The Annals of Applied Probability
2014, Vol. 24, No. 3, 1225–1267
DOI: 10.1214/13-AAP946

STOCHASTICALLY-INDUCED BISTABILITY IN CHEMICAL REACTION SYSTEMS[1]

BY JOHN K. MCSWEENEY AND LEA POPOVIC

Rose-Hulman Institute of Technology and Concordia University

We study a stochastic two-species chemical reaction system with two mechanisms. One mechanism consists of chemical interactions which govern the overall drift of species amounts in the system; the other mechanism consists of resampling, branching or splitting which makes unbiased perturbative changes to species amounts. Our results show that in a system with a large but bounded capacity, certain combinations of these two types of interactions can lead to stochastically-induced bistability. Depending on the relative magnitudes of the rates of these two sets of interactions, bistability can occur in two distinct ways with different dynamical signatures.

1. Introduction. Recent advances in measurement technology have enabled scientists to observe molecular dynamics in single cells and to study the cell-to-cell variability (Brehm-Stecher and Johnson [5]). Many studies have shown that variability observed in genetically identical cells is due to noise that is inherent to biochemical reactions happening within each cell (McAdams and Arkin [22], Elowitz et al. [10]). Understanding how intracellular mechanisms are affected by this intrinsic noise is an important challenge for systems biology. Determining what role this noise plays in creating phenotypic heterogeneity has many practical consequences (Avery [2]).

An important feature in cellular dynamics is bistability, the alternation between two different stable states for a molecular species. This feature is present in many gene-expression systems, where a gene alternates between two types of states ("on" and "off") regulating the production of a protein. It is also present in many phosphorylation switches in signaling pathways. Causes for bistable behavior can be deterministic, but many bistable switching patterns are enabled by stochastic fluctuations. It is often assumed that it follows from the existence of two stable equilibria in the deterministic drift and the ability of infrequent large fluctuations to pull the system from a basin of attraction of one equilibirum to the other. There are also cases of chemical dynamics in which bistability is not possible in the deterministic model, but is possible in the stochastic model of the same chemical

Received May 2012; revised June 2013.

[1]Supported by the New Researchers Grant (F00637-126232) from FQRNT (Fonds de recherche du Québec Nature et technologies).

MSC2010 subject classifications. 60J27, 60J28, 60J60, 60F10, 60F17, 92C45, 92C37, 80A30.

Key words and phrases. Reaction networks, chemical reactions, resampling, bistability, Markov chains, large deviations, stochastic switching, scaling limits.

reaction system (Samoilov et al. [25], Bishop and Qian [4]). Metastable behavior is also sometimes observed (Robert et al. [24]).

In addition to noise inherent to biochemical reactions, cells also experience fluctuations in molecular composition due to cell division. This source of noise is significant, and also difficult to separate from the noise due to biochemical reactions (Huh and Paulsson [14, 15]). In this paper we investigate under what conditions a system of chemical reactions in a cell can use these two sources of noise to exhibit bistable or metastable behavior in their molecular composition.

We would like to emphasize a couple of points observed in the literature. First, the rate of switching between two states is important for cellular development and survival (Acar et al. [1]). Time-scales on which transitions between stable states happen varies whole orders of magnitude over different systems. For example, in the lysogenic state of *E. coli* the time-scale of switching between states is slow (Zong et al. [29])—once per 10^8 cell generations—as determined from the activity of a controlling protein. In the case of gene expression in *S. cerevisiae* the switching time-scale is fast (Kaufmann et al. [17])—once per 8.33 generations—and switching times between mother and daughter cells are correlated in a way that takes several generations to dissipate. Second, both the strength and the distribution of noise affects whether bistability will occur and what the final outcomes will be. Samoilov et al. [25] and Bishop and Qian [4] show that auxiliary chemical reactions can induce a dynamic switching behavior in the enzymatic PdP cycle, and that final dynamics is determined by the noise of the additional reactions. In the bistable switch of lactose operon of *E. coli* Robert et al. [24] show that both cellular growth rate and the molecular concentration levels influence the ability to switch. Huh and Paulsson [15] showed that the type of the cellular division mechanism also plays an important role in the form of the final dynamics. We interpret these observations vis-a-vis our results in the Discussion section.

Finally, we note that bistablility in a stochastic population system is not limited to chemical dynamics. In a genetic population, mutation and selection may lead to alternating fixation in one of two genotypes. In an ecological population, interactions between species can lead to dynamics where two competing species are switching for dominance. We note that our analysis and results apply to any population model described by a density dependent Markov jump process.

1.1. *Outline of results.* We examine qualitatively different ways in which switching between stable states is a result of a stochastic effect in a population modeled by density dependent Markov jump processes. In addition to noise inherent to the reaction system, we include an intrinsically noisy splitting/resampling mechanism in the system. In many stochastic branching models an entity will (upon reproduction, division, duplication, etc.) produce offspring identical to itself. Here we model the division as unbiased but variable. When a cell divides its molecules are randomly allocated to its daughter cells, only on average replicating the parent's molecular composition. We will show that introducing such a splitting

process at a sufficiently high rate can produce switching dynamics in which previously unattainable states become attainable. We will exploit the fact that these two sets of mechanisms (reactions in the system and changes due to unbiased resampling/splitting of the system) may operate on different time-scales.

We consider the following question: which qualitatively different types of behavior can we observe and under which time scaling regimes? The short answer is as follows: (1) If the resampling mechanism is "slower" than the reaction dynamics, then the system behavior will entirely depend on the nonlinear dynamics of the reactions: in case the underlying deterministic system has multiple stable equilibria, the stochastic process will behave as a Markov chain switching between these states. (2) If the resampling mechanism is much "faster" than the reaction dynamics, then the system behavior will not depend on the details of the reaction dynamics, and will behave as a Markov chain switching between two extremes (zero and capacity) of the system. We define a single parameter based on the rates of the two mechanisms that makes the meaning of "faster" and "slower" in the statements above mathematically precise.

We show that a fast but unbiased resampling mechanism may be necessary to produce bistable behavior that the reaction dynamics cannot exhibit. We further show that the two cases, (1) and (2), produce qualitatively different dynamical signatures, in terms of switching times and stable points. Since our analysis only depends on general features (unbiasedness and time-scale of the rate) of the resampling mechanism, one can also use a set of auxiliary reactions instead of resampling. There are other types of noisy mechanisms that one could consider; however, our goal is to stress that adding noise with even small changes (relative to the size of the system) can produce bistable behavior. The additional noise achieves this either by: (1) introducing small perturbations to a dynamical system that already has the required properties for bistability or (2) occurring so frequently that the details of the dynamical system are irrelevant and the system is pushed to its extreme (zero or capacity) amounts.

2. Description of the process.

2.1. *Stochastic model for reaction dynamics.* In the customary notation for interaction of chemical species labeled $A, B, \ldots,$

$$\{a_i A + b_i B + \cdots \longrightarrow a_i' A + b_i' B + \cdots\}_{i=1,\ldots,k} \tag{1}$$

denotes a system of reactions indexed by $i = 1, \ldots, k$ in which $a_i, b_i, \ldots \in \mathbb{Z}^+$ molecules of types $A, B, \ldots$ respectively react and produce $a_i', b_i', \ldots \in \mathbb{Z}^+$ molecules of these types. Each reaction i has a reaction rate λ_i, a time and state-dependent rate of occurrences of this reaction. If $X_A(t), X_B(t), \ldots$ denote the number of molecules of type $A, B, \ldots$ respectively at time $t \geq 0$, then $X(t) = (X_A(t), X_B(t), \ldots)$ evolves as a Markov Jump Process with jump sizes $\{(a_i' - a_i, b_i' - b_i, \ldots)\}$ occurring at rates $\lambda_i(X_A(t), X_B(t), \ldots)$.

The reaction formalism (1) can also be used to describe other systems of interacting entities under a well-mixed spatial assumption. For example, evolution of an SIS epidemic is expressed as $I + S \to 2I$ (infection), $I \to S$ (recovery); a two-allele Moran model with mutation from population genetics can be expressed as $A \to B$, $B \to A$ (mutation), $A + B \to 2A$, $A + B \to 2B$ (resampling).

For simplicity, we consider the effect of a system of chemical reactions on essentially a single molecular species A. We include the effect of only one other species B which satisfies a conservation relation with A. This means that every reaction involving A and B is of the form $aA + bB \to (a + \zeta)A + (b - \zeta)B$ for some $\zeta \in \{-a, \dots, b\}$, and it ensures that the state space of our system is one-dimensional determined by $(X_A(t), t \geq 0)$. The rationale for such a conservation law could come from a cellular environment which is limited (by a factor such as space, or availability of nutrients or catalysts), or a molecular species whose type can take two different forms (e.g., a gene that has two allelic types).

We also assume the following properties for the reaction dynamics:

(1) The amount of species $X_A(t)$ is bounded above by the system capacity N and below by 0. The rate of any reaction that decreases the amount of A is zero when $X_A = 0$, and the rate of any reaction that increases the amount of A is zero when $X_A = N$.

(2) The drift at 0 and N of the overall reaction dynamics is directed toward the interior

$$\frac{d}{ds}\mathbf{E}\big[X_A(s)|X_A(t) = 0\big]|_{s=t} > 0, \qquad \frac{d}{ds}\mathbf{E}\big[X_A(s)|X_A(t) = N\big]|_{s=t} < 0.$$

(3) The form of reaction rates λ is governed by the law of (stochastic) mass-action kinetics. A reaction of the form

$$aA + bB \xrightarrow{\kappa} a'A + b'B$$

has rate $\lambda(X(t)) = \kappa(X_A(t))_a(X_B(t))_b = \kappa(X_A(t))_a(N - X_A(t))_b$. Here $(Z)_c$ denotes the falling factorial $(Z)_c = Z(Z-1)\cdots(Z-c+1)$. When we renormalize $X_A(t)$ by its maximum value N, we will also need the "scaled falling factorial" $(z)_{c,N}$ defined by

$$(z)_{c,N} := N^{-c}(Nz)_c = z\left(z - \frac{1}{N}\right)\left(z - \frac{2}{N}\right)\cdots\left(z - \frac{c-1}{N}\right), \tag{2}$$

$$0 \leq z \leq 1.$$

Note that $\lim_{N\to\infty}(z)_{c,N} = z^c$ for fixed z and c. The constant $\kappa > 0$ is independent of the state $(X_A(t), X_B(t) = N - X_A(t))$ but will depend on the scaling parameter N, $\kappa = \kappa(N)$. We do not necessarily assume that $\kappa(N)$ has the "standard" scaling form $\kappa(N) = \tilde{\kappa}N^{1-(a+b)}$.

