5.11
STOCHASTIC CHEMICAL KINETICS

Daniel T. Gillespie
Dan T Gillespie Consulting, 30504 Cordoba Place, Castaic, CA 91384

The time evolution of a well-stirred chemically reacting system is traditionally described by a set of coupled, first-order, ordinary differential equations. Obtained through heuristic, phenomenological reasoning, these equations characterize the evolution of the molecular populations as a *continuous, deterministic* process. But a little reflection reveals that the system actually possesses neither of those attributes: Molecular populations are whole numbers, and when they change they always do so by discrete, integer amounts. Furthermore, in excusing ourselves from the arduous task of tracking the positions and velocities of all the molecules in the system, which we hope to justify on the grounds that the system is "well-stirred", we preclude a deterministic description of the system's evolution; because, a knowledge of the system's current molecular populations is not by itself sufficient to predict with certainty the future molecular populations. Just as rolled dice are essentially random or "stochastic" when we do not precisely track their positions and velocities and all the forces acting on them, so is the time evolution of a well-stirred chemically reacting system for all practical purposes stochastic. That said, discreteness and stochasticity are usually not noticeable in chemical systems of "test-tube" size or larger, and for most such systems the traditional continuous deterministic description seems to be adequate. But if the molecular populations of some reactant species are very small, as is often the case for instance in cellular systems in biology, discreteness and stochasticity can sometimes play an important role. Whenever that happens, the ordinary differential equations approach will not be able to accurately describe the true behavior of the system.

Stochastic chemical kinetics attempts to describe the time evolution of a well-stirred chemically reacting system as an overtly *discrete*, *stochastic* process, evolving in real (continuous) time. And it tries to do this in a way that accurately reflects how chemical reactions physically occur at the molecular level. This article will outline the theoretical foundations of stochastic chemical kinetics, and then derive and interrelate its principle equations and

S. Yip (ed.),
Handbook of Materials Modeling, 1735–1752.

computational methods. It will also show how it happens that the resulting discrete, stochastic description usually gives way to the traditional continuous, deterministic description in a special limiting approximation.

1. Microphysical Foundations of Stochastic Chemical Kinetics

We consider a well-stirred system of molecules of N chemical species $\{S_1, \ldots, S_N\}$, which interact through M chemical reaction channels $\{R_1, \ldots, R_M\}$. We assume the system to be confined to a constant volume Ω, and to be in thermal (but not necessarily chemical) equilibrium at some constant absolute temperature T. We let $X_i(t)$ denote the number of molecules of species S_i in the system at time t. Our goal is to estimate, as best we can, the state vector $\mathbf{X}(t) = (X_1(t), \ldots, X_N(t))$, given that the system was in state $\mathbf{X}(t_0) = \mathbf{x}_0$ at some initial time $t_0 < t$.[1]

Each reaction channel R_j is assumed to be "elemental" in the sense that it describes a distinct physical event which happens essentially instantaneously. This assumption restricts us to two general types of reaction: *Unimolecular* reactions of the form $S_i \rightarrow$ product(s); and *bimolecular* reactions of the form $S_i + S_{i'} \rightarrow$ product(s), where in the latter i' may or may not be the same as i.[2]

A given reaction channel R_j is characterized mathematically by two quantities. The first is its *state-change vector* $\boldsymbol{\nu_j} = (\nu_{1j}, \ldots, \nu_{Nj})$, where ν_{ij} is defined to be the change in the S_i molecular population caused by one R_j reaction event; thus, if the system is in state $\mathbf{x}$ and an R_j reaction occurs, the system immediately jumps to state $\mathbf{x} + \boldsymbol{\nu_j}$. The two-dimensional array $\{\nu_{ij}\}$ is commonly known as the *stoichiometric matrix*. Its elements are practically always confined to the values 0, ± 1 and ± 2.

The other defining quantity for reaction channel R_j is its *propensity function* a_j. It is defined as follows:

$$a_j(\mathbf{x})\,\mathrm{d}t \stackrel{\Delta}{=} \text{the } \textit{probability}, \text{ given } \mathbf{X}(t) = \mathbf{x}, \text{ that one } R_j \text{ reaction will occur somewhere inside } \Omega \text{ in the next } \textit{infinitesimal} \text{ time interval } [t, t + \mathrm{d}t). \tag{1}$$

[1] Boldface variables will always be understood here to be N-component vectors, with the components corresponding to the N chemical species in the system.

[2] A set of three elemental reactions of the form $S_1 + S_2 \rightleftarrows S_4$ and $S_3 + S_4 \rightarrow S_5$ can often be regarded as the single *trimolecular* reaction $S_1 + S_2 + S_3 \rightarrow S_5$ if the first two reactions are much faster than the third. But this is always an *approximation*.

This definition might be said to be the *fundamental premise* of stochastic chemical kinetics, because everything else follows from it. It is important to recognize that this probabilistic definition has a solid basis in physical theory, more solid in fact than the reasoning that is traditionally used to justify the deterministic differential equations mentioned earlier. Since the microphysical basis of Eq. (1) ultimately determines the forms of the propensity functions, it is appropriate to describe it briefly here.

If R_j is the *unimolecular* reaction $S_1 \rightarrow$ product(s), the underlying physics, which might be quantum mechanical, generally dictates the existence of some constant which we shall call c_j such that $c_j\,\mathrm{d}t$ gives the *probability* that any *particular* S_1 molecule will so react in the next infinitesimal time $\mathrm{d}t$. If there is currently a finite number x_1 of S_1 molecules in the system, we can take $\mathrm{d}t$ to be so small that no more than one of them will undergo that reaction in the next $\mathrm{d}t$. This allows us to invoke the addition law of probability theory for mutually exclusive events, and so calculate the probability for *any* S_1 molecule in the system to undergo the R_j reaction by simply summing the individual reaction probabilities. That sum gives $x_1 \times c_j \mathrm{d}t$, from which we may conclude that the propensity function in Eq. (1) is $a_j(\mathbf{x}) = c_j x_1$.

