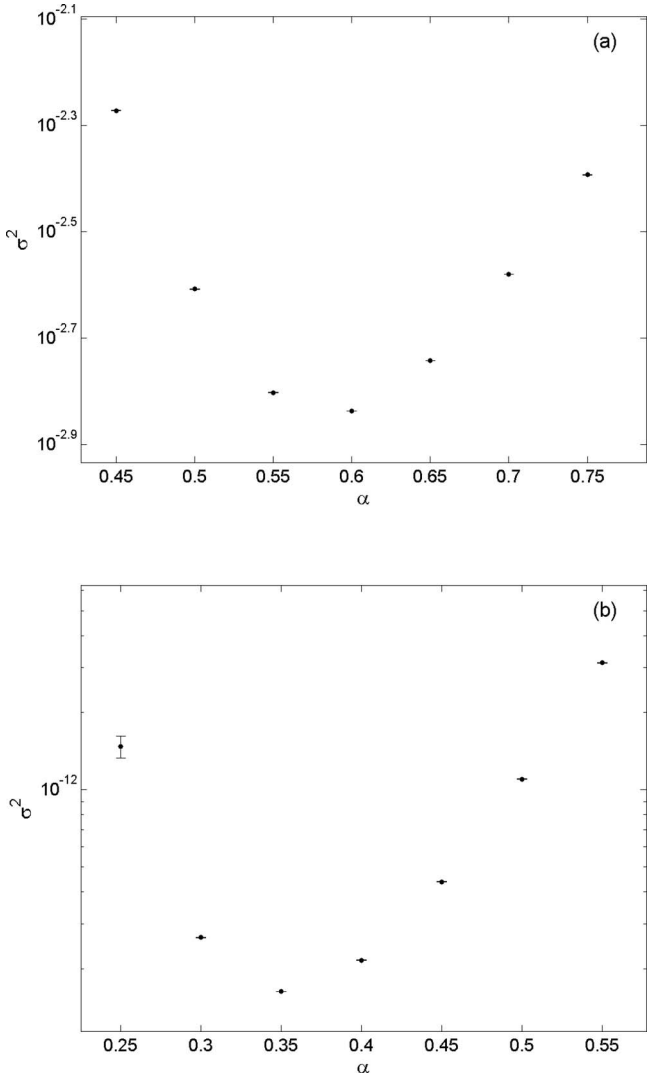

FIG. 2. (a) A plot of σ^2 vs α obtained in wSSA runs of reactions (16) that were designed to determine $p(\mathbf{x}_0,\varepsilon_5;100)$ for $\varepsilon_5=40$ using the biasing scheme $\gamma_3=\alpha$ and $\gamma_6=1/\alpha$. Each vertical bar shows the estimated mean and one standard deviation of σ^2 at that α value as found in four $n=10^5$ runs of the modified wSSA procedure in Sec. III. The optimal α value here is seen to be 0.60. (b) A similar plot for $\varepsilon_5=25$. The optimal α value now is 0.35, which gives a stronger bias than was optimal for the case in (a).

$$p(\mathbf{x}_0,40;100) = 0.042\ 21 \pm 0.000\ 02 \quad (95\%\ \text{confidence}). \tag{17}$$

For this value of $\hat{p}$ and a one-standard uncertainty of $u^{\text{wSSA}}=0.000\ 01$, formula (11) yields $n^{\text{SSA}}=4.22\times10^8$. This implies a gain in computational efficiency over the unweighted SSA of $g=42$.

For the case $\varepsilon_5=25$, the σ^2 versus α plot is shown in Fig. 2(b). As in Fig. 2(a), each vertical bar shows the result of

four wSSA runs with $n=10^5$. This plot shows that the optimal α value is now 0.35, which is more biased (i.e., further from 1) than the optimal α value 0.60 for the case $\varepsilon_5=40$ and also more biased than the value 0.50 that was used in Ref. 1. A final longer wSSA run with $\alpha=0.35$ and $n=10^7$ yielded

$$p(\mathbf{x}_0,25;100) = 1.747 \times 10^{-7} \pm 0.003 \times 10^{-7} \quad (95\% \text{ confidence}). \tag{18}$$

For this value of $\hat{p}$ and a one-standard uncertainty of $u^{\text{wSSA}}=0.0015\times 10^{-7}$, formula (11) yields $n^{\text{SSA}}=7.76\times 10^{12}$, which implies a gain in computational efficiency for the wSSA of $g=7.76\times 10^5$.