(4) The effect on A from any other species in the system is subsumed into the values of the rate constants κ, and are assumed to be state-independent.

Assumption (1) ensures that $X_A(\cdot) \in \{0, \ldots, N\}$ where N serves as the system-size parameter, while assumption (2) ensures the reaction system does not get absorbed at either boundary $\{0, N\}$. Assumption (3) is not essential, but with an explicit scaling of the rate $\kappa(N)$ in terms of N, the polynomial form of the rates λ will make it easy to also establish the scaling of the rates $\lambda(X(t))$ in terms of N under a rescaling of the species amounts X_A [we will occasionally use the notation $\kappa(N)$ for κ when awareness of dependence on N is key]. Assumption (4) is made to absorb the effect of the environment and other species, the changes of which we will not keep track of explicitly.

Under these assumptions, our reaction network system can now be expressed by

$$\{aA + bB \xrightarrow{\kappa_\zeta^{ab}} (a+\zeta)A + (b-\zeta)B\}_{a\in\{0,\ldots,N\},b\in\{0,\ldots,N\},\zeta\in\{-a,\ldots,b\}} \tag{3}$$

with reaction rates of the form $\lambda_\zeta^{ab}(x) = \kappa_\zeta^{ab} \cdot (x)_a (N-x)_b$.

Since the dynamics of the system depends on its overall drift, it will be useful to distinguish a subset of reactions whose combined effect on $\mathbf{E}[X_A(t)]$ is zero, irrespective of the value of $X_A(t)$. In other words, we will group reactions into a subset which contributes zero to the drift ("balanced"), and the rest which are responsible for all of the drift ("biased"). Note that the definition of balance below is made for subsets of reactions—one cannot determine for a single reaction on its own whether it is balanced or not—in order for a reaction to be balanced it needs to belong to a balanced subset.

Let $\mathcal{I}$ denote the set of all triples (a, b, ζ) for which a reaction as written in (3) is present in the system. A subset of reactions is defined as "balanced" $\mathcal{I}^{\mathrm{bal}} \subset \mathcal{I}$ if for some fixed reactant amounts a, b, it satisfies

$$\sum_{\zeta:(a,b,\zeta)\in\mathcal{I}^{\mathrm{bal}}} \zeta \lambda_\zeta^{ab}(x) = 0 \qquad \forall x \quad \Longleftrightarrow \quad \sum_{\zeta\,:\,(a,b,\zeta)\in\mathcal{I}^{\mathrm{bal}}} \zeta \kappa_\zeta^{ab} = 0.$$

A reaction $(a, b, \zeta) \in \mathcal{I}$ that is part of some balanced subset is called "balanced" and all the remaining reactions that are not part of any balanced subset are called "biased," $\mathcal{I}^{\mathrm{bia}} = \mathcal{I} - \mathcal{I}^{\mathrm{bal}}$. Note that our notion of balance is very restrictive and is not related to standard notions of chemical reactions.

For any balanced reaction $(a, b, \zeta) \in \mathcal{I}^{\mathrm{bal}}$, there is necessarily a reaction $(a, b, \zeta') \in \mathcal{I}^{\mathrm{bal}}$ with ζ, ζ' having opposite signs (though not necessarily of the same size). Hence, a reaction $(a, b, \zeta) \in \mathcal{I}^{\mathrm{bal}}$ cannot have nontrivial rate at the boundaries of the system: if $\lambda_\zeta^{a,b}(0) > 0$ for some $\zeta > 0$, then the balance condition would imply the existence of some $\zeta' < 0$ for which $\lambda_{\zeta'}^{a,b}(0) > 0$, which would violate assumption (1) by allowing X_A to drop below 0 upon a single further (a, b, ζ') reaction. Consequently, the boundaries 0 and N are absorbing for the balanced subsystem of reactions, and for all $(a, b, \zeta') \in \mathcal{I}^{\mathrm{bal}}$ we must have both $a > 0$ and $b > 0$. Since assumption (2) does not allow the boundary $\{0, N\}$ to be absorbing

for the full dynamics, this further implies that there is at least one biased reaction $(a,b,\zeta)\in\mathcal{I}^{\mathrm{bia}}$ with $\zeta>0$ and $\zeta\lambda_\zeta^{ab}(0)>0$, hence $a=0,b>0$; and there is at least one biased reaction $(a,b,\zeta')\in\mathcal{I}^{\mathrm{bia}}$ with $\zeta'<0$, and $\zeta'\lambda_{\zeta'}^{a,0}(N)<0$, hence $a>0,b=0$.

The continuous-time Markov jump process model for the reaction dynamics can be expressed in terms of a set of Poisson processes under a random time change. Given a collection $\{Y_\zeta^{ab}\}_{(a,b,\zeta)\in\mathcal{I}}$ of independent unit-rate Poisson processes, the state of the system can be expressed as a solution to the stochastic equation (see [21] or [3] for details)

$$\begin{aligned}X_A(t)&=X_A(0)+\sum_{(a,b,\zeta)\in\mathcal{I}}\zeta Y_\zeta^{ab}\biggl(\int_0^t\lambda_\zeta^{ab}\bigl(X_A(s)\bigr)\,ds\biggr)\\&=X_A(0)+\sum_{(a,b,\zeta)\in\mathcal{I}}\zeta\hat{Y}_\zeta^{ab}\biggl(\int_0^t\lambda_\zeta^{ab}\bigl(X_A(s)\bigr)\,ds\biggr)+\int_0^t F\bigl(X_A(s)\bigr)\,ds,\end{aligned}$$

where $\{\hat{Y}_\zeta^{ab}\}_{(a,b,\zeta)\in\mathcal{I}}$ are centered Poisson processes $\hat{Y}(\lambda t):=Y(\lambda t)-\lambda t$, and

$$F(x)=\sum_{(a,b,\zeta)\in\mathcal{I}}\zeta\lambda_\zeta^{ab}(x)=\sum_{(a,b,\zeta)\in\mathcal{I}^{\mathrm{bia}}}\zeta\kappa_\zeta^{ab}(x)_a(N-x)_b.$$

Since the capacity N of the system may be arbitrarily large, we will consider a "standard" rescaling of the system; see, for example, [11], Chapter 11.2. Let $X_N(t)=N^{-1}X_A(t)$, then

$$\begin{aligned}X_N(t)&=X_N(0)\\&\quad+\sum_{(\zeta,a,b)\in\mathcal{I}}N^{-1}\zeta\hat{Y}_\zeta^{ab}\Bigl(N^{a+b}\kappa_\zeta^{ab}\int_0^t\bigl(X_N(s)\bigr)_{a,N}\bigl(1-X_N(s)\bigr)_{b,N}\,ds\Bigr)\\&\quad+\int_0^t F_N\bigl(X_N(s)\bigr)\,ds,\end{aligned}\tag{4}$$

where the local drift of the renormalized system is given by

$$F_N(x)=\sum_{(\zeta,a,b)\in\mathcal{I}^{\mathrm{bia}}}N^{a+b-1}\zeta\kappa_\zeta^{ab}(x)_{a,N}(1-x)_{b,N}.$$

The most important feature of the Markov jump process model is the relationship of the variance to the drift. Note that we can write (4) as

$$X_N(t)=X_N(0)+M_N(t)+\int_0^t F_N\bigl(X_N(s)\bigr)\,ds,$$

where the second term from (4), a weighted sum of time-changed centered Poisson processes, is a martingale $M_N(t)$ whose quadratic variation satisfies

$$[M_N]_t=\sum_{(\zeta,a,b)\in\mathcal{I}}N^{-2}\zeta^2Y_\zeta^{ab}\Bigl(N^{a+b}\kappa_\zeta^{ab}\int_0^t\bigl(X_N(s)\bigr)_{a,N}\bigl(1-X_N(s)\bigr)_{b,N}\,ds\Bigr).$$

Hence, if $\mathcal{F}_t = \sigma(X(s), 0 \le s \le t)$ denotes the natural filtration of the process, then

$$\begin{aligned}
&\frac{d}{ds}\mathbf{E}\big[X_N(s)|\mathcal{F}_t\big]|_{s=t} \\
&\quad = \mathbf{E}\big[F_N\big(X_N(t)\big)|\mathcal{F}_t\big] \\
&\quad = \sum_{(\zeta,a,b)\in\mathcal{I}^{\mathrm{bia}}} N^{a+b-1}\zeta\kappa_\zeta^{ab}\big(X_N(t)\big)_{a,N}\big(1-X_N(t)\big)_{b,N}, \\
&\frac{d}{ds}\mathbf{E}\big[\big(X_N(s)-\mathbf{E}\big[X_N(s)\big]\big)^2|\mathcal{F}_t\big]|_{s=t} \\
&\quad = \frac{d}{ds}\mathbf{E}\big[[M_N]_s|\mathcal{F}_t\big]|_{s=t} \\
&\quad = \sum_{(\zeta,a,b)\in\mathcal{I}^{\mathrm{bal}}\cup\mathcal{I}^{\mathrm{bia}}} N^{a+b-2}\zeta^2\kappa_\zeta^{ab}\big(X_N(t)\big)_{a,N}\big(1-X_N(t)\big)_{b,N}.
\end{aligned}$$

Recall that the reaction rates $\kappa_\zeta^{ab} = \kappa_\zeta^{ab}(N)$ also depend on the scaling parameter N. The standard scaling for a reaction constant is $\kappa_\zeta^{ab}(N) = \tilde{\kappa}_\zeta^{ab}N^{1-(a+b)}$ for some N-independent constant $\tilde{\kappa}_\zeta^{ab}$. However, regardless of the chosen scaling of κ_ζ^{ab}, for biased reactions $\mathcal{I}^{\mathrm{bia}}$ the order of magnitude for each summand in the infinitesimal variance $\frac{d}{ds}\mathbf{E}[[M_N]_s|\mathcal{F}_t]|_{s=t}$ is N^{-1} times smaller than the corresponding summand in the infinitesimal drift $\mathbf{E}[F_N(X_N(s))|\mathcal{F}_t]|_{s=t}$. This constrains the possible limiting dynamics of X_N. Suppose the scaling of the rates is $\kappa_\zeta^{ab} = N^{1-(a+b)}\tilde{\kappa}_\zeta^{ab}$, and note that then $F_N(x) \to \sum_{(\zeta,a,b)\in\mathcal{I}^{\mathrm{bia}}}\zeta\tilde{\kappa}_\zeta^{ab}x^a(1-x)^b$ uniformly for $x \in [0,1]$. As established in [20], in the limit as $N \to \infty$ the drift overpowers the noise and, provided $X_N(0) \Rightarrow x(0)$, the renormalized process $(X_N(t), t \ge 0)$ converges in distribution (in the Skorokhod topology of cadlag paths) to a solution $(x(t), t \ge 0)$ of the ordinary differential equation

$$x(t) = x(0) + \int_0^t \sum_{(\zeta,a,b)\in\mathcal{I}^{\mathrm{bia}}} \zeta\tilde{\kappa}_\zeta^{ab}x(s)^a\big(1-x(s)\big)^b\,ds. \tag{5}$$

In fact, if the scaling of the reaction constants κ_ζ^{ab} is not standard, but is consistent for both balanced and biased reactions in terms of the polynomial order of the rate function λ_ζ^{ab}, then the same deterministic limit is obtained under an appropriate time rescaling.