If R_j is a *bimolecular* reaction of the form $S_1 + S_2 \rightarrow$ product(s), stochasticity manifests itself in two ways, both stemming from the fact that we do not know the exact position and velocity of any molecule in the system: First, we can predict only the probability that an S_1 molecule and an S_2 molecule will collide in the next $\mathrm{d}t$. And second, we can predict only the probability that such a collision will actually produce an R_j reaction. Consider a randomly chosen pair of S_1 and S_2 molecules. The assumption of thermal equilibrium implies that the S_2 molecule will see the S_1 molecule moving with an average relative speed $\bar{v}_{12} = \sqrt{8k_\mathrm{B}T/\pi m_{12}}$, where k_B is Boltzmann's constant and $m_{12} = m_1 m_2/(m_1 + m_2)$. Denote the effective collision cross section of the molecular pair by σ_{12} (which would equal $\pi(r_1 + r_2)^2$ if the molecules were hard spheres with radii r_1 and r_2). In the next infinitesimal time $\mathrm{d}t$, the S_1 molecule will sweep out relative to the S_2 molecule an infinitesimally small "collision volume" of size $(\bar{v}_{12}\,\mathrm{d}t)\sigma_{12}$ – so called because if the center of the S_2 molecule happens to lie inside that volume then the two molecules will collide in the next $\mathrm{d}t$. (We take $\mathrm{d}t$ to be so small that there is virtually no chance that the collision will be preempted by an earlier collision with some third molecule.) By our assumption that the system is "well-stirred" – a condition that can be secured either by an externally driven stirrer or by the inevitable self-stirring effects of the many non-reactive (bounce-off) molecular collisions that typically occur in such a system – the *probability* that the center of the S_2 molecule will lie inside the collision volume is just the ratio of that volume to the total system volume: $(\bar{v}_{12}\mathrm{d}t)\sigma_{12}/\Omega$. This ratio is therefore the probability that the pair will *collide* in the next $\mathrm{d}t$. Denoting by p_j the probability that a *colliding* S_1-S_2 molecular pair will actually

react according to R_j, we conclude by the multiplication law of probability theory that

$$\frac{(\bar{v}_{12}\mathrm{d}t)\sigma_{12}}{\Omega} \times p_j = \left(\frac{\bar{v}_{12}\sigma_{12}p_j}{\Omega}\right)\mathrm{d}t \triangleq c_j\mathrm{d}t \tag{2}$$

gives the probability that a randomly chosen S_1-S_2 molecular pair will und-ergo the R_j reaction in the next $\mathrm{d}t$. Now taking $\mathrm{d}t$ to be so small that no more than one of the x_1x_2 S_1-S_2 pairs in the system will react in the next $\mathrm{d}t$, we can invoke the addition law of probability for mutually exclusive events to compute the probability for *some* pair to so react as $x_1x_2 \times c_j\mathrm{d}t$. Thus we conclude that the propensity function in Eq. (1) is $a_j(\mathbf{x}) = c_jx_1x_2$. If this bimolecular reaction had instead been $S_1 + S_1 \rightarrow$ product(s), we would have reckoned the number of *distinct* S_1 molecular pairs to be $x_1(x_1 - 1)/2$, and so obtained for the propensity function $a_j(\mathbf{x}) = c_j(1/2)x_1(x_1 - 1)$, which properly vanishes if there is only one S_1 molecule.

The foregoing analysis shows two things: First, an elemental reaction channel R_j can indeed be described by a function $a_j(\mathbf{x})$ in the manner prescribed by Eq. (1). And second, $a_j(\mathbf{x})$ can usually be written as the product of some constant c_j, called the *specific reaction probability rate constant*, times the number of distinct combinations of R_j reactant molecules that are available when the system's state is $\mathbf{x}$. Our subsequent work here will depend critically on the first point, but will tolerate considerable variance with respect to the second; hence, we shall be less concerned here with the forms of the propensity functions than with the fact that they exist and satisfy Eq. (1).

But we should note in passing that the task of evaluating c_j entirely from first principles is a very challenging one. An interesting result for bimolecular reactions emerges in the idealized case in which the colliding molecules will react if and only if the kinetic energy associated with the component of their relative velocity along their line of centers at contact exceeds some threshold value ε^*; in that case, it can be proved from elementary kinetic theory that the *conditional reaction probability* p_j in Eq. (2) is given by $p_j = -\varepsilon^*/k_\mathrm{B}T$, thus providing a physically transparent interpretation of the familiar Arrhenius factor.[3] If it were also the case that the reaction can occur only if the point of collisional contact between the two molecules lies inside specific solid angles ω_1 on molecule 1 and ω_2 on molecule 2, then in the absence of any orienting forces p_j would contain the additional probability factors $(\omega_1/4\pi)$ and $(\omega_2/4\pi)$.

It turns out that c_j for a *unimolecular* reaction is numerically equal to the *reaction rate constant* k_j of conventional deterministic chemical kinetics, while c_j for a *bimolecular* reaction is equal to k_j/Ω if the reactants are

[3] See R. Present, *Kinetic Theory of Gases* (McGraw-Hill, New York, 1958), and D. Gillespie, *Physica A* 188, 404–425, 1992.

different species, or $2k_j/\Omega$ if they are the same. Contemplating this result by itself, one might be tempted to conclude that the fundamental premise (1) and the mathematical forms of the propensity functions all follow from some simple heuristic "stochastic extrapolation" of the mass-action equations of deterministic chemical kinetics. But the foregoing analysis shows that that is not the case: The existence and forms of the propensity functions are rooted in the realities of molecular dynamics. The equations of stochastic chemical kinetics cannot be derived in a logically rigorous way from the equations of deterministic chemical kinetics; rather, as we shall see later, the derivation goes the other way.

In what follows, we shall simply assume that the propensity functions $a_j(\mathbf{x})$, like the state-change vectors $\boldsymbol{\nu}_j$, are all given.