All the results obtained here are consistent with the values reported in Ref. 1. The added value here is the confidence intervals, which were absent in Ref. 1, and also the assurance that these results were obtained in a computationally efficient way. We should note that the results obtained here are probably more accurate than would be required in practice, e.g., if we were willing to give up one decimal of accuracy in result (18), then the value of n used to get that result could be reduced from 10^7 to 10^5, which would translate into a 100-fold reduction in the wSSA's computing time.

VI. FIRST PASSAGE TIME THEORY: STABLE STATE TRANSITIONS

Rare events in a stochastic context have traditionally been studied in terms of mean first passage times. The time $T(\mathbf{x}_0,\mathcal{E})$ required for the system, starting in state $\mathbf{x}_0$, to first reach some state in the set $\mathcal{E}$ is a random variable, and its mean $\langle T(\mathbf{x}_0,\mathcal{E})\rangle$ is often of interest. Since the cumulative distribution function $F(t;\mathbf{x}_0,\mathcal{E})$ of $T(\mathbf{x}_0,\mathcal{E})$ is, by definition, the probability that $T(\mathbf{x}_0,\mathcal{E})$ will be less than or equal to t, it follows from Eq. (3) that

$$F(t;\mathbf{x}_0,\mathcal{E}) = p(\mathbf{x}_0,\mathcal{E};t). \tag{19}$$

Therefore, since the derivative of $F(t;\mathbf{x}_0,\mathcal{E})$ with respect to t is the probability density function of $T(\mathbf{x}_0,\mathcal{E})$, the mean of the first passage time $T(\mathbf{x}_0,\mathcal{E})$ is given by

$$\langle T(\mathbf{x}_0,\mathcal{E})\rangle = \int_0^\infty t\left(\frac{dp(\mathbf{x}_0,\mathcal{E};t)}{dt}\right)dt = \int_0^\infty (1-p(\mathbf{x}_0,\mathcal{E};t))dt, \tag{20}$$

where the last step follows from an integration by parts.

In light of this close connection between the mean first passage time $\langle T(\mathbf{x}_0,\mathcal{E})\rangle$ and the probability $p(\mathbf{x}_0,\mathcal{E};t)$ that the wSSA aims to estimate, it might be thought that the wSSA also provides an efficient way to estimate $\langle T(\mathbf{x}_0,\mathcal{E})\rangle$. But that turns out not to be so. The reason is that, in order to compute $\langle T(\mathbf{x}_0,\mathcal{E})\rangle$ from Eq. (20), we must compute $p(\mathbf{x}_0,\mathcal{E};t)$ for times t that are on the order of $\langle T(\mathbf{x}_0,\mathcal{E})\rangle$. But for a truly rare event that time will be very large, and since the wSSA does not shorten the elapsed time t, it will not be feasible to make runs with the wSSA for that long a time.

From a practical point of view though, it seems likely that a knowledge of the very small value of $p(\mathbf{x}_0,\mathcal{E};t)$ for reasonable values of t might be just as useful as a knowledge of the very large value of $\langle T(\mathbf{x}_0,\mathcal{E})\rangle$. In other words, in practice it may be just as helpful to know how likely it is for the rare event $\mathbf{x}_0\rightarrow\mathcal{E}$ to happen within a time frame t of practical interest as to know how long a time on average we would have to wait in order to see the event occur. To the extent that that is true, the inability of the wSSA to accurately estimate $\langle T(\mathbf{x}_0,\mathcal{E})\rangle$ will not be a practical drawback.