The only way to get a stochastic limiting object for X_N is for at least one subset of balanced reactions to have a rate constant with a different scaling in N. This different scaling needs to be such that the noise term due to this subset of reactions will be of the same order of magnitude as the overall drift from the biased reactions. This would require a specific separation of time-scales for balanced versus biased reactions. Although we do not exclude this possibility from our analysis

(see definition of ε_A at the end of this section), our emphasis in this paper is on separating the time-scales in terms of contribution of an additional source of noise, and its ability to produce nontrivial random limiting objects for X_N.

2.2. *Stochastic model for resampling, branching or splitting.* We now introduce the additional mechanism in the system that describes changes to species amounts due to the effect of splitting, branching or resampling, which also effects the species count. For intracellular molecular populations, our first model of splitting was motivated by a simple double-then-divide principle: the cell will first double in size by replicating its constituent molecular species, and then allocate approximately one half of this doubled material into each daughter cell—the allocation mechanism is not perfect and will make random error from the original (undoubled) amount. For genetic populations, common models for resampling follow the Wright–Fisher or the Moran neutral reproduction law: each individual of the offspring population chooses at random from the diploid version of the current population's genes what to inherit—the resampling mechanism is such that an allele of one type in one generation may at random be replaced in the subsequent generation by an allele of the other type. These two are both examples of a general mechanism with the following key properties that we assume for splitting/resampling:

(5) The splitting/resampling occurs at rate $\gamma(x, N)$ that depends on: the current state $X_A = x$ of the system and the scaling parameter N; conditional on $X_A = x$ it is independent of reactions.

(6) The change in the species amount X_A due to a splitting/resampling event has the distribution $p_{x,y} = \mathbf{P}[X_A(t) = y | X_A(t-) = x]$ that have absorbing boundaries $p_{0,0} = 1$, $p_{N,N} = 1$, and that are unbiased

$$\mu_N(x) = \sum_y y p_{x,y} = x \qquad \forall x \in \{0, 1, \ldots, N\}.$$

We also assume, for some of our results, that the rate $\gamma(x, N)$ and distribution $\{p_{x,y}\}$ are such that:

(7*) The change sizes are asymptotically uniformly bounded,

(7*.a) $$\forall \Delta > 0 \qquad \sup_x \gamma(x, N) \sum_{y: N^{-1}|y-x| \geq \Delta} p_{x,y} \to 0 \qquad \text{as } N \to \infty,$$

and the change size variance $\sigma_N^2(x) = \sum_y (y - x)^2 p_{x,y}$ is asymptotically given by

(7*.b) $$\sup_x \left| \gamma(x, N) N^{-2} \sigma_N^2(x) - \tilde{\gamma}^2 \tilde{\sigma}^2 \left(N^{-1} x \right) \right| \to 0 \qquad \text{as } N \to \infty$$

for some constant $\tilde{\gamma} > 0$ and function $\tilde{\sigma}(\cdot)$ that are independent of N, and such that $x \mapsto \tilde{\sigma}^2(x)$ is continuous with $\tilde{\sigma}^2(x) > 0, \forall x \in (0, 1)$ and $\tilde{\sigma}^2(0) = \tilde{\sigma}^2(1) = 0$.

Unbiasedness in assumption (6) could be replaced by an "asymptotic unbiasedness" assumption $N^{-1}|\mu_N(x) - x| \to 0$ as $N \to \infty$, but for the sake of simplicity we assume $\mu_N(x) = x$. Absorption in assumption (6) implies splitting is noiseless on the boundaries regardless of its time-scale. When the additional assumption (7*) holds (as we will assume for our results in Section 2.3), the splitting mechanism contributes diffusively to the limit of the renormalized species count X_N. However, we will also examine the case when the rate of the splitting mechanism is on a slower time-scale (in Section 3.1), as well as the case when it is on a faster time-scale (in Section 4.1). The condition that $\tilde{\sigma}^2$ has boundary values $\tilde{\sigma}^2(0) = \tilde{\sigma}^2(1) = 0$ is natural given that any splitting or resampling mechanism should absorb at the boundaries as indicated by $p_{0,0} = p_{N,N} = 1$.

EXAMPLE (HG). One example of a splitting mechanism would be to completely randomly reallocate the doubled content of a parent cell into daughter cells. If the initial content is $(X_A, X_B) = (x, N - x)$, and the doubled content $(2x, 2(N - x))$ is partitioned in a single swoop (draw without replacement) into two sets of N molecules (one for each daughter cell), then the content in each daughter cell has the hypergeometric distribution (below we keep track of an arbitrarily chosen single lineage)

$$p_{x,y} = \mathbf{P}\big[X_A(t) = y | X_A(t-) = x\big] = \frac{\binom{2x}{y}\binom{2N-2x}{N-y}}{\binom{2N}{N}}, \qquad 0 \vee (2x - N) \le y \le 2x \wedge N.$$

The change in the species count is clearly unbiased $\mu_N(x) = \sum_{y=0\vee N-2x}^{2x\wedge N} y p_{x,y} = x$, with variance

$$\sigma_N^2(x) = \sum_{y=0\vee N-2x}^{2x\wedge N} (y - x)^2 p_{x,y} = \frac{N 2x(2N - 2x)}{4N^2}\left(1 - \frac{N-1}{2N-1}\right) = \frac{x(N-x)}{2N-1}.$$

Then assumption (7*) will hold if $\gamma(x, N) = \tilde{\gamma}^2 N$ and $\tilde{\sigma}^2(x) = \frac{1}{2}x(1 - x)$, since for (7*.a) we have

$$\sup_x |\gamma(x, N) N^{-2} \sigma_N^2(x) - \tilde{\gamma}^2 \tilde{\sigma}^2(N^{-1}x)| = \tilde{\gamma}^2 \sup_x \left| \frac{x(N-x)}{N(2N-1)} - \frac{1}{2}\frac{x}{N}\left(1 - \frac{x}{N}\right)\right| \to 0$$

and using tail bounds for the hypergeometric distribution [6] $\sum_{y=x+N\Delta}^{N} p_{x,y} \le e^{-2\Delta^2 N}$ independently of x, and for (7*.b) we have

$$\gamma(x, N) \sup_x \sum_{y:|y-x|\ge N\Delta} p_{x,y} \le 2\tilde{\gamma}^2 N e^{-2\Delta^2 N} \to 0.$$

EXAMPLE (Bin). Another example would be to sample with replacement from the population in which each offspring picks its type randomly from any individual in the parent generation. If the initial count is $X_A = x$, then the count in the next generation has the binomial distribution

$$p_{x,y} = \mathbf{P}[X_A(t) = y | X_A(t-) = x] = \binom{N}{y} \left(\frac{x}{N}\right)^y \left(1 - \frac{x}{N}\right)^{N-y}, \qquad 0 \le y \le N.$$

This form of resampling is used in (the haploid version of) the Wright–Fisher model for genetic drift (e.g., [9] Section 1.2). It is also used as the prototype of a splitting mechanism of simple "independent segregation" of division of cells [15]. This distribution is again unbiased, and assumption (7*) will hold if $\gamma(x, N) = \frac{1}{2}\tilde{\gamma}^2 N$ for some constant $\tilde{\gamma}^2 > 0$. Using similar arguments as above, it is then easy to show that both (7*.a) and (7*.b) will hold with $\tilde{\sigma}^2(x) = \frac{1}{2}x(1-x)$.

EXAMPLE (Bern). Finally, the simplest example of a splitting/resampling mechanism is to have a single amount error in the daughter cell (or the next generation), and to have the rate at which the error occurs be proportional to both the current amount $X_A = x$ and the amount of $X_B = N - x$. Errors from imperfect division will result in $\pm$ change with equal probability

$$p_{x,x-1} = p_{x,x+1} = 1/2.$$

This distribution is clearly unbiased, and assumption (7*) will hold if the rate of error occurrences is $\gamma(x, N) = \frac{1}{2}\tilde{\gamma}^2 N^2 \frac{x}{N}(1 - \frac{x}{N})$ for some $\tilde{\gamma}^2 > 0$, with the limiting variance $\tilde{\sigma}^2(x) = \frac{1}{2}x(1-x)$.
This form of resampling is used in the Moran model for genetic drift (e.g., [9] Section 1.5). It is also used in [15] as an example of an "ordered segregation" splitting mechanism for cell division (self volume exclusion partitioning error, [15] Supporting Information). In a cellular system it could also be described as a set of balanced reactions $A + B \to 2A$, $A + B \to 2B$ with mass-action dynamics and appropriately scaled rate constants.

We note that, from the perspective of limiting results, the differences in the specific details of the mechanism will not be important. The only feature of relevance will be the order of magnitude of the prelimiting rate $\gamma(x, N)$ and the form of the limiting variance $\tilde{\sigma}^2(x)$. There are many other types of splitting, branching or resampling mechanisms, yielding a different form for the limiting variance. They are easy to construct in case of small changes that result in single count errors, via a range of birth–death probability distributions. We shall see, in both Section 3 and Section 4, how the actual form for the variance $\tilde{\sigma}^2(x)$ affects the qualitative behavior of the limit of the renormalized process.