2. The Chemical Master Equation

Although the probabilistic nature of Eq. (1) precludes making an exact prediction of $\mathbf{X}(t)$ given that $\mathbf{X}(t_0) = \mathbf{x}_0$ for any $t > t_0$, we might reasonably hope to infer the *probability*

$$P(\mathbf{x}, t \mid \mathbf{x}_0, t_0) \stackrel{\Delta}{=} \text{Prob}\left\{\mathbf{X}(t) = \mathbf{x}, \text{ given } \mathbf{X}(t_0) = \mathbf{x}_0\right\}. \tag{3}$$

In technical terms, $P(\mathbf{x}, t \mid \mathbf{x}_0, t_0)$ is the "probability density function" of the time-dependent "random variable" $\mathbf{X}(t)$, and $\mathbf{X}(t)$ in turn is, by virtue of the dynamics prescribed by Eq. (1), a "jump Markov process".[4]

It is not difficult to deduce a time-evolution equation for the function (3) by using the laws of probability theory to write $P(\mathbf{x}, t + \mathrm{d}t \mid \mathbf{x}_0, t_0)$ as the *sum* of the probabilities of all the mutually exclusive ways in which the system could evolve from state $\mathbf{x}_0$ at time t_0 to state $\mathbf{x}$ at time $t + \mathrm{d}t$, via *specified* states at time t:

$$P(\mathbf{x}, t + \mathrm{d}t \mid \mathbf{x}_0, t_0) = P(\mathbf{x}, t \mid \mathbf{x}_0, t_0) \times \left[1 - \sum_{j=1}^{M} \left(a_j(\mathbf{x})\mathrm{d}t\right)\right] + \sum_{j=1}^{M} P(\mathbf{x} - \boldsymbol{\nu}_j, t \mid \mathbf{x}_0, t_0) \times \left(a_j(\mathbf{x} - \boldsymbol{\nu}_j)\mathrm{d}t\right).$$

[4] A *stochastic process* – a random variable that depends on time – is said to be *Markov* if its future values depend on its past values only through its present value. (A Markov *process* is distinguished from a Markov *chain* by the fact that time is a real or continuous variable in a process and an integer variable in a chain.) A *jump* Markov process changes discontinuously at isolated instants in time, and remains constant between such jumps. There are also *continuous* Markov processes, which evolve in a way that is mathematically continuous but often not differentiable.

Here, the first term on the right is the probability that the system is already in state $\mathbf{x}$ at time t and then no reaction of any kind occurs in $[t, t+\mathrm{d}t)$. And the generic second term is the probability that the system is one R_j reaction removed from state $\mathbf{x}$ at time t and then one R_j reaction occurs in $[t, t+\mathrm{d}t)$. That these $M+1$ routes to the final state $\mathbf{x}$ are mutually exclusive and collectively exhaustive is ensured by taking dt to be so small that no more than one reaction of any kind can occur in $[t, t+\mathrm{d}t)$. Subtracting $P(\mathbf{x}, t \mid \mathbf{x}_0, t_0)$ from both sides of the above equation, dividing through by $\mathrm{d}t$, and then taking the limit $\mathrm{d}t \to 0$, we obtain what is know as the *chemical master equation* (CME):

$$\frac{\partial P(\mathbf{x}, t \mid \mathbf{x}_0, t_0)}{\partial t} = \sum_{j=1}^{M} \big[a_j(\mathbf{x} - \boldsymbol{\nu}_j) P(\mathbf{x} - \boldsymbol{\nu}_j, t \mid \mathbf{x}_0, t_0) - a_j(\mathbf{x}) P(\mathbf{x}, t \mid \mathbf{x}_0, t_0)\big]. \tag{4}$$

In principle, the CME completely determines the function $P(\mathbf{x}, t \mid \mathbf{x}_0, t_0)$. But a closer inspection of Eq. (3) reveals that it is actually a *set* of coupled, ordinary differential equations in t; in fact, there is one equation for each possible value $(0,1,2,\ldots)$ of each of the M components of the variable $\mathbf{x}$ – roughly as many equations as there are combinations of molecules in the system! So it is perhaps not surprising that the CME can be solved analytically for only a very few very simple systems, and numerical solutions are usually prohibitively difficult in other cases.

One might hope, less ambitiously, to learn something from the CME about the behavior of functional *averages* like $\langle f(\mathbf{X}(t))\rangle \equiv \sum_{\mathbf{x}} f(\mathbf{x}) P(\mathbf{x}, t \mid \mathbf{x}_0, t_0)$, but this too turns out to be practically impossible if any of the reaction channels are bimolecular. For example, it can be proved from Eq. (4) that

$$\frac{\mathrm{d}\langle \mathbf{X}(t)\rangle}{\mathrm{d}t} = \sum_{j=1}^{M} \boldsymbol{\nu}_j \left\langle a_j(\mathbf{X}(t))\right\rangle. \tag{5}$$

Now, if all the reactions were monomolecular, the propensity functions would all be linear in the state variables, and we would have $\left\langle a_j(\mathbf{X}(t))\right\rangle = a_j(\langle \mathbf{X}(t)\rangle)$. Equation (5) would then become a closed ordinary differential equation for the first moment or *mean* $\langle \mathbf{X}(t)\rangle$. But if any reaction is bimolecular, the right hand side of Eq. (5) will contain at least one quadratic moment of the form $\langle X_i(t) X_{i'}(t)\rangle$, and Eq. (5) would then be merely the first of an infinite, openended *set* of equations for *all* the moments.

In the hypothetical case in which there are *no fluctuations*, i.e., if $\mathbf{X}(t)$ were a deterministic or *sure* process, we would have $\langle f(\mathbf{X}(t))\rangle = f(\mathbf{X}(t))$ for all functions f. Equation (5) would then reduce to

$$\frac{\mathrm{d}\mathbf{X}(t)}{\mathrm{d}t} = \sum_{j=1}^{M} \boldsymbol{\nu}_j a_j(\mathbf{X}(\mathbf{t})). \tag{6}$$

This is just the well known *reaction rate equation* (RRE) of traditional *deterministic* chemical kinetics – a set of coupled first-order ordinary differential equations for the components $X_i(t)$, which are now *continuous* (real) variables. The RRE is more commonly written in terms of the concentration variable $\mathbf{Z}(t) \stackrel{\Delta}{=} \mathbf{X}(t)/\Omega$, but that simple scalar transformation is inconsequential for our purposes here.

Although the foregoing analysis shows that the deterministic RRE (6) would be valid if all fluctuations were simply ignored, it does not tell us how or why the fluctuations might ever be "ignorable". We shall later prove that the RRE can actually be *derived* from Eq. (1) through a series of physically transparent approximating assumptions.