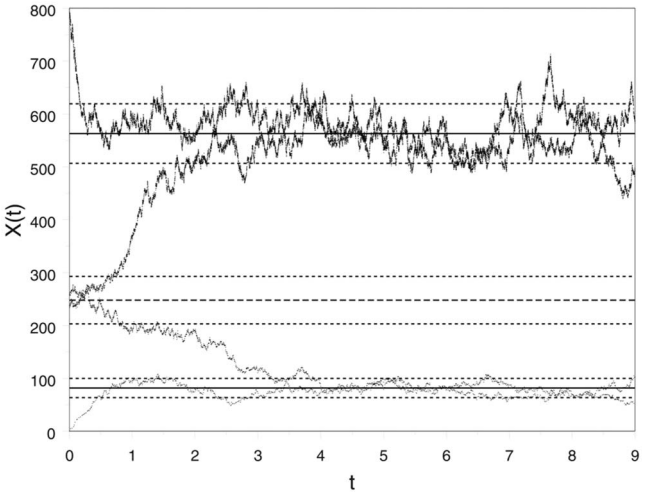

FIG. 3. Four SSA runs of the Schlögl reaction set (21) using the parameter values (22) and the initial states indicated. (From Ref. 6.) The S population $X(t)$ is plotted out here after every fifth reaction event. Starting values below the barrier region between $x=200$ and $x=300$ tend to wind up fluctuating about the lower stable state $x_1=82$, while starting values above the barrier region tend to wind up fluctuating about the upper stable state $x_2=563$. The dotted lines around the two stable states show their theoretically predicted widths, which are evidently consistent with these simulations. Spontaneous transitions between the two states will inevitably occur if the system is allowed to run long enough.

An illustration of these points is provided by the phenomenon of spontaneous transitions between the stable states of a bistable system. A well known simple model of a bistable system is the Schlögl reaction set

$$B_1 + 2S \underset{c_2}{\overset{c_1}{\rightleftarrows}} 3S, \qquad B_2 \underset{c_4}{\overset{c_3}{\rightleftarrows}} S, \tag{21}$$

where species B_1 and B_2 are assumed to be buffered so that their molecular populations N_1 and N_2 remain constant. For the parameter values

$$c_1 = 3\times 10^{-7}, \quad c_2 = 10^{-4}, \quad c_3 = 10^{-3}, \quad c_4 = 3.5, \quad N_1 = 10^5, \quad N_2 = 2\times 10^5, \tag{22}$$

the molecular population X of species S can be shown[6] to have two stable states, $x_1=82$ and $x_2=563$. Figure 3 shows four exact SSA simulations for these parameter values with four different initial states. In each of these simulation runs, X has been plotted after every five reaction events. The solid horizontal lines locate the stable states x_1 and x_2, and the adjacent dotted lines show the theoretically predicted

"widths" of those stable states. The other three horizontal lines in the figure locate the "barrier region" that separates the two stable states. (See Ref. 6 for details.) Using first passage time theory, it can be shown that the mean time for a transition from x_1 to x_2 is[6]

$$\langle T(x_1,x_2)\rangle = 5.031 \times 10^4 \tag{23}$$

and further that the associated standard deviation has practically the same value. This implies that we would usually have to run the simulations in Fig. 3 for times of order 10^4 before witnessing a spontaneous transition from x_1 to x_2, and that is a very long time on the scale of Fig. 3. But it might also be interesting to know the probability of seeing an x_1-to-x_2 transition occur within a time span that is comparable to that of Fig. 3, say, in time $t=5$.

Finding an effective importance sampling strategy to compute $p(82,563;5)$ turned out to be more difficult than we anticipated. We suspect the reason for this is the extreme sensitivity of the Schlögl reactions (21) to the values of its reaction parameters in the vicinity of the bistable configuration. For example, a 5% reduction in the value of c_3 from the value given in (22) will cause the upper steady state x_2 to disappear, while a 5% increase will cause the lower steady state x_1 to disappear. This means that in the importance sampling strategy of Eq. (4), small changes in the γ_j values can result in major changes in the dynamical structure of the system. This made finding a good biasing strategy more difficult than in the two examples considered in Sec. V. Nevertheless, we found that taking $\gamma_3=\alpha$ and $\gamma_4=1/\alpha$ with $\alpha=1.05$ produced the following estimate with $n=4\times10^7$ runs:

$$p(82,563;5) = 4.56 \times 10^{-7} \pm 0.25 \times 10^{-7} \quad (95\%\ \text{confidence}). \tag{24}$$

For this value of $\hat{p}$ and a one-standard uncertainty of $u^{\text{wSSA}}=0.125\times10^{-7}$, formula (11) yields $n^{\text{SSA}}=2.9\times10^9$. Dividing that by $n^{\text{wSSA}}=4\times10^7$ gives a gain in computational efficiency of $g=73$.