The changes due to this additional mechanism can also be expressed in terms of a Poisson processes under a random time change. Let Y_γ be a counting process

with state-dependent rate $\gamma(x, N)$, and $\{Z(x, s)\}_{0\le x\le N}$ be independent random variables with probability distribution $p_{x,\cdot}$ for any $s \ge 0$. A change due to splitting or resampling can be represented as a stochastic integral $\int_0^t (Z(X(s-), s) - X(s-))\, dY_\gamma(s)$. The evolution in species count due to both reaction dynamics and splitting is

$$X_A(t) = X_A(0) + \sum_{(\zeta,a,b)\in I} \zeta \hat{Y}_\zeta^{ab}\Big(\int_0^t \lambda_\zeta^{ab}\big(X_A(s)\big)\, ds\Big) + \int_0^t F\big(X_A(s)\big)\, ds + \int_0^t \big(Z\big(X_A(s-), s\big) - X_A(s-)\big)\, dY_\gamma(s);$$

hence for the rescaled system $X_N = N^{-1}X_A$ we have

$$\begin{aligned} X_N(t) &= X_N(0) \\ &\quad + N^{-1}\sum_{(\zeta,a,b)\in\mathcal{I}} \zeta \hat{Y}_\zeta^{ab}\Big(N^{a+b}\kappa_\zeta^{ab}\int_0^t \big(X_N(s)\big)_{a,N}\big(1 - X_N(s)\big)_{b,N}\, ds\Big) \\ &\quad + \int_0^t F_N\big(X(s)\big)\, ds \\ &\quad + \int_0^t \big(N^{-1}Z\big(NX_N(s-), s\big) - X_N(s-)\big)\, d\hat{Y}_\gamma(s) \\ &\quad + \int_0^t \big(N^{-1}Z\big(NX_N(s-), s\big) - X_N(s-)\big)\gamma\big(NX_N(s), N\big)\, ds \end{aligned} \tag{6}$$

$$\begin{aligned} &= X_N(0) + M_{N,\gamma}(t) + \int_0^t F_N\big(X(s)\big)\, ds \\ &\quad + \int_0^t N^{-1}\big(Z\big(NX_N(s-), s\big) - NX_N(s-)\big)\gamma\big(NX_N(s), N\big)\, ds. \end{aligned} \tag{6$'$}$$

We still have

$$F_N(x) = \sum_{(\zeta,a,b)\in\mathcal{I}^{\mathrm{bia}}} N^{a+b-1}\zeta\kappa_\zeta^{ab}(N)(x)_{a,N}(1-x)_{b,N}, \tag{7}$$

but now $M_{N,\gamma}$ denotes the martingale formed by the second and fourth summand in (6) whose quadratic variation is

$$\begin{aligned} &[M_{N,\gamma}]_t \\ &= \sum_{(\zeta,a,b)\in\mathcal{I}} N^{-2}\zeta^2 Y_\zeta^{ab}\Big(N^{a+b}\kappa_\zeta^{ab}(N)\int_0^t \big(X_N(s)\big)_{a,N}\big(1 - X_N(s)\big)_{b,N}\, ds\Big) \\ &\quad + \int_0^t N^{-2}\big(Z\big(NX_N(s-), s\big) - NX_N(s-)\big)^2\, dY_\gamma(s). \end{aligned} \tag{8}$$

Note that since the two mechanisms are driven by independent Poisson processes, there is no quadratic covariation contribution. Since $\mathbf{E}_{p_{x,\cdot}}[Z(x,s)-x]=0$ for all $s\geq 0$ and $x\in\{0,\dots,N\}$, the infinitesimal mean still satisfies

$$\begin{aligned}&\frac{d}{ds}\mathbf{E}\big[X_N(s)|\mathcal{F}_t\big]\Big|_{s=t}\\ &\quad=\mathbf{E}\big[F_N\big(X_N(t)\big)|\mathcal{F}_t\big]\\ &\quad=\sum_{(\zeta,a,b)\in\mathcal{I}^{\mathrm{bia}}}N^{a+b-1}\zeta\kappa_\zeta^{ab}(N)\big(X_N(t)\big)_{a,N}\big(1-X_N(t)\big)_{b,N};\end{aligned}\tag{9}$$

on the other hand, the infinitesimal variance now satisfies

$$\begin{aligned}&\frac{d}{ds}\mathbf{E}\big[\big(X_N(s)-\mathbf{E}\big[X_N(s)\big]\big)^2|\mathcal{F}_t\big]\Big|_{s=t}\\ &\quad=\frac{d}{ds}\mathbf{E}\big[[M_N]_s|\mathcal{F}_t\big]\Big|_{s=t}\\ &\quad=\sum_{(\zeta,a,b)\in\mathcal{I}^{\mathrm{bal}}\cup\mathcal{I}^{\mathrm{bia}}}N^{a+b-2}\zeta^2\kappa_\zeta^{ab}(N)\big(X_N(s)\big)_{a,N}\big(1-X_N(s)\big)_{b,N}\\ &\qquad\qquad+\gamma\big(NX_N(s),N\big)N^{-2}\sigma_N^2\big(NX_N(s)\big).\end{aligned}\tag{10}$$

2.3. *Possible qualitative behaviors.* In order to determine the role that the rate of the splitting/resampling mechanism may play, we first establish the possible behavior of the system when N is large. The decisive quantity for the qualitative behavior of the system is

$$\varepsilon_A:=\lim_{N\to\infty}\varepsilon_A(N),\qquad \varepsilon_A(N):=\frac{c_{\sigma^2}(N)}{c_\mu(N)},\tag{11}$$

where

$$c_{\sigma^2}(N):=\sum_{(a,b,\zeta)\in\mathcal{I}^{\mathrm{bal}}}N^{a+b-2}\kappa_\zeta^{ab}(N)+\sup_{x\in[0,1]}\gamma(Nx,N)N^{-2}\sigma_N^2(Nx)\tag{12}$$

and

$$c_\mu(N):=\sum_{(a,b,\zeta)\in\mathcal{I}^{\mathrm{bia}}}N^{a+b-1}\kappa_\zeta^{ab}(N);\tag{13}$$

ε_A relates the magnitude of the variance due to the splitting mechanism (or possibly a faster set of balanced reactions) to the magnitude of the drift due to reaction dynamics. If they are of the same order of magnitude, then the rescaled process will converge to a diffusion. In other words, if $\varepsilon_A\in(0,\infty)$, then we can assume (by rescaling time as necessary) that both scaling constants (12) and (13) satisfy $\tilde{c}_{\sigma^2}=\lim_{N\to\infty}c_{\sigma^2}(N)\in(0,\infty)$, $\tilde{c}_\mu=\lim_{N\to\infty}c_\mu(N)\in(0,\infty)$ and $\tilde{c}_{\sigma^2}=\varepsilon_A\tilde{c}_\mu$. If assumption (7*) is satisfied, the noise of the splitting mechanism is such that the limiting behavior of the system is diffusive, instead of being deterministic, as in (5) when only reactions are present.

PROPOSITION 2.1. *If* $\varepsilon_A \in (0,\infty)$, *assumption* (7*) *holds for* $X_A = NX_N$, *and* $X_N(0) \Rightarrow \tilde{X}(0) \in [0,1]$, *then* $X_N \Rightarrow \tilde{X}$ *as* $N \to \infty$ *in distribution on the Skorokhod space of cadlag paths on* $[0,1]$, *where* $\tilde{X}$ *is a diffusion with drift and diffusion coefficients given by*

$$\begin{aligned}\tilde{\phi}(x) &= \sum_{(a,b,\zeta)\in\mathcal{I}^{\mathrm{bia}}} \zeta \tilde{\kappa}^{ab}_{\zeta,\mu} x^a (1-x)^b,\\ \tilde{a}(x) &= \sum_{(a,b,\zeta)\in\mathcal{I}^{\mathrm{bal}}} \zeta^2 \tilde{\kappa}^{ab}_{\zeta,\sigma^2} x^a (1-x)^b + \tilde{\gamma}^2 \tilde{\sigma}^2(x),\end{aligned} \tag{14}$$

where for each $(a,b,\zeta) \in \mathcal{I}^{\mathrm{bia}}$

$$\tilde{\kappa}^{ab}_{\zeta,\mu} = \lim_{N\to\infty} N^{a+b-1} \kappa^{ab}_{\zeta}(N)$$

for each $(a,b,\zeta) \in \mathcal{I}^{\mathrm{bal}}$

$$\tilde{\kappa}^{ab}_{\zeta,\sigma^2} = \lim_{N\to\infty} N^{a+b-2} \kappa^{ab}_{\zeta}(N)$$

and for some $\tilde{\gamma}^2 > 0$

$$\tilde{\gamma}^2 \tilde{\sigma}^2(x) = \lim_{N\to\infty} \gamma(Nx,N) N^{-2} \sigma_N^2(Nx).$$

If all reaction rates have standard scaling $\kappa^{ab}_{\zeta} = \tilde{\kappa}^{ab}_{\zeta} N^{1-(a+b)}$, *then* $\tilde{\kappa}^{ab}_{\zeta,\sigma^2} = 0$ *and* $\tilde{a}(x) = \tilde{\gamma}^2 \tilde{\sigma}^2(x)$.

PROOF. This is a direct consequence of standard theorems for convergence of Markov processes to a diffusion (see, e.g., [8] Section 8.7) based on locally uniform convergence of the infinitesimal mean and variance to the limiting drift and diffusion coefficients, respectively, and convergence of jumps so that they disappear in the limit. Recall that the infinitesimal mean of the rescaled process X_N from (6′) is given by (9) and its infinitesimal variance by (10). Since the process takes values in $[0,1]$, we can check convergence uniformly on the whole space, and moreover $M_{N,\gamma}(t) = X_N(t) - \mathbf{E}[X_N(t)]$, whose quadratic variation is given in (8), is then a square integrable martingale. For the contributions by the splitting mechanism, the convergence of the infinitesimal mean and variance, as well as the control of the jumps, are easy to check from the three requirements on the splitting mechanism made in assumptions (6) and (7*). For the contributions by the reaction dynamics the convergence of the infinitesimal mean and variance, and the control over jumps, follow from the scaling properties of the counting processes used in their representation and from the fact that the rates for these counting processes are Lipschitz and bounded. These same conditions have been checked, in the case when reaction rates have a more general form, for law of large numbers and central limit theorem results for rescaled population-dependent Markov processes [20]. Alternatively, one could also check that the Markov process X_N

satisfies all the conditions required for convergence of more general Markov jump processes to a diffusion as stated in Theorem 2.11 of [18] and Theorem 3.1 of [19]. The only thing left to check is whether a diffusion with coefficients as given exists and is unique in law. This follows easily from the fact that the contributions to $\tilde{a}(x)$ and $\tilde{\phi}(x)$ from reaction rates are polynomial, and we have assumed that $\tilde{\sigma}^2(x)$ is Lipschitz. □

A diffusion may or may not hit its boundary points, but it never spends a disproportionate amount of time at any point in its range, including the boundaries, unless they are absorbing. Hence, we really need to consider the behavior of the process when either $\varepsilon_A \to 0$ or $\varepsilon_A \to \infty$ (as a function of an additional asymptotic parameter which will be discussed below in Section 3.1). The only remark we make when ε_A remains bounded away from 0 and ∞ is that the behavior of $\tilde{X}$ at the boundary $\{0, 1\}$ depends on the form for the limiting variance of the splitting mechanism. As a consequence of assumption (2), and of the properties of the splitting variance at $\{0, 1\}$, we are only guaranteed that $\tilde{\phi}(0) > 0, \tilde{\phi}(1) < 0$ and $\tilde{a}(0) = \tilde{a}(1) = 0$. Hence, $\{0, 1\}$ are neither absorbing nor natural, but it remains to determine whether they are entrance or regular boundary points. Further conditions on the reaction and splitting mechanisms for reaching the boundary (i.e., for $\{0, 1\}$ to be regular boundary points) are guaranteed by interpreting Feller's test for explosion; see, for example, [8], Section 6.2. or [16], Section 15.6.