3. The Stochastic Simulation Algorithm

Since the CME (4) is practically never useful for calculating the probability density function of $\mathbf{X}(t)$, we need another approach. Let us look for a way to construct a *numerical realization* of $\mathbf{X}(t)$, i.e., a simulated trajectory of $\mathbf{X}(t)$ vs. t. Note that this is *not* the same as solving the CME numerically; however, much the same effect can be achieved by either histogramming or averaging the results of many realizations. For example, the nth moment $\langle X_i^n(t_1)\rangle$, which would be given in terms of the solution to the CME as $\sum_{\mathbf{x}} x_i^n P(\mathbf{x}, t_1 \mid \mathbf{x}_0, t_0)$, can also be estimated by generating L trajectories $\mathbf{x}^{(1)}(t), \ldots, \mathbf{x}^{(L)}(t)$ from state $\mathbf{x}_0$ at time t_0 to time t_1, and then computing $L^{-1} \sum_{l=1}^{L} \left[x_i^{(l)}(t_1)\right]^n$. This estimate will have an associated uncertainty which decreases with the number of realizations L like $L^{-1/2}$. In practice, it is often found that as few as two or three simulated trajectories can convey as good a picture of the dynamical behavior of $\mathbf{X}(t)$ as would be afforded by an exact expression for $P(\mathbf{x}, t \mid \mathbf{x}_0, t_0)$.

The key to generating simulated trajectories of $\mathbf{X}(t)$ is actually not the CME, but rather a new probability function, $p(\tau, j \mid \mathbf{x}, t)$, which is defined as follows:

$$p(\tau, j \mid \mathbf{x}, t)\, \mathrm{d}\tau \stackrel{\Delta}{=} \text{the probability, given } \mathbf{X}(t) = \mathbf{x}, \text{ that the } \textit{next} \text{ reaction in the system will occur in the infinitesimal time interval } [t + \tau, t + \tau + \mathrm{d}\tau), \textit{and} \text{ will be an } R_j \text{ reaction.} \tag{7}$$

Formally, this function is the joint probability density function of the two random variables "time to the next reaction" (τ) and "index of the next reaction" (j), given that the system is currently in state $\mathbf{x}$. If we can derive an analytical expression for this function, we could use Monte Carlo techniques

to generate simultaneous samples of τ and j, and that would enable us to advance the system in time from one reaction event to the next. Happily, it turns out that we can do all this fairly easily, and without having to make any approximations.

To derive an analytical expression for $p(\tau, j \mid \mathbf{x}, t)$, we begin by introducing yet another probability function, $P_0(\tau \mid \mathbf{x}, t)$, which is defined as the probability, given $\mathbf{X}(t) = \mathbf{x}$, that *no* reaction of any kind occurs in the time interval $[t,\ t+\tau)$. By the definition (1) and the laws of probability theory, this function satisfies

$$P_0(\tau + \mathrm{d}\tau \mid \mathbf{x}, t) = P_0(\tau \mid \mathbf{x}, t) \times \left[1 - \sum_{j'=1}^{M} (a_{j'}(\mathbf{x})\mathrm{d}\tau)\right],$$

since the right side gives the probability that no reaction occurs in $[t, t+\tau)$ and then no reaction occurs in $[t+\tau,\ \ t+\tau+\mathrm{d}\tau)$ (as usual we take the infinitesimal time span $\mathrm{d}\tau$ to be so small that it can contain no more than one reaction). A simple algebraic rearrangement of this equation and passage to the limit $\mathrm{d}\tau \to 0$ results in the differential equation

$$\frac{\mathrm{d}P_0(\tau \mid \mathbf{x}, t)}{\mathrm{d}\tau} = -a_0(\mathbf{x})\ P_0(\tau \mid \mathbf{x}, t),$$

where we have defined

$$a_0(\mathbf{x}) \triangleq \sum_{j'=1}^{M} a_{j'}(\mathbf{x}). \tag{8}$$

The solution to this differential equation for the initial condition $P_0(\tau = 0 \mid \mathbf{x}, t) = 1$ is

$$P_0(\tau \mid \mathbf{x}, t) = \exp\left(-a_0(\mathbf{x})\ \tau\right).$$

Now we observe that the probability defined in Eq. (7) can be written

$$p(\tau, j \mid \mathbf{x}, t)\,\mathrm{d}\tau = P_0(\tau \mid \mathbf{x}, t) \times (a_j(\mathbf{x})\mathrm{d}\tau),$$

since the right side gives the probability that no reactions occur in $[t,\ t+\tau)$ and then one R_j reaction occurs in $[t+\tau, t+\tau+\mathrm{d}\tau)$. When we insert the above formula for $P_0(\tau \mid \mathbf{x}, t)$ into this last equation and cancel the $\mathrm{d}\tau$'s, we obtain

$$p(\tau, j \mid \mathbf{x}, t) = a_j(\mathbf{x})\ \exp\left(-a_0(\mathbf{x})\ \tau\right), \tag{9a}$$

or equivalently

$$p(\tau, j \mid \mathbf{x}, t) = a_0(\mathbf{x}) \ \exp(-a_0(\mathbf{x})\,\tau) \times \frac{a_j(\mathbf{x})}{a_0(\mathbf{x})}. \tag{9b}$$

Equation (9a) is the desired explicit formula for the joint probability density function of τ and j. The equivalent form (9b) shows that this joint density function can be factored as the product of a τ-density function and a j-density function; more precisely, it shows that τ is an exponential random variable with mean and standard deviation $1/a_0(\mathbf{x})$, while j is a statistically independent integer random variable with point probabilities $a_j(\mathbf{x})/a_0(\mathbf{x})$. There are several exact Monte Carlo procedures for generating samples of these random variables. Perhaps the most direct is the procedure that follows by applying to each of the two probability density functions in Eq. (9b) the so-called inversion generating method:[5] Draw two random numbers r_1 and r_2 from the uniform distribution in the unit-interval, and take

$$\tau = \frac{1}{a_0(\mathbf{x})} \ln\left(\frac{1}{r_1}\right), \tag{10a}$$

$$j = \text{the smallest integer satisfying } \sum_{j'=1}^{j} a_{j'}(\mathbf{x}) > r_2\, a_0(\mathbf{x}). \tag{10b}$$

And so we arrive at the following exact procedure for constructing a numerical realization of the process $\mathbf{X}(t)$, a procedure called the *stochastic simulation algorithm* (SSA):

0. Initialize the time $t = t_0$ and the system's state $\mathbf{x} = \mathbf{x}_0$.
1. With the system in state $\mathbf{x}$ at time t, evaluate all the $a_j(\mathbf{x})$ and their sum $a_0(\mathbf{x})$.
2. Generate values for τ and j using Eqs. (10) (or an equivalent procedure).
3. Effect the next reaction by replacing $t \leftarrow t + \tau$ and $\mathbf{x} \leftarrow \mathbf{x} + \boldsymbol{\nu}_j$.
4. Record $(\mathbf{x}, t)$ as desired. Return to Step **1**, or else end the simulation.