Results (23) and (24) refer to the same transition $x_1 \rightarrow x_2$, and both results are informative but in different ways. However, there does not appear to be a reliable procedure for inferring either of these results from the other; in particular, the wSSA result (24) is a new result, not withstanding the known result (23). We hope to explore more fully the problem of finding optimal wSSA weighting strategies for bistable systems in a future publication.

VII. CONCLUSIONS

The numerical results reported in Secs. V and VI support our expectation that the refinements to the original wSSA[1] made possible by the variance computation significantly improve the algorithm: The benefit of being able to quantify the uncertainty in the wSSA's estimate of $p(\mathbf{x}_0,\mathcal{E};t)$ is obvious. And having an unambiguous measure of the optimality of a given set of values of the importance sampling parameters $\{\gamma_1,\ldots,\gamma_M\}$ makes possible the task of minimizing that uncertainty. But much work remains to be done in order to develop a practical, systematic strategy for deciding how best to parametrize the set $\{\gamma_1,\ldots,\gamma_M\}$ in terms of a smaller number of parameters, and, more generally, for deciding which reaction channels in a large network of reactions should be encouraged and which should be discouraged through importance sampling. More enlightenment on these matters will clearly be needed if the wSSA is to become easily applicable to more complicated chemical reaction networks.

We described in Sec. VI the relationship between the probability $p(\mathbf{x}_0,\mathcal{E};t)$ computed by the wSSA and the mean first passage time $\langle T(\mathbf{x}_0,\mathcal{E})\rangle$, which is the traditional way of analyzing rare events. We showed that in spite of the closeness of this relationship, if the former is very "small" and the latter is very "large," then neither can easily be inferred from the other. But in practice, knowing $p(\mathbf{x}_0,\mathcal{E};t)$ will often be just as useful, if not more useful, than knowing $\langle T(\mathbf{x}_0,\mathcal{E})\rangle$.

We conclude by commenting that, in spite of the demonstration in Sec. V of how much more efficiently the wSSA computes the probability $p(\mathbf{x}_0,\mathcal{E};t)$ than the SSA when $p(\mathbf{x}_0,\mathcal{E};t)\ll1$, it would be inaccurate and misleading to view the wSSA and the SSA as "competing" procedures which aim to do the same thing. This becomes clear when we recognize two pronounced differences between those two procedures: First, whereas the wSSA always requires the user to exercise insight and judgment in choosing an importance sampling strategy, the SSA never imposes such demands on the user. Second, whereas the SSA usually plots out the state trajectories of its runs, since those trajectories reveal how the system typically behaves in time, the trajectories of the wSSA are of no physical interest because they are artificially biased. The SSA and the wSSA really have different, but nicely complementary, goals: The SSA is concerned with revealing the *typical* behavior of the system, showing how the molecular populations of *all* the species *usually* evolve with time. In contrast, the wSSA is concerned with the *atypical* behavior of the system, and more particularly with estimating the value of a single scalar quantity: the probability that a specified event will occur within a specified limited time when that probability is very small.