The diffusive case $\varepsilon_A \in (0, \infty)$ separates two other types of behavior. When $\varepsilon_A \approx 0$ and $\varepsilon_A \approx \infty$, the rate of splitting is either slower or faster, respectively, than prescribed by assumption (7*). Both cases lead to behavior which exhibits a type of stochastic bistability, in which the system spends almost all of its time at two points, or very near them. This bistability is, in the two cases $\varepsilon_A \approx 0$ and $\varepsilon_A \approx \infty$, caused by completely different effects of the two stochastic mechanisms in our model, which we investigate separately in the next two sections.

3. Bistable behavior from slow splitting. Let us consider the case $\varepsilon_A \approx 0$, and assume that time has been rescaled so that $\tilde{c}_\mu = \lim_{N\to\infty} c_\mu(N) \in (0, \infty)$ and $\tilde{c}_{\sigma^2} = \lim_{N\to\infty} c_{\sigma^2}(N) \approx 0$. In modeling this is a relatively conventional scaling, in which a small amount of noise (from balanced reactions and splitting) will affect the predominantly deterministic behavior due to drift (of biased reactions). A precise statement of this depends on how fast $\varepsilon_A(N) = \frac{c_{\sigma^2}(N)}{c_\mu(N)}$ approaches 0 as a function of N, and we examine it more carefully by first introducing a separate perturbation parameter ε and then relating it to the scaling parameter N.

3.1. *Small diffusive noise effects.* The simplest way to model small diffusive effects is with an enforced separation of time-scales between reactions and splitting using a perturbation parameter. Suppose all the reaction constants $\kappa_\zeta^{ab}(N)$ depend only on the scale of the system N and have the standard scaling

$\kappa_\zeta^{ab} = \tilde{\kappa}_\zeta^{ab} N^{1-(a+b)}$ for some constants $\tilde{\kappa}_\zeta^{ab}$. Suppose the splitting rate, in addition to N, also depends on a small parameter $\varepsilon > 0$, so that the splitting rate is $\gamma(x, N, \varepsilon) = \varepsilon^2 \gamma(x, N)$ where $\gamma(x, N)$ satisfies assumption (7*). The fact that the splitting rate is slower than diffusive is expressed in terms of the fact that we will consider the behavior of the system as $\varepsilon \to 0$. In this case the quantitiy ε_A defined in (11) is just a constant multiple of ε^2

$$\begin{aligned}\varepsilon_A &:= \lim_{N\to\infty} \frac{\sum_{(a,b,\zeta)\in I^{\mathrm{bal}}} N^{a+b-2}\kappa_\zeta^{ab}(N) + \varepsilon^2 \sup_{x\in[0,1]} N^{-2}\gamma(Nx, N)\sigma_N^2(Nx)}{\sum_{(a,b,\zeta)\in I^{\mathrm{bia}}} N^{a+b-1}\kappa_\zeta^{ab}(N)} \\ &= \frac{\varepsilon^2 \tilde{c}_{\sigma^2}}{\tilde{c}_\mu},\end{aligned}$$

where $\tilde{c}_{\sigma^2} = \tilde{\gamma}^2 \sup_{x\in[0,1]} \tilde{\sigma}^2(x)$ and $\tilde{c}_\mu = \sum_{(a,b,\zeta)\in I^{\mathrm{bia}}} \tilde{\kappa}_{\zeta,\mu}^{ab}$.

We could also assume the rates of balanced reactions depend on the additional parameter ε^2, in the sense that $\kappa_\zeta^{ab} = \varepsilon^2 \tilde{\kappa}_\zeta^{ab} N^{2-(a+b)}$ for $(a, b, \zeta) \in \mathcal{I}^{\mathrm{bal}}$. In this case

$$\tilde{c}_{\sigma^2} = \sum_{(a,b,\zeta)\in I^{\mathrm{bal}}} \tilde{\kappa}_{\zeta,\mu}^{ab} + \tilde{\gamma}^2 \sup_{x\in[0,1]} \tilde{\sigma}^2(x).$$

However, if we make no special separation in the way balanced and biased reactions are scaled, then the assumption of standard scaling $\kappa_\zeta^{ab} = \tilde{\kappa}_\zeta^{ab} N^{1-(a+b)}$ implies that this is only possible if the parameter ε satisfies $\varepsilon^2 = N^{-1}$, on which we remark further in the next subsection.

By Proposition 2.1, for any fixed $\varepsilon > 0$, the process obtained in the limit $X_N \Rightarrow \tilde{X}_\varepsilon$ is a diffusion with coefficients $\tilde{\phi}(x)$ as in (14) and $\tilde{a}_\varepsilon(x) = \varepsilon^2 \tilde{\gamma}^2 \tilde{\sigma}^2(x)$ (we will use the subscript ε in the notation of the limiting diffusion to stress its dependence on the small parameter ε). $\tilde{X}_\varepsilon$ is a solution of the stochastic differential equation

$$d\tilde{X}_\varepsilon(t) = \tilde{\phi}\big(\tilde{X}_\varepsilon(t)\big)\,dt + \varepsilon\tilde{\gamma}\tilde{\sigma}\big(\tilde{X}_\varepsilon(t)\big)\,dB(t), \qquad \tilde{X}_\varepsilon \in [0, 1], \tag{15}$$

where B is a standard Brownian motion, a classical case of a diffusion with small diffusion coefficient.

For many such diffusions $\varepsilon \approx 0$ will have little qualitative effect relative to $\varepsilon = 0$; however, suppose that $\tilde{\phi}$ has two stable and one unstable equilibria, and thus the potential Φ defined by $\Phi = -\int \tilde{\phi}$ is a double-well potential. Since $\tilde{\phi}$ is a polynomial, this is an assumption on the number and type of zeros of $\tilde{\phi}$. Explicitly, we will assume that

$$\begin{aligned}&\exists 0 < x_1 < x_2 < x_3 < 1 : \tilde{\phi}(x_i) = 0, \\ &\qquad i = 1, 2, 3 \text{ and } \tilde{\phi}'(x_1) < 0, \tilde{\phi}'(x_2) > 0, \tilde{\phi}'(x_3) < 0.\end{aligned} \tag{16}$$

Recall also that assumption (2) implies that at the boundaries we have $\tilde{\phi}(0) > 0, \tilde{\phi}(1) < 0$. As a consequence, $\tilde{X}_\varepsilon$ is a process whose mean behavior involves

monotone convergence to one of two stable equilibria (determined by the initial conditions), but where the small amount of noise allows the process to switch from one equilibrium to the other, creating a bistable system. Precise statements of this behavior are described by Freidlin–Wentzell theory for random perturbations of dynamical systems by diffusive noise, [12], which can also cover processes with metastability, [13]. We will follow closely the notation of [13], as these results apply most directly to $\tilde{X}_\varepsilon$. We first need a transformation to handle the state dependence $\tilde{\sigma}^2(x)$ of the diffusion coefficient, easily done using [7], Section 5.6. or [23] Section 2.5.

For $\tilde{X}_\varepsilon$ satisfying (15), large deviation theory for Gaussian perturbations of dynamical systems, Dembo and Zeitouni ([7] Theorem 5.6.7 and Exercise 5.6.25), state that deviations of $\tilde{X}_\varepsilon$ away from an ε-sized neighborhood of x_1 and x_3 are characterized by the large deviation rate function for $\tilde{X}_\varepsilon$ given by the quasipotential (with respect to x_i and x_2)

$$I_{x_i,x_2}(\tilde{\phi},\tilde{\gamma}\tilde{\sigma}) := \inf_{s>0}\inf_{\xi}\left\{\int_0^s L\big(\xi(u),\xi'(u)\big)\,du \,\middle|\, \xi\in C^1\big([0,s]\big),\xi(0)=x_i,\xi(s)=x_2\right\}, \quad i=1,3,$$

where L is the action functional

$$L(\xi,\xi') = \left(\frac{\xi'-\tilde{\phi}(\xi)}{\tilde{\gamma}\tilde{\sigma}(\xi)}\right)^2.$$

This identifies the most likely paths which leave a neighborhood of x_1 or x_3, since every path between x_1 and x_3 of the one-dimensional $\tilde{X}_\varepsilon$ has to pass through x_2. We can write $L(\xi,\xi')$ in this form for all such paths because $\tilde{X}_\varepsilon$ is nonsingular away from the boundaries, that is, $\tilde{\sigma}^2(x)>c,\forall x\in[x_1,x_3]$ for some $c>0$. If the diffusion coefficient were constant $\tilde{\gamma}\tilde{\sigma}\equiv 1$, then $L(\xi,\xi')=(\xi'-\tilde{\phi}(\xi))^2$ and $I_{x_i,x_2}(\tilde{\phi},1)$ would be simply a constant multiple of the potential, $I_{x_i,x_2}(\tilde{\phi},1)=2(\Phi(x_2)-\Phi(x_i))$, for $i=1,3$. The quasipotential would be determined by the height of the potential barrier which $\tilde{X}_\varepsilon$ needs to overcome in order to pass from one equilibrium to the basin of attraction of the other.