The $\mathbf{X}(t)$ trajectory that is produced by the SSA might be thought of as a "stochastic version" of the trajectory that would be obtained by solving the RRE (6). (But note that the time step τ in the SSA is exact, and is *not* a finite approximation to some infinitesimal $\mathrm{d}t$, as is the time step in most numerical solvers for the RRE.) If it is found that every SSA-generated trajectory is practically indistinguishable from the RRE trajectory, then we may conclude that microscale randomness is negligible for this system. But if the SSA trajectories are found to deviate significantly from the RRE trajectory, then we must

[5]It can be proved that a sample x of the random variable X can be obtained from a sample r of the unit-interval uniform random variable by solving $\int_{-\infty}^{x} P(x')\mathrm{d}x' = r$, where P is the density function of X. This is known as the "inversion" generating procedure.

conclude that microscale randomness is not negligible, and the deterministic RRE does not provide an accurate description of the system's true behavior.

The SSA and the CME are logically equivalent to each other, since each is derived without approximation from premise (1). But even when the CME is completely intractable, the SSA is quite straightforward to implement. In fact, as a numerical procedure, the SSA is even simpler than the procedures that are typically used to numerically solve the RRE (6). The catch is that the SSA is often *very slow*. The source of this slowness can be traced to the factor $1/a_0(\mathbf{x})$ in Eq. (10a), which as mentioned earlier is the mean of the random variable τ: Since $a_0(\mathbf{x})$ is at least linear and more commonly quadratic in the reactant populations, $a_0(\mathbf{x})$ can be very large, and $\langle\tau\rangle$ correspondingly very small, whenever *any* reactant species is present in large numbers, and that is nearly always the case in practice.

One notable attempt to speed up the SSA is the *Gibson-Bruck procedure*, which advances the system in exact accord with the function $p(\tau, j \,|\, \mathbf{x}, t)$ in Eq. (9) but using a different scheme than Eqs. (10).[6] Although this procedure is more complicated to code than the procedure described above, it is significantly faster and more efficient for systems having many species and many reaction channels.

But *any* procedure that simulates *every* reaction event, exactly and one at a time, will inevitably be too slow for many practical applications. This prompts us to consider the possibility of giving up some of the exactness of the SSA in return for greater simulation speed.

4. Tau Leaping

One approximate accelerated simulation strategy is *tau-leaping*, which tries to advance the system by a *pre-selected* time interval τ that encompasses *more than one* reaction event. To properly accomplish that feat when the system is in state $\mathbf{x}$ at time t, we need to know how to generate sample values of the M random variables

$$K_j(\tau; \mathbf{x}, t) \stackrel{\Delta}{=} \text{the number of times reaction channel } R_j \text{ fires in } [t, t+\tau), \text{ given that } \mathbf{X}(t) = \mathbf{x} \quad (j = 1, \ldots, M). \tag{11}$$

For then, we could simply insert those sample values into the update formula

$$\mathbf{X}(t+\tau) = \mathbf{x} + \sum_{j=1}^{M} K_j(\tau; \mathbf{x}, t)\, \boldsymbol{\nu}_j \tag{12}$$

[6]For details, see M. Gibson and J. Bruck, *J. Phys. Chem.*, 104, 1876–1889, 2000.

to "leap" the system ahead by the chosen time τ. Unfortunately, that is easier said than done. In general, the M random variables (11) are statistically dependent, and it is not altogether clear even how to calculate their joint probability density function, much less generate random samples according to that function.

Suppose, however, that τ is chosen small enough that the following *Leap Condition* is satisfied: The expected state change induced by the leap is sufficiently small that no propensity function changes its value by a significant amount. In that case, we should be able to *approximate* each $K_j(\tau; \mathbf{x}, t)$ by a statistically independent *Poisson* random variable:

$$K_j(\tau; \mathbf{x}, t) \approx \mathcal{P}_j(a_j(\mathbf{x}), \tau) \; (j = 1, \ldots, M). \tag{13}$$

This is because the generic Poisson random variable $\mathcal{P}(a, \tau)$ is by definition the number of events that will occur in time τ, given that $a\,\mathrm{d}t$ is the probability that an event will occur in any infinitesimal time $\mathrm{d}t$, where a may be any positive *constant* (hence the need for the Leap Condition).[7] Therefore, if we can find a value for τ that is small enough that the Leap Condition is satisfied, yet large enough that many reaction events occur in time τ, we may indeed have a faster, albeit approximate, simulation strategy.