ACKNOWLEDGMENTS

The authors acknowledge with thanks financial support as follows: D.T.G. was supported by the California Institute of Technology through Consulting Agreement No. 102-1080890 pursuant to Grant No. R01GM078992 from the National Institute of General Medical Sciences and through Contract No. 82-1083250 pursuant to Grant No. R01EB007511 from the National Institute of Biomedical Imaging and Bioengineering, and also from the University of California at Santa Barbara under Consulting Agreement No. 054281A20 pursuant to funding from the National Institutes of Health. M.R. and L.R.P. were supported by Grant No. R01EB007511 from the National Institute of Biomedical Imaging and Bioengineering, Pfizer Inc., DOE Grant No. DE-FG02-04ER25621, NSF IGERT Grant No. DG02-21715, and the Institute for Collaborative Biotechnologies through Grant No. DFR3A-8-447850-23002 from the U.S. Army Research Office. The content of this work is solely the responsibility of the authors and does not necessarily reflect the official

views of any of the aforementioned institutions.

APPENDIX: MONTE CARLO AVERAGING AND IMPORTANCE SAMPLING

If X is a random variable with probability density function P and f is any integrable function, then the "average of f with respect to X," or equivalently the "average of the random variable $f(X)$," can be computed as either

$$\langle f(X)\rangle = \int_{-\infty}^{\infty} f(x)P(x)dx \tag{A1}$$

or

$$\langle f(X)\rangle = \lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n} f(x^{(i)}), \tag{A2}$$

where the $x^{(i)}$ in Eq. (A2) are statistically independent samples of X. *Monte Carlo averaging* is a numerical procedure for computing $\langle f(X)\rangle$ from Eq. (A2) but using a *finite* value for n. But using a finite n renders the computation *inexact*:

$$\langle f(X)\rangle \approx \frac{1}{n}\sum_{i=1}^{n} f(x^{(i)}) \quad (n<\infty). \tag{A3}$$

To estimate the *uncertainty* associated with this approximation, we reason as follows.

Let Y be any random variable with a well-defined mean and variance, and let $Y_1,\ldots,Y_n$ be n statistically independent copies of Y. Define the random variable Z_n by

$$Z_n \equiv \frac{1}{n}\sum_{i=1}^{n} Y_i. \tag{A4}$$

This means, by definition, that a sample z_n of Z_n can be obtained by generating n samples $y^{(1)},\ldots,y^{(n)}$ of Y and then taking

$$z_n = \frac{1}{n}\sum_{i=1}^{n} y^{(i)}. \tag{A5}$$

Now take n large enough so that, by the central limit theorem, Z_n is approximately *normal*. In general, the normal random variable $\mathcal{N}(m,\sigma^2)$ with mean m and variance σ^2 has the property that a random sample s of $\mathcal{N}(m,\sigma^2)$ will fall within $\pm\gamma\sigma$ of m with probability 68% if $\gamma=1$, 95% if $\gamma=2$, and 99.7% if $\gamma=3$. (For more on normal confidence interval theory, see the article by Welch.[7]) This implies that s will "estimate the mean" of $\mathcal{N}(m,\sigma^2)$ to within $\pm\gamma\sigma$ with those respective probabilities, a statement that we can write more compactly as $m\approx s\pm\gamma\sigma$. In particular, since Z_n is approximately normal, we may estimate its mean as

$$\langle Z_n\rangle \approx z_n \pm \gamma\sqrt{\mathrm{var}\{Z_n\}}. \tag{A6}$$

It is not difficult to prove that the mean and variance of Z_n as defined in Eq. (A4) can be computed in terms of the mean and variance of Y by

$$\langle Z_n\rangle = \langle Y\rangle \quad \text{and} \quad \mathrm{var}\{Z_n\} = \frac{\mathrm{var}\{Y\}}{n}. \tag{A7}$$

With Eqs. (A7) and (A5), we can rewrite the estimation formula (A6) as

$$\langle Y\rangle \approx \frac{1}{n}\sum_{i=1}^{n} y^{(i)} \pm \gamma\sqrt{\frac{\mathrm{var}\{Y\}}{n}}. \tag{A8}$$

This formula is valid for any random variable Y with a well-defined mean and variance provided n is sufficiently large (so that normality is approximately achieved).