To solve the variational problem in our case, we can use a transformation of the path space $\xi=g(\psi)$ to get an action functional of the form $L(\xi,\xi')=(\psi'-\tilde{\phi}(\psi))^2$, from which we can deduce the explicit form of the rate function I_{x_i,x_2} for state-dependent $\tilde{\gamma}\tilde{\sigma}(x)$. For any monotone C^1 function g which for all s is surjective from $C^1([0,s])$ to $C^1([0,s])$, we have

$$I_{x_i,x_2}(\tilde{\phi},\tilde{\gamma}\tilde{\sigma}) = \inf_{s>0}\inf_{\psi}\left\{\int_0^s L\big(g(\psi(u)),[g(\psi(u))]'\big)\,du \,\middle|\, \psi\in C^1\big([0,s]\big), \psi(0)=g^{-1}(x_i),\psi(s)=g^{-1}(x_2)\right\}.$$

We take g which satisfies the (autonomous) first-order ODE $g'(y) = \tilde{\gamma}\tilde{\sigma}(g(y))$, so that

$$L\big(g(\psi), [g(\psi)]'\big) = \left(\frac{g'(\psi)\psi' - \tilde{\phi}(g(\psi))}{\tilde{\gamma}\tilde{\sigma}(g(\psi))}\right)^2 = \left(\psi' - \frac{\tilde{\phi}(g(\psi))}{\tilde{\gamma}\tilde{\sigma}(g(\psi))}\right)^2.$$

Note that $\tilde{\gamma}\tilde{\sigma}(x) > 0, \forall x \in (0,1)$ ensures that g is in fact strictly increasing on $(0,1)$. Let $h(x) = g^{-1}(x)$. Then, if $\tilde{\phi}$ is the vector field of a double-well potential, so is α defined as $\alpha = \frac{\tilde{\phi}\circ g}{\tilde{\gamma}\tilde{\sigma}\circ g}$, for the following reasons. Let $y_i = g^{-1}(x_i) = h(x_i)$; these will be the equilibria for α, since $\alpha(y_i) = \tilde{\phi}(g(y_i))/\tilde{\gamma}\tilde{\sigma}(g(y_i)) = \tilde{\phi}(x_i)/\tilde{\gamma}\tilde{\sigma}(x_i) = 0$. As for their stability, we have

$$\begin{aligned}\alpha'(y_i) &= \frac{\tilde{\phi}'(g(y_i))g'(y_i)\tilde{\gamma}\tilde{\sigma}(y_i) - \tilde{\phi}(g(y_i))\tilde{\gamma}\tilde{\sigma}'(g(y_i))g'(y_i)}{\tilde{\gamma}^2\tilde{\sigma}^2(g(y_i))} \\ &= \frac{\tilde{\phi}'(g(y_i))g'(y_i)}{\tilde{\gamma}\tilde{\sigma}(g(y_i))} = \tilde{\phi}'\big(g(y_i)\big),\end{aligned}$$

where the first equality holds since $\tilde{\phi}(g(y_i)) = 0$, and the second by definition of g. Therefore, for each i, the stability of x_i under the vector field $\tilde{\phi}$ is the same as that of y_i with vector field α; we may therefore define $A = -\int \alpha$ to be the (double-well) potential associated with α. Since $L(\xi, \xi')$ is now in the form $L(g(\psi), [g(\psi)]') = (\psi' - \alpha(g(\psi)))^2$, we can conclude that

$$I_{x_i,x_2}(\tilde{\phi}, \tilde{\gamma}\tilde{\sigma}) = I_{y_i,y_2}(\alpha, 1) = 2\big(A(y_2) - A(y_i)\big), \qquad i = 1, 3. \tag{17}$$

We can now interpret the results of [13] to characterize the behavior of the process $\tilde{X}_\varepsilon$ [defined in (15)] as $\varepsilon \to 0$. Let D_i denote basins of attraction for the deterministic process (5) driven by the drift $\tilde{\phi}$, that is, $D_1 = [0, x_2) \ni x_1$, $D_2 = \{x_2\}$, $D_3 = (x_2, 1] \ni x_3$, and $B_c(x_i)$ denote closed balls of radius $c > 0$ around x_1, x_3 such that $B_c(x_1) \subset D_1$, $B_c(x_3) \subset D_3$. If the wells of the transformed potential A are not at equal depth $A(y_1) \neq A(y_3)$, we will without loss of generality assume $A(y_1) < A(y_3)$. Let

$$T_\varepsilon = \inf\big\{t > 0 : \tilde{X}_\varepsilon(t) \in B_c(x_1)\big\}, \qquad \tilde{T}_\varepsilon = \inf\big\{t > T_\varepsilon : \tilde{X}_\varepsilon(t) \in B_c(x_3)\big\}$$

denote the first hitting time of the neighborhood of the stable equilibrium with the deeper basin, and the subsequent first hitting time of the neighborhood of the other stable equilibrium. Let β_ε be the time-scale on which transitions from D_3 to the neighborhood of x_1 happen, defined by $\mathbf{P}[T_\varepsilon > \beta_\varepsilon | \tilde{X}_\varepsilon(0) = x_3] = e^{-1}$, and $\tilde{\beta}_\varepsilon$ the one on which the reverse transition happen, defined by $\mathbf{P}[\tilde{T}_\varepsilon > \tilde{\beta}_\varepsilon | \tilde{X}_\varepsilon(0) = x_1] = e^{-1}$. The next result establishes that the transition from one stable equilibrium to the other happens on a time-scale of order $O(e^{\varepsilon^{-2}(A(y_2)-A(y_i))})$ with $i = 3$ and $i = 1$, respectively, and that in the limit as $\varepsilon \to 0$ the transition times have an exponential distribution.

PROPOSITION 3.1. *If $\tilde{\phi}$ satisfies (16), then the transitions of $\tilde{X}_\varepsilon$ from D_3 to $B_c(x_1)$ and from D_1 to $B_c(x_3)$ satisfy*:

$$\text{(i)} \quad \lim_{\varepsilon\to 0} \mathbf{P}\big[T_\varepsilon > t\beta_\varepsilon | \tilde{X}_\varepsilon(0) = x \in D_3\big] = e^{-t} \qquad \forall t > 0,$$

$$\lim_{\varepsilon\to 0} \mathbf{P}\big[\tilde{T}_\varepsilon > t\tilde{\beta}_\varepsilon | \tilde{X}_\varepsilon(0) = x \in D_1\big] = e^{-t} \qquad \forall t > 0;$$

$$\text{(ii)} \quad \lim_{\varepsilon\to 0} \varepsilon^2 \ln \beta_\varepsilon = I_{x_3,x_2}(\tilde{\phi}, \tilde{\gamma}\tilde{\sigma}) = 2\big(A(y_2) - A(y_3)\big),$$

$$\lim_{\varepsilon\to 0} \varepsilon^2 \ln \tilde{\beta}_\varepsilon = I_{x_1,x_2}(\tilde{\phi}, \tilde{\gamma}\tilde{\sigma}) = 2\big(A(y_2) - A(y_1)\big).$$

PROOF. (i) is a restatement of Theorem 1 in [13]. (ii) follows from Theorem 4.2 of Chapter 4 in [12], which states that for any $\delta > 0$, $\lim_{\varepsilon\to 0} \mathbf{P}[|\varepsilon^2 \ln T_\varepsilon - I_{x_3,x_2}(\tilde{\phi}, \tilde{\gamma}\tilde{\sigma})| > \delta | \tilde{X}_\varepsilon(0) = x_3] = 0$ and $\lim_{\varepsilon\to 0} \mathbf{P}[|\varepsilon^2 \ln \tilde{T}_\varepsilon - I_{x_1,x_2}(\tilde{\phi}, \tilde{\gamma}\tilde{\sigma})| > \delta | \tilde{X}_\varepsilon(0) = x_1] = 0$, and from our explicit calculation of the value of the quasipotential in (17). □

The following result characterizes the long-term behavior on the natural time-scale (determined by β_ε) for transition to the stable point with the deeper basin. Let $R_\varepsilon = e^{\varepsilon^{-2}a}$ for some $a \in (0, 2(A(y_2) - A(y_3)))$, so that $R_\varepsilon \to \infty$ while $R_\varepsilon/\beta_\varepsilon \to 0$ as $\varepsilon \to 0$. Again following [13], define the measure-valued process $(\nu_t^\varepsilon)_{t\geq 0}$ by

$$\nu_t^\varepsilon(f) = \frac{1}{R_\varepsilon} \int_{\beta_\varepsilon t}^{\beta_\varepsilon t + R_\varepsilon} f\big(\tilde{X}_\varepsilon(s)\big)\, ds$$

for any (bounded) continuous function f on $[0, 1]$. The measure ν_t^ε approximates the law for the location of $\tilde{X}_\varepsilon(T)$ on the time-scale $T = \beta_\varepsilon t$.

Note that if $A(y_1) < A(y_3)$, then the results of (ii) imply that $\tilde{\beta}_\varepsilon/\beta_\varepsilon \to \infty$ as $\varepsilon \to 0$, so that $\inf_{x\in D_1} \mathbf{P}[\tilde{T}_\varepsilon/\beta_\varepsilon > t | \tilde{X}_\varepsilon(0) = x] \to 1$. Hence, in this case metastability is characterized by the fact that the transitions into the deeper well are on an exponentially faster time-scale, relative to which the transitions back into the less deep well will not be noticed. Let $\mathbf{P}_x[\cdot]$ denote $\mathbf{P}[\cdot | \tilde{X}_\varepsilon(0) = x]$.

PROPOSITION 3.2. *For each $x \in D_3$, continuous function f on $[0, 1]$, and $\delta > 0$ we have*

$$\lim_{\varepsilon\to 0} \mathbf{P}_x\Big[\sup_{s\in[0,(T_\varepsilon - 3R_\varepsilon)/R_\varepsilon]} \big|\nu_t^\varepsilon(f) - f(x_3)\big| > \delta\Big] = 0,$$

$$\lim_{\varepsilon\to 0} \mathbf{P}\Big[\sup_{s\in[T_\varepsilon/\beta_\varepsilon,(\tilde{T}_\varepsilon - 3R_\varepsilon)/R_\varepsilon]} \big|\nu_t^\varepsilon(f) - f(x_1)\big| > \delta\Big] = 0.$$

Moreover, we have convergence in law on the space of cadlag paths (with the Skorokhod topology) of $(\nu_t^\varepsilon)_{t\geq 0}$ to a jump process $(\nu_t)_{t\geq 0}$ such that:

(i) (*Metastability*). *If* $A(x_1) < A(x_3)$, *then* $(\nu_t)_{t\geq 0}$ *is given by*

$$\nu_t = \begin{cases} \delta_{x_3}, & t < T, \\ \delta_{x_1}, & t \geq T, \end{cases}$$

where T *is an exponential mean* 1 *random variable*.