The practical question arises, how can we determine in advance the *largest* value of τ that is compatible with the Leap Condition? Although there is as yet no unequivocal answer to this question, the following recipe for choosing τ will approximately ensure that no propensity function is likely to change its value in the leap by more than $\varepsilon a_0(\mathbf{x})$, where ε is some pre-chosen *accuracy control parameter* satisfying $0 < \varepsilon \ll 1$: With

$$f_{jj'}(\mathbf{x}) \stackrel{\Delta}{=} \sum_{i=1}^{N} \frac{\partial a_j(\mathbf{x})}{\partial x_i} \nu_{ij'} \quad (j, j' = 1, \ldots, M) \tag{14a}$$

and

$$\left.\begin{aligned} \mu_j(\mathbf{x}) &\stackrel{\Delta}{=} \sum_{j'=1}^{M} f_{jj'}(\mathbf{x})\, a_{j'}(\mathbf{x}) \\ \sigma_j^2(\mathbf{x}) &\stackrel{\Delta}{=} \sum_{j'=1}^{M} f_{jj'}^2(\mathbf{x})\, a_{j'}(\mathbf{x}) \end{aligned}\right\} \quad (j = 1, \ldots, M), \tag{14b}$$

take[8]

$$\tau = \min_{j \in [1,M]} \left\{ \frac{\varepsilon\, a_0(\mathbf{x})}{|\mu_j(\mathbf{x})|}, \; \frac{\varepsilon^2 a_0^2(\mathbf{x})}{\sigma_j^2(\mathbf{x})} \right\}. \tag{15}$$

[7] It can be shown that the probability that the random variable $\mathcal{P}(a, \tau)$ as so defined will equal any non-negative integer n is $e^{-a\tau}(a\tau)^n \big/ n!$, and also that the *mean* and *variance* of $\mathcal{P}(a, \tau)$ are both equal to $a\tau$.

[8] For a derivation, see D. Gillespie and L. Petzold, *J. Chem. Phys.*, 119, 8229–8234, 2003.

The *explicit tau-leaping simulation procedure* thus goes as follows:

1. In state $\mathbf{x}$ at time t, and with a value chosen for ε, evaluate τ from Eq. (15).
2. For $j = 1, \ldots, M$, generate the number of firings k_j of reaction R_j in time τ as a sample of the Poisson random variable $\mathcal{P}\left(a_j(\mathbf{x}), \tau\right)$.[9]
3. Leap, by replacing $t \leftarrow t + \tau$ and $\mathbf{x} \leftarrow \mathbf{x} + \sum_{j=1}^{M} k_j \boldsymbol{\nu}_j$.

Smaller values of ε will result in a better satisfaction of the Leap Condition, and hence a leap that is more accurate, but of course shorter. In the limit $\varepsilon \to 0$, tau-leaping becomes mathematically equivalent to the SSA; however, tau-leaping will be very inefficient in that limit because all the k_j's will usually be zero, giving a very small time step without any change of state. Therefore, it is advisable to abort the above procedure after Step 1 if τ is found to be less than a few multiples of $1/a_0(\mathbf{x})$, the mean time to the next reaction, and instead use the SSA to step directly to that next reaction.

A variation on the foregoing tau-leaping procedure allows us to advance the system to the moment of the *next* firing of some *particular* reaction channel R_α, which perhaps initiates some critical sequence of events in the system. To do that, we start by computing a *tentative* τ from Eq. (15), and then computing $a_\alpha(\mathbf{x})\,\tau$, the *expected* number of R_α firings in that time τ. If $a_\alpha(\mathbf{x})\,\tau < 1$, we should *not* try to leap ahead to the next R_α reaction because that would violate the Leap Condition. But if $a_\alpha(\mathbf{x})\,\tau \geq 1$, then a leap with $k_\alpha = 1$ should be okay. In that case, we would generate the *actual* time τ to the next R_α reaction as $\tau = a_\alpha^{-1}(\mathbf{x}) \ln(1/r)$,where r is a unit-interval uniform random number. Using that value for τ, we would then generate Poisson values for all the other $k_{j\neq\alpha}$ as in Step 2, and finally effect the leap as in Step 3.

If the system happens to be "dynamically stiff" – meaning that it has widely varying dynamical modes, the fastest of which are stable – the explicit tau-leaping procedure will be computationally unstable for time steps that are larger than the fastest time scale, and that may severely restrict the size of τ. Stiffness is very common in chemical systems. Recently, an *implicit* tau-leaping procedure has been proposed which shows promise of overcoming the instability problem for stiff systems.[10]

It should be noted that tau-leaping is not as foolproof as the SSA. If one takes leaps that are too large, bad things can happen; e.g., some species populations might be driven negative. The underlying philosophy of tau-leaping is to leap over "unimportant" reaction events but not the "important" ones, and

[9]Numerical procedures for generating Poisson random numbers can be found, for instance, in W. Press, B. Flannery, S. Teukolsky, and W. Vetterling, *Numerical Recipes: The Art of Scientific Computing*, Cambridge University Press, New York, 1986.
[10]For details, see M. Rathinam, L. Petzold, Y. Cao, and D. Gillespie, *J. Chem. Phys.*, 119, 12784–12794, 2003.

in some circumstances special measures must be taken to ensure that outcome. Much more work in this area is needed.

5. The Chemical Langevin Equation

In the previous section we noted that, when the system is in state $\mathbf{x}$ at time t, if we choose a time-step Δt that is small enough that none of the propensity function values changes significantly during Δt, then the system's state at time $t + \Delta t$ can be decently approximated by

$$\mathbf{X}(t + \Delta t) \doteq \mathbf{x} + \sum_{j=1}^{M} \mathcal{P}_j\left(a_j(\mathbf{x}), \Delta t\right) \boldsymbol{\nu_j}, \tag{16}$$

where the $\mathcal{P}_j$'s are statistically independent Poisson random variables. Suppose the system admits a Δt that satisfies not only that condition, but also the condition that the expected (or mean) number of firings of each reaction channel in time Δt is $\gg 1$; i.e.,

$$a_j(\mathbf{x})\,\Delta t \gg 1 \quad \text{for all } j = 1, \ldots, M. \tag{17}$$

It will usually be possible to find such a Δt if the molecular populations of *all* reactant species are "sufficiently large". Now, it is well know that the Poisson random variable $\mathcal{P}(a, \tau)$, which has mean and variance $a\tau$, can be *approximated* when $a\tau \gg 1$ by the *normal* random variable with the same mean and variance.[11] Therefore, denoting the normal random variable with mean m and variance σ^2 by by $\mathcal{N}(m, \sigma^2)$, condition (17) allows Eq. (16) to be *further* approximated as follows:

$$\begin{aligned}\mathbf{X}(t + \Delta t) &\doteq \mathbf{x} + \sum_{j=1}^{M} \mathcal{N}_j\left(a_j(\mathbf{x})\Delta t, a_j(\mathbf{x})\Delta t\right) \boldsymbol{\nu_j}, &\text{(18a)}\\ &= \mathbf{x} + \sum_{j=1}^{M} \left(a_j(\mathbf{x})\Delta t + \sqrt{a_j(\mathbf{x})\Delta t}\mathcal{N}_j(0,1)\right) \boldsymbol{\nu_j},\end{aligned}$$

$$\mathbf{X}(t + \Delta t) \doteq \mathbf{x} + \sum_{j=1}^{M} \boldsymbol{\nu_j} a_j(\mathbf{x})\Delta t + \sum_{j=1}^{M} \boldsymbol{\nu_j}\sqrt{a_j(\mathbf{x})}\mathcal{N}_j(0,1)\sqrt{\Delta t}\,, \tag{18b}$$

where the second line invokes the fact that $\mathcal{N}(m, \sigma^2) = m + \sigma\mathcal{N}(0, 1)$.