Setting $Y=f(X)$ in Eq. (A8), we obtain

$$\langle f(X)\rangle \approx \frac{1}{n}\sum_{i=1}^{n} f(x^{(i)}) \pm \gamma\sqrt{\frac{\mathrm{var}\{f(X)\}}{n}}. \tag{A9}$$

This formula evidently quantifies the uncertainty in the estimate (A3). Again, the values $\gamma=1,2,3$ correspond to respective "confidence intervals" of 68%, 95%, and 99.7%. But formula (A9) as it stands is not useful in practice because we do not know $\mathrm{var}\{f(X)\}$. It is here that we indulge in a bit of bootstrapping logic: We *estimate*

$$\mathrm{var}\{f(X)\} \approx \frac{1}{n}\sum_{i=1}^{n} (f(x^{(i)}))^2 - \left(\frac{1}{n}\sum_{i=1}^{n} f(x^{(i)})\right)^2. \tag{A10}$$

This estimate evidently makes the assumption that n is already large enough that the n-sample first and second moments of f provide reasonably accurate estimates of $\langle f\rangle$ and $\langle f^2\rangle$. In practice, we need to test this assumption by demanding "reasonable closeness" among several n-run computations of the right hand side of Eq. (A10). *Only* when n is large enough for that to be so can we reliably invoke formulas (A9) and (A10) to infer an estimate of $\langle f(X)\rangle$ and an estimate of the uncertainty in that estimate from the two sums $\Sigma_{i=1}^{n} f(x^{(i)})$ and $\Sigma_{i=1}^{n}(f(x^{(i)}))^2$.

The most obvious way to decrease the size of the uncertainty term in Eq. (A9) is to increase n; indeed, in the limit $n\to\infty$, Eq. (A9) reduces to the exact formula (A2). But the time available for computation usually imposes a practical upper limit on n. However, we could also make the uncertainty term in Eq. (A9) smaller if we could somehow decrease the variance. Several "variance-reducing" strategies with that goal have been developed, and one that has proved to be effective in many scientific applications is called importance sampling.

Importance sampling arises from the fact that we can write Eq. (A1) as

$$\begin{aligned}\langle f(X)\rangle &= \int_{-\infty}^{\infty} f(x)P(x)\left(\frac{Q(x)}{Q(x)}\right)dx \\ &= \int_{-\infty}^{\infty}\left(\frac{f(v)P(v)}{Q(v)}\right)Q(v)dv,\end{aligned} \tag{A11}$$

where Q is the probability density function of some new random variable V. Defining still another random variable $g(V)$ by

$$g(V) \equiv \frac{f(V)P(V)}{Q(V)}, \tag{A12}$$

it follows from Eq. (A11) that

$$\langle g(V)\rangle = \langle f(X)\rangle. \tag{A13}$$

But although the two random variables $f(X)$ and $g(V)$ have the same *mean*, they will *not* generally have the same *variance*. In fact, *if we choose the function* $Q(v)$ *so that it varies with* v *in roughly the same way that* $f(v)P(v)$ *does*, then the sample values of $g(V)$ will not show as much variation as the sample values of $f(X)$. That would imply that

$$\mathrm{var}\{g(V)\} < \mathrm{var}\{f(X)\}. \tag{A14}$$

In that case, we will get a more accurate estimate of $\langle f(X)\rangle$ if we use, instead of Eq. (A9),

$$\langle f(X)\rangle = \langle g(V)\rangle \approx \frac{1}{n}\sum_{i=1}^{n} g(v^{(i)}) \pm \gamma\sqrt{\frac{\mathrm{var}\{g(V)\}}{n}}, \tag{A15}$$

where

$$\mathrm{var}\{g(V)\} \approx \frac{1}{n}\sum_{i=1}^{n} (g(v^{(i)}))^2 - \left(\frac{1}{n}\sum_{i=1}^{n} g(v^{(i)})\right)^2. \tag{A16}$$

Of course, if one is not careful in selecting the function Q, the inequality in Eq. (A14) could go the other way, and Eq. (A15) would then show a *larger* uncertainty than Eq. (A9). The key to having Eq. (A14) hold is to choose the function $Q(v)$ so that it tends to be large (small) where $f(v)P(v)$ is large (small). When that is so, generating samples $v^{(i)}$ according to Q will sample the real axis most heavily in those "important" regions where the integrand in Eq. (A1) is large. But at the same time, Q must be simple enough that it is not too difficult to generate those samples.