(ii) (*Bistability*). *If* $A(x_1) = A(x_3)$ *and a sequence of transition times is defined by* $\tilde{T}_\varepsilon^0 = 0$, *and*

$$T_\varepsilon^i = \inf\{t > \tilde{T}_\varepsilon^{i-1} : \tilde{X}_\varepsilon(t) \in B_c(x_1)\},$$
$$\tilde{T}_\varepsilon^i = \inf\{t > T_\varepsilon^i : \tilde{X}_\varepsilon(t) \in B_c(x_3)\}, \qquad i = 1, 2, \ldots,$$

then $(\nu_t)_{t\geq 0}$ *is given by*

$$\nu_t = \begin{cases} \delta_{x_3}, & T_{2i} \leq t < T_{2i+1}, \\ \delta_{x_1}, & T_{2i+1} \leq t < T_{2i+2}, \end{cases} \qquad i = 0, 1, 2, \ldots,$$

where $T_0 = 0$, *and* $\{T_i\}_{i\geq 0}$ *are arrival times in a rate* 1 *Poisson process*.

PROOF. (i) is simply a restatement of the main result Theorem 2 in [13], and (ii) is an easy extension of this result. Since $A(y_1) = A(y_3)$, we have $\beta_\varepsilon = \tilde{\beta}_\varepsilon$ and the transitions from one stable equilibrium to the other happen on the same exponential time-scale. By Proposition 3.1(i), on the time-scale $T = \beta_\varepsilon t$, in the limit as $\varepsilon \to 0$, T_ε^1 is exponentially distributed with parameter 1, and $\tilde{X}^\varepsilon(T_\varepsilon^1) \in D_1$. By the strong Markov property of $\tilde{X}_\varepsilon$, the time increment to the subsequent transition $\tilde{T}_\varepsilon^1 - T_\varepsilon^1$ is independent of T_ε^1, and the same Theorem implies that on the time-scale $T = \tilde{\beta}_\varepsilon t = \beta_\varepsilon t$, in the limit as $\varepsilon \to 0$, $\tilde{T}_\varepsilon^1 - T_\varepsilon^1$ is also exponentially distributed with parameter 1, and $\tilde{X}^\varepsilon(\tilde{T}_1^\varepsilon) \in D_3$. The rest now follows from the same arguments as in the proof of Theorem 2 in [13]. □

3.2. *Finite-system-size effects*. The above results relied on using an additional parameter ε to separate the scaling of the noise from the scaling of the drift, obtaining a diffusion approximation for the limiting process first, then applying large deviation techniques for the diffusion (15) with small perturbation coefficient ε. A priori, there is no reason why the limits need be taken in that order. Another approach is to apply large deviations techniques directly to the rescaled process $X_N = N^{-1}X_A$, and obtain results that describe the large time-scale behavior of X_N relative to the equilibrium points of the limiting drift (16). It is natural to compare these results to those for the associated diffusion with small diffusion coefficient. We will identify the exact relation of time-scales of the reaction system and the splitting mechanism for which large deviation rates of these two methods can be compared.

This question is most easily answered when the reactions and the splitting/resampling mechanism make only unit net changes at each step, so that X_A is

a birth–death process. Assume, as before, all the reaction constants have the standard scaling $\kappa_\zeta^{ab} = \tilde{\kappa}_\zeta^{ab} N^{1-(a+b)}$, and assume again the splitting rate is of the form $\gamma(x, N, \varepsilon) = \varepsilon_N^2 \gamma(x, N)$, except now the parameter ε_N^2 depends on N as well. Since, by assumption (6), the change due to splitting is unbiased, we have $p_{x,x+1} = p_{x,x-1} = \frac{1}{2}$, and the splitting variance is $\sigma_N^2 = 1$. As earlier $\gamma(x, N)$ is assumed to satisfy condition in assumption (7*), that is, $\sup_x |\gamma(x, N)N^{-2} - \tilde{\gamma}^2\tilde{\sigma}^2(\frac{x}{N})| \to 0$.

Suppose X_A is a Markov jump process with rates $N\tilde{r}_+(x)\,dt = \mathbf{P}[X_A(t+dt) = x+1 | X_A(t) = x]$ and $N\tilde{r}_-(x)\,dt = \mathbf{P}[X_A(t+dt) = x-1 | X_A(t) = x]$ such that $\ln \tilde{r}_+, \ln \tilde{r}_-$ are bounded Lipschitz continuous functions, and $X_N = N^{-1}X_A$ is its rescaled version. Then, according to the Freidlin–Wentzell large deviation theory for Markov jump processes, [27] Theorem 6.17, since transitions between the two stable equilibria x_1, x_3 of $\tilde{\phi}$ are uniquely achieved by crossing the potential barrier at x_2, the deviations of X_N away from neighborhoods of x_1 and x_3 are characterized by the large deviation rate function for X_N given by the quasipotential (with respect to x_i and x_2),

$$\iota_{x_i,x_2}(\tilde{r}_+, \tilde{r}_-) := \inf_{T>0} \inf_{\xi} \left\{ \int_0^T \ell\big(\xi(u), \xi'(u)\big)\,du \,\middle|\, \xi \in C^1\big([0,T]\big), \xi(0) = x_i, \xi(T) = x_2 \right\}, \qquad i = 1, 3,$$

where ℓ is the action functional in variational form

$$\ell(x, y) = \sup_\theta \big\{ \theta y - \big(\tilde{r}_+(x)\big(e^\theta - 1\big) + \tilde{r}_-(x)\big(e^{-\theta} - 1\big)\big) \big\}$$

determined from the jump rates of the process $\tilde{r}_+$ and $\tilde{r}_-$. Calculus of variations results, see [27] Theorem 11.15, give an explicit expression for the quasipotential as

$$\iota_{x_i,x_2}(\tilde{r}_+, \tilde{r}_-) = \int_{x_i}^{x_2} \ln\left(\frac{\tilde{r}_-(x)}{\tilde{r}_+(x)}\right) dx, \qquad i = 1, 3. \tag{18}$$

If X_A is a birth–death process whose rates $r_+(x), r_-(x)$ are such that $r_+^N(x) = N^{-1}r_+(Nx) \to \tilde{r}_+(x)$ and $r_-^N(x) = N^{-1}r_-(Nx) \to \tilde{r}_-(x)$ uniformly in $x \in [0, 1]$, then the logarithmic moment-generating function $g_N(x, \theta)$ of the jump measure $\mu_N(x, \cdot) = r_+^N(x)\delta_1 + r_-^N(x)\delta_{-1}$, for fixed θ, also converges uniformly in $x \in [0, 1]$

$$\begin{aligned} g_N(x, \theta) &= \int \big(e^{\theta z} - 1\big)\mu_N(x, dz) = r_+^N(x)\big(e^\theta - 1\big) + r_-^N(x)\big(e^{-\theta} - 1\big) \\ &\xrightarrow[N\to\infty]{} \tilde{r}_+(x)\big(e^\theta - 1\big) + \tilde{r}_-(x)\big(e^{-\theta} - 1\big) = \int \big(e^{\theta z} - 1\big)\mu(x, dz) = g(x, \theta) \end{aligned}$$

to the logarithmic moment generating function of the jump measure $\mu(x, \cdot) = \tilde{r}_+(x)\delta_1 + \tilde{r}_-(x)\delta_{-1}$. Since the Legendre transform $\ell_N(x, y)$ of $g_N(x, y)$ has the

explicit form

$$\begin{aligned}\ell_N(x,y) &= \sup_\theta\{\theta y - g_N(x,\theta)\}\\ &= \ln\left(\frac{y+\sqrt{y^2+4r_+^N(x)r_-^N(x)}}{2r_+^N(x)}\right)\\ &\quad -\sqrt{y^2+4r_+^N(x)r_-^N(x)}+r_+^N(x)+r_-^N(x)\end{aligned}$$

for fixed y, we also have uniform convergence in $x\in[\delta,1-\delta]$, for any $\delta>0$,

$$\ell_N(x,y)=\sup_\theta\{\theta y-g_N(x,\theta)\}\underset{N\to\infty}{\longrightarrow}\sup_\theta\{\theta y-g(x,\theta)\}=\ell(x,y).$$

Consequently, the large deviation behavior for $X_N=N^{-1}X_A$ is determined by the same action functional $\ell(x,y)$ and exit times in terms of the same quasipotential $\iota_{x_i,x_2}(\tilde r_+,\tilde r_-)$, $i=1,3$ as above.

For the system of reactions and splitting, birth and death rates for the process X_N, r_+ and r_-, respectively, are of the form

$$r_+(x)=N\sum_{(a,b,1)\in\mathcal{I}}\tilde\kappa_1^{ab}\left(\frac{x}{N}\right)_{a,N}\left(1-\frac{x}{N}\right)_{b,N}+\frac{1}{2}\varepsilon_N^2\gamma(x,N), \tag{19}$$

$$r_-(x)=N\sum_{(a,b,-1)\in\mathcal{I}}\tilde\kappa_{-1}^{ab}\left(\frac{x}{N}\right)_{a,N}\left(1-\frac{x}{N}\right)_{b,N}+\frac{1}{2}\varepsilon_N^2\gamma(x,N). \tag{20}$$

We wish to obtain results for the time-scale of exit from a neighborhood of a stable equilibrium for the rescaled process X_N that are analogous to those for $\tilde X_\varepsilon$ obtained in Proposition 3.1. To this end, we will have to make some assumptions about the behavior of r_+ and r_- in order to use the quasipotential $\iota_{x_i,x_2}(\tilde r_+,\tilde r_-)$. Let β_{ε_N} and $\tilde\beta_{\varepsilon_N}$ denote time-scales of the transitions of the process X_N from D_3 to $B_c(x_1)$, and from D_1 to $B_c(x_3)$, respectively, in the analogous way as β_ε and $\tilde\beta_\varepsilon$ were for the singularly perturbed diffusion. The next result establishes the time-scale of transition for X_N from one stable equilibrium to the other.

PROPOSITION 3.3. *If X_A is a birth–death chain, whose rates satisfy*

$$\frac{r_+(N\cdot)}{N}\to\tilde r_+(\cdot),\qquad \frac{r_-(N\cdot)}{N}\to\tilde r_-(\cdot)\qquad \textit{uniformly in } [0,1] \tag{21}$$

such that $\tilde\phi=\tilde r_+-\tilde r_-$ satisfies (16), then the mean times β_{ε_N} and $\tilde\beta_{\varepsilon_N}$ for transitions of X_N from D_3 to $B_c(x_1)$, and from D_1 to $B_c(x_3)$, respectively, are given in terms of $\iota_{x_i,x_2}(\tilde r_+,\tilde r_-)$ from (18) by

$$\lim_{N\to\infty}\frac{1}{N}\ln\beta_{\varepsilon_N}=\iota_{x_3,x_2}(\tilde r_+,\tilde r_-),\qquad \lim_{N\to\infty}\frac{1}{N}\ln\tilde\beta_{\varepsilon_N}=\iota_{x_1,x_2}(\tilde r_+,\tilde r_-).$$

PROOF. This is just the statement of results for the exit problem for the jump Markov chain X_N in terms of its quasipotential, obtained by Freidlin and Wentzell [12]; see Theorems 1.2 and 2.1 of Chapter 5, the discussion at the beginning of Section 4 and Theorem 4.3 of Chapter 5; also see Theorem 5.7.11 of Chapter 5 in [7]. Uniform convergence of the action potential, that is, the Legendre transform ℓ_N, is necessary in order to express the quasipotential ι_{x_i,x_2} in terms of the limiting rates $\tilde{r}_+, \tilde{r}_-$. All of the assumptions on the equilibrium points of $\tilde{\phi}(x) = \tilde{r}_+(x) - \tilde{r}_-(x)$ in (16) are also necessary, since $\tilde{\phi}$ determines the fluid limit of the jump Markov chain X_N. □

Finally, we can establish the time-scale separation under which the switching results for the rescaled jump process X_N and the diffusion $\tilde{X}_\varepsilon$ with the small diffusion coefficient can be compared.

THEOREM 3.1. *If the reaction system has increments of size* $\{1, -1\}$ *only, its rates have standard scaling* $\kappa_\zeta^{ab} = \tilde{\kappa}_\zeta^{ab} N^{1-(a+b)}$, *its limiting drift* $\tilde{\phi}$ *satisfies* (16) *and if the splitting mechanism has increments of size* $\{1, -1\}$, *its rate is* $\varepsilon_N^2 \gamma(x, N)$ *where* $\gamma(x, N)$ *satisfies assumption* (7*) *and*

$$N\varepsilon_N^2 \to 1,$$

then results based on large deviations for X_N *in Proposition* 3.3 *are more informative than results based on large deviations for the diffusion* $\tilde{X}_\varepsilon$ *with the small perturbation parameter* ε_N *in Proposition* 3.1, *that is*,

$$\iota_{x_i,x_2}(\tilde{r}_+, \tilde{r}_-) \le I_{x_i,x_2}(\tilde{\phi}, \tilde{\gamma}\tilde{\sigma}).$$

PROOF. For $r_+^N(x) = N^{-1} r_+(Nx)$ and $r_-^N(x) = N^{-1} r_-(Nx)$ by (19)–(20), we have

$$r_+^N(x) = \sum_{(a,b,1)\in\mathcal{I}} \tilde{\kappa}_1^{ab}(x)_{a,N}(1-x)_{b,N} + \frac{1}{2} N^{-1} \varepsilon_N^2 \gamma(Nx, N),$$

$$r_-^N(x) = \sum_{(a,b,-1)\in\mathcal{I}} \tilde{\kappa}_{-1}^{ab}(x)_{a,N}(1-x)_{b,N} + \frac{1}{2} N^{-1} \varepsilon_N^2 \gamma(Nx, N).$$

Since $\gamma(x, N)$ is such that $|\gamma(Nx, N)N^{-2} - \tilde{\gamma}^2\tilde{\sigma}^2(x)| \to 0$ uniformly in $x \in \{0, \frac{1}{N}, \dots, 1\}$, then given that $N\varepsilon_N^2 \to 1$, we have uniform convergence of $r_+^N \to \tilde{r}_+$ and $r_+^N \to \tilde{r}_+$ to

$$\tilde{r}_+(x) = \sum_{(a,b,1)\in\mathcal{I}} \tilde{\kappa}_1^{ab} x^a (1-x)^b + \frac{1}{2}\tilde{\gamma}^2\tilde{\sigma}^2(x),$$

$$\tilde{r}_-(x) = \sum_{(a,b,-1)\in\mathcal{I}} \tilde{\kappa}_{-1}^{ab} x^a (1-x)^b + \frac{1}{2}\tilde{\gamma}^2\tilde{\sigma}^2(x).$$

Let $\omega(x) = 1 - \frac{\tilde{r}_-(x)}{\tilde{r}_+(x)}$, so

$$\omega(x) = \frac{\tilde{r}_+(x) - \tilde{r}_-(x)}{\tilde{r}_+(x)} = \frac{\tilde{\phi}(x)}{\sum_{(a,b,1)\in\mathcal{I}} \tilde{\kappa}_1^{ab} x^a (1-x)^b + (1/2)\tilde{\gamma}^2\tilde{\sigma}^2(x)},$$

$$\frac{\omega(x)}{1-\omega(x)} = \frac{\tilde{r}_+(x) - \tilde{r}_-(x)}{\tilde{r}_-(x)} = \frac{\tilde{\phi}(x)}{\sum_{(a,b,-1)\in\mathcal{I}} \tilde{\kappa}_{-1}^{ab} x^a (1-x)^b + (1/2)\tilde{\gamma}^2\tilde{\sigma}^2(x)}$$

and (18) implies that $\iota_{x_i,x_2}(\tilde{r}_+, \tilde{r}_-) = \int_{x_i}^{x_2} \ln(1-\omega(x))\,dx$ satisfies

$$\begin{aligned} &-\int_{x_i}^{x_2} \frac{\tilde{\phi}(x)\,dx}{\sum_{(a,b,-1)\in\mathcal{I}} \tilde{\kappa}_{-1}^{ab} x^a (1-x)^b + (1/2)\tilde{\gamma}^2\tilde{\sigma}^2(x)} \\ &\quad\le \iota_{x_i,x_2}(\tilde{r}_+, \tilde{r}_-) \\ &\quad\le -\int_{x_i}^{x_2} \frac{\tilde{\phi}(x)\,dx}{\sum_{(a,b,1)\in\mathcal{I}} \tilde{\kappa}_1^{ab} x^a (1-x)^b + (1/2)\tilde{\gamma}^2\tilde{\sigma}^2(x)}. \end{aligned}$$

On the other hand by (17) and the fact that $g'(y) = \tilde{\gamma}\tilde{\sigma}(g(y))$ we also have

$$\begin{aligned} I_{x_i,x_2}(\tilde{\phi}, \tilde{\sigma}) &= -2\int_{y_i}^{y_2} \alpha(y)\,dy = -2\int_{y_i}^{y_2} \frac{\tilde{\phi}(g(y))\,dy}{\tilde{\gamma}\tilde{\sigma}(g(y))} = -\int_{x_i}^{x_2} \frac{\tilde{\phi}(x)\,dx}{(1/2)\tilde{\gamma}^2\tilde{\sigma}^2(x)} \\ &\ge \iota_{x_i,x_2}(\tilde{r}_+, \tilde{r}_-). \end{aligned}$$

Hence if $N\varepsilon_N^2 \to 1$, we get a comparison using quasipotentials for X_N and $\tilde{X}_\varepsilon$ of the time-scales for transitions between stable equilibria, as

$$\ln \beta_\varepsilon \approx \frac{1}{\varepsilon_N^2} I_{x_i,x_2}(\tilde{\phi}, \tilde{\sigma}) \gtrsim N\iota_{x_i,x_2}(\tilde{r}_+, \tilde{r}_-) \approx \ln \beta_{\varepsilon_N}. \qquad \square$$

If $\varepsilon_N^2 = N^{-1}$, transitions between stable equilibria are more often due to finite-system-size effects than due to the effects of an additional mechanism. This is understandable in light of the fact that the diffusion $\tilde{X}_\varepsilon$ is a limit of the rescaled process X_N in which the contribution of any subdiffusive noise disappears. As remarked earlier, when $\varepsilon_N^2 = N^{-1}$, we could use this informally prior to obtaining a diffusion limit $\tilde{X}_\varepsilon$. If, for rates of balanced reactions we write $\kappa_\zeta^{ab}(N) = N^{1-(a+b)}\tilde{\kappa}_\zeta^{ab} = \varepsilon_N^2 N^{2-(a+b)}\tilde{\kappa}_\zeta^{ab}$, then the diffusion coefficient would become $\tilde{a}_\varepsilon(x) = \varepsilon_N^2(\sum_{(a,b,\zeta)\in I^{\mathrm{bal}}} \tilde{\kappa}_{\zeta,\mu}^{ab} x^a (1-x)^b + \tilde{\gamma}^2\tilde{\sigma}^2(x))$. However, even this "adjusted" diffusion coefficient would not change the conclusion of Theorem 3.1, since the contribution of the rates from biased reactions is still missing from the quasipotential of $\tilde{X}_\varepsilon$.

If $N^{-1} \ll \varepsilon_N^2 \ll 1$, it is clear from Theorem 3.1 that the noise of the splitting is the dominant factor in effecting the transitions, while if $\varepsilon_N^2 \ll N^{-1}$, the noise from reactions dominates, and both rates $\tilde{r}_+, \tilde{r}_-$ and the quasipotential ι_{x_i,x_2} are determined by the reaction system only.

3.3. *Example*: *bistable behavior from slow splitting*. Here is an example of a simple reaction system that yields a limiting system with a double-well potential:

$$A \overset{\kappa_{-1}^{10}}{\to} B, \tag{22}$$

$$B \overset{\kappa_{1}^{01}}{\to} A, \tag{23}$$

$$A + B \overset{\kappa_{-1}^{11}}{\to} 2B, \tag{24}$$

$$2A + B \overset{\kappa_{1}^{21}}{\to} 3A. \tag{25}$$

The trimolecular reaction (25) produces a term in the drift which is cubic in X_A, which is needed in order to obtain the three desired equilibria. With standard mass-action scaling $\kappa_\zeta^{ab} = N^{1-(a+b)}\tilde{\kappa}_\zeta^{ab}$, the limit of $F_N(X_N(t)) = \mathbf{E}[X_N(t)] = \mathbf{E}[X_A(t)/N] \in [0,1]$ as $N \to \infty$ is

$$\tilde{\phi}(x) = \lim_{N\to\infty} F_N(x) = -\tilde{\kappa}_{-1}^{10}x + \tilde{\kappa}_1^{01}(1-x) - \tilde{\kappa}_{-1}^{11}x(1-x) + \tilde{\kappa}_1^{21}x^2(1-x), \qquad x \in [0,1].$$

With the special choice of $\tilde{\kappa}_{-1}^{10} = \tilde{\kappa}_1^{01} = 1, \tilde{\kappa}_{-1}^{11} = \frac{16}{3}, \tilde{\kappa}_1^{21} = \frac{32}{3}$ we have

$$\tilde{\phi}(x) = \tfrac{1}{3}\big(3 - 22x + 48x^2 - 32x^3\big) = -\tfrac{32}{3}\big(x - \tfrac{1}{4}\big)\big(x - \tfrac{1}{2}\big)\big(x - \tfrac