[11] That $e^{-a\tau}(a\tau)^n/n! \approx (2\pi a\tau)^{-1/2}\exp(-(n - a\tau)^2/2a\tau)$ when $a\tau \gg 1$ follows from Stirling's approximation and the small-x approximation for $\ln(1 + x)$.

We have thus established the following result: If the system admits a *macroscopically infinitesimal* time increment $\mathrm{d}t$, defined so that during $\mathrm{d}t$ (i) no propensity function changes its value significantly yet (ii) every reaction channel fires many more times that once, then we can approximate the t to $t + \mathrm{d}t$ system evolution by

$$\mathbf{X}(t + \mathrm{d}t) \doteq \mathbf{X}(t) + \sum_{j=1}^{M} \boldsymbol{\nu}_j a_j\,(\mathbf{X}(t))\,\mathrm{d}t + \sum_{j=1}^{M} \boldsymbol{\nu}_j \sqrt{a_j\,(\mathbf{X}(t))}\,N_j(t)\sqrt{\mathrm{d}t}, \tag{19}$$

where the $N_j(t)$ are M statistically independent, temporally uncorrelated, *normal* random variables with means 0 and variances 1. Equation (19) is called the *chemical Langevin equation* (CLE). The dot over its equal sign reminds us that it is an *approximation*, valid only to the extent that $\mathrm{d}t$ is *small* enough to satisfy condition (i) and simultaneously *large* enough to satisfy condition (ii). It is usually possible to find such a $\mathrm{d}t$ if all the reactant populations are sufficiently large. But if that is not possible, Eq. (19) has no basis and should not be invoked.

The approximate character of the CLE (19) is underscored in the fact that the state vector $\mathbf{X}(t)$ therein is no longer discrete (integer-valued), but instead is continuous (real-valued); in fact, the name "Langevin" is applied because Eq. (19) has the exact mathematical form of the like-named generic equation that governs the time-evolution of any *continuous Markov process*. For the sake of completeness, two pertinent but unobvious results from the formal theory of continuous Markov processes should be noted here:[12] First, Eq. (19) can be written in the mathematically equivalent form

$$\frac{\mathrm{d}\mathbf{X}(t)}{\mathrm{d}t} \doteq \sum_{j=1}^{M} \boldsymbol{\nu}_j\, a_j\,(\mathbf{X}(t)) \;+ \sum_{j=1}^{M} \boldsymbol{\nu}_j \sqrt{a_j\,(\mathbf{X}(t))}\,\Gamma_j(t)\,. \tag{20}$$

Here, $\Gamma_j(t)$ are statistically independent "Gaussian white noise" processes satisfying $\langle \Gamma_j(t)\;\Gamma_{j'}(t')\rangle = \delta_{jj'}\,\delta(t - t')$, where the first delta function is Kronecker's and the second is Dirac's. Equation (20) is called the "white noise form" of the CLE. Second, the time evolution of $\mathbf{X}(t)$ prescribed by Eq. (19)

[12]For a proof of the equivalence of the mathematical forms (19), (20) and (21), see D. Gillespie, *Am. J. Phys.*, 64, 1246–1257, 1996.

induces a time evolution in the probability density function of $\mathbf{X}(t)$ according to the partial differential equation

$$\begin{aligned}\frac{\partial P(\mathbf{x}, t|\mathbf{x}_0, t_0)}{\partial t} \doteq & -\sum_{i=1}^{N} \frac{\partial}{\partial x_i}\left[\left(\sum_{j=1}^{M} \nu_{ij} a_j(\mathbf{x})\right) P(\mathbf{x}, t \mid \mathbf{x}_0, t_0)\right] \\ & + \frac{1}{2}\sum_{i=1}^{N} \frac{\partial^2}{\partial x_i^2}\left[\left(\sum_{j=1}^{M} \nu_{ij}^2 a_j(\mathbf{x})\right) P(\mathbf{x}, t \mid \mathbf{x}_0, t_0)\right] \\ & + \sum_{\substack{i,i'=1 \\ (i<i')}}^{N} \frac{\partial^2}{\partial x_i \partial x_{i'}}\left[\left(\sum_{j=1}^{M} \nu_{ij}\nu_{i'j} a_j(\mathbf{x})\right) P(\mathbf{x}, t \mid \mathbf{x}_0, t_0)\right].\end{aligned} \tag{21}$$

This equation is called the *chemical Fokker–Planck equation* (CFPE). Essentially, we have approximated the *jump* Markov process governed by the master Eq. (4) by the *continuous* Markov process governed by the Fokker–Planck Eq. (21).

All this somewhat complicated and possibly unfamiliar mathematics should not be allowed to obscure the genuine simplicity of the logical arguments underlying the foregoing derivation of the CLE (19): Condition (i) allowed us to infer, essentially from the fundamental premise (1), the *Poisson* approximation (16), and condition (ii) then allowed us to make the *normal* approximation (18), whence the CLE (19).[13]

Before examining some interesting theoretical implications of the CLE, we should note that it has a very practical numerical application: In either of its forms (18), the CLE enables us to approximately advance the system in time by a *macroscopically infinitesimal* time increment Δt. By virtue of condition (17), that would allow us to leap over very many individual reactions, thus producing a very substantial increase in simulation speed over the SSA. The Langevin update formula (18) is computationally more attractive than the explicit tau-leaping update formula (16) simply because normal random numbers are easier to generate than Poisson random numbers.[14] But it should be clear that the Langevin update formula (18) is really just a limiting approximation of the explicit tau-leaping update formula (16): Whenever conditions (17) hold, tau-leaping inevitably reduces to Langevin leaping.

[13] Note that this derivation of the CLE does not proceed in the *ad-hoc* manner of many Langevin equation derivations, in which the forms of the coefficients of $\Gamma_j(t)$ in Eq. (20) are simply *assumed* with an eye to obtaining some pre-conceived outcome.

[14] See the reference cited in footnote 9.

6. The Reaction Rate Equation Limit

In practice, most chemical systems contain huge numbers of molecules, and are thus well on their way to the so-called *thermodynamic limit*, in which the species populations X_i and the system volume Ω all approach infinity in such a way that the species concentrations X_i/Ω remain constant. The large molecular populations of such systems usually means that their dynamical behavior is well described by the CLE (19).

An inspection of the CLE (19) shows that it separates the state increment $\mathbf{X}(t + \mathrm{d}t) - \mathbf{X}(t)$ in a macroscopically infinitesimal time step $\mathrm{d}t$ into *two components*: a *deterministic* component proportional to $\mathrm{d}t$, and a *fluctuating* component proportional to $\sqrt{\mathrm{d}t}$. The deterministic component is evidently linear in the propensity functions, while the fluctuating component is proportional to the square root of the propensity functions. Now it happens that all propensity functions grow, in the thermodynamic limit, in *direct proportion* to the size of the system. For a unimolecular propensity function of the form $c_j x_i$ this is obvious; for a bimolecular propensity function of the form $c_j x_i x_{i'}$ this follows because c_j is inversely proportional to the system volume Ω [cf. Eq. (2)], which offsets one of the population variables. Therefore, as the thermodynamic limit is approached, the *deterministic* component of the state increment in the CLE (19) grows like the system size, whereas the *fluctuating* component grows like the *square root* of the system size. The fluctuating component thus scales, relative to the deterministic component, as the *inverse* square root of the system size. This establishes, in a logically deductive way, the conventional rule-of-thumb that *relative fluctuations in a chemically reacting system typically scale as the inverse square root of the system size*.

This scaling behavior also implies that, in the full thermodynamic limit, the fluctuating term in the CLE (19) usually becomes vanishingly small compared to the deterministic term, and hence can be dropped. The CLE therefore becomes in the full thermodynamic limit,

$$\mathbf{X}(t + \mathrm{d}t) \doteq \mathbf{X}(t) + \sum_{j=1}^{M} \boldsymbol{\nu}_j a_j\left(\mathbf{X}(t)\right) \mathrm{d}t. \tag{22}$$

This is just the conventional RRE (6). But we have now *derived* it within the theoretical framework of stochastic chemical kinetics.

Notice how our description of the system's dynamical behavior has progressed: The CME (4) and the SSA (9) describe $\mathbf{X}(t)$ as a *discrete stochastic* process. The CLE (19) and the CFPE (21) describe $\mathbf{X}(t)$ as a *continuous stochastic* process. And the RRE (22) describes $\mathbf{X}(t)$ as a *continuous deterministic* process. At each level, the description is an approximation of the description at the previous level, valid only under certain specific conditions.

One instance in which the limiting form (22) can be misleading is when the sum on the right hand side is zero, which happens whenever the system evolves to a "stable state". In such a circumstance, the fluctuating term in the CLE (19) will inevitably be larger than the deterministic term, and hence not entirely negligible. Another instance of inadequacy of the RRE concerns the long-time behavior of an open or driven system that has more than one stable state: Such a system will in fact perpetually visit all of those stable states, whereas the RRE contrarily implies that the system will go to the nearest (downhill) stable state and stay there forever. But in the many cases where the approximating assumptions leading to the RRE are warranted, the RRE provides a very efficient description of the system's temporal behavior.

7. The Chemical Kinetics Modeling Hierarchy

We conclude by summarizing the hierarchy of schemes that are available for modeling the time evolution of a chemically reacting system, proceeding from the slowest and most accurate to the fastest and most approximate.

The most exact procedure for simulating the time evolution of a chemically reacting system is *molecular dynamics* (MD), wherein the position and velocity of every molecule in the system are tracked precisely. This results in a simulation of every molecular collision that occurs in the system, not only the reactive collisions but also the non-reactive (elastic) collisions. MD is thus able to show very accurately the evolution of not only the species populations, but also their spatial distributions. But of course, this essentially exact approach requires an enormous investment of computation time and resources.

If the system is such that reactive collisions are usually separated in time by many non-reactive collisions, and the predominant effect of the latter is simply to "stir" the system, then we may back away from an MD simulation and use instead the SSA. The SSA simulates only the reactive collisions. Because it skips over all the non-reactive collisions, and also avoids computing spatial distributions (which are assumed to be uniform in the statistical sense), the SSA is computationally much faster than MD.

Tau-leaping is based on the same assumptions as the SSA, but it proceeds *approximately* from those assumptions; more specifically, it uses a special *Poisson* approximation to advance the system by a *pre-selected* time τ during which more than one reaction event may occur. The size of τ is restricted by the condition that no propensity function may change its value during τ by a "significant" amount. Whenever that condition is satisfied and at least some of the reaction channels fire very many times in τ, tau-leaping will be faster than the SSA.

A tau-leap in which *all* of the reaction channels fire many more times than once is approximately described by the CLE. In a Langevin-leap, the number

of firings of each reaction channel is approximated by a *normal* random number instead of a Poisson random number. Since by hypothesis many reaction events are skipped over, and since also normal random numbers are easier to generate than Poisson random numbers, Langevin-leaping is faster than ordinary tau-leaping.

Finally, if the system admits a description by the CLE and is for all practical purposes at the thermodynamic or "large system" limit, then the random term in the CLE will usually be negligibly small compared to the deterministic term. The CLE then reduces to the deterministic RRE. This RRE limit is usually justified for macroscopic systems, and when it is, it provides the most efficient way to simulate the evolution of the system.

Acknowledgments

This work was supported by the Air Force Office of Scientific Research and the California Institute of Technology under DARPA Award No. F30602-01-2-0558, and also by the Molecular Sciences Institute under Contract No. 244725 with the Sandia National Laboratories and the Department of Energy's "Genomes to Life Program."