In practice, once a functional form for Q has been chosen, one or more parameters in Q are varied in a series of test runs to find the values that minimize variance (A16). Then a final run is made using the minimizing parameter values and as large a value of n as time will allow to get the most accurate possible estimate of $\langle f(X)\rangle$.

The connection of the foregoing general theory to the application considered in the main text can be roughly summarized by the following correspondences:

$X \leftrightarrow$ an unbiased (SSA) state trajectory,

$f(X) \leftrightarrow$ statistical weight of an unbiased trajectory,

$V \leftrightarrow$ a biased (wSSA) state trajectory,

$g(V) \leftrightarrow$ statistical weight of a biased trajectory,

$$\langle f(X)\rangle = \langle g(V)\rangle \leftrightarrow p(\mathbf{x}_0, \mathcal{E}; T),$$

$$\frac{P(v)}{Q(v)} \leftrightarrow w_k = \prod_{\substack{\text{all reaction events}\\ \text{comprising trajectory } k}} \frac{a_j/a_0}{b_j/b_0},$$

$$\sum_{k=1}^{n} g(v^{(k)}) \leftrightarrow m_n^{(1)}, \quad \sum_{k=1}^{n} (g(v^{(k)}))^2 \leftrightarrow m_n^{(2)}.$$

[1] H. Kuwahara and I. Mura, J. Chem. Phys. **129**, 165101 (2008).

[2] The computation of σ^2 in Eq. (7) evidently involves taking the difference between two usually large and, in the best of circumstances, nearly equal numbers. This can give rise to numerical inaccuracies. Since, with $\mu_m \equiv n^{-1}\Sigma_{k=1}^{n} w_k^m$, it is so that $\mu_2 - \mu_1^2$ is mathematically identical to $n^{-1}\Sigma_{k=1}^{n}(w_k - \mu_1)^2$, the form of the latter as a sum of non-negative numbers makes it less susceptible to numerical inaccuracies. Unfortunately, using this more accurate formula is much less convenient than formula (7), whose two sums can be computed on the fly without having to save the w_k values. But unless the two sums in Eq. (7) are computed with sufficiently high numerical precision, use of the alternate formula is advised.

[3] See, for instance, J. V. Sengers, D. T. Gillespie, and J. J. Perez-Esandi, Physica **90A**, 365 (1978); D. T. Gillespie, J. Opt. Soc. Am. A **2**, 1307 (1985).

[4] Result (9a) for the uncertainty when *no* importance sampling is used can also be deduced through the following line of reasoning: Abbreviating $p(\mathbf{x}_0, \mathcal{E}; t) \equiv p$, the n runs are analogous to n tosses of a coin that have probability p of being successful. We know from elementary statistics that the number of successful runs should then be the *binomial* (or Bernoulli) random variable with mean np and variance $np(1-p)$. When n is very large, that binomial random variable can be approximated by the normal random variable with the same mean and variance. Multiplying that random variable by n^{-1} gives the *fraction* of the n runs that are successful. Random variable theory tells us that it too will be (approximately) normal but with mean $n^{-1}p = p/n$ and variance $(n^{-1})^2 np(1-p) = p(1-p)/n$, and hence standard deviation $\sqrt{p(1-p)/n}$. The latter, with $p = m_n/n$, is precisely uncertainty (9a). Essentially this argument was given in Appendix B of Ref. 1. But there is apparently no way to generalize this line of reasoning to the case where the weights of the successful runs are not all unity; hence the need for the procedure described in the text.

[5] See, for instance, C. W. Gardiner, *Handbook of Stochastic Methods for Physics, Chemistry and the Natural Sciences* (Springer-Verlag, Berlin, 1985), pp. 238–240.

[6] D. T. Gillespie, *Markov Processes: An Introduction for Physical Scientists* (Academic, New York, 1992), pp. 520–529.

[7] P. D. Welch, in *The Computer Performance Modeling Handbook*, edited by S. Lavenberg (Academic, New York, 1983), pp. 268–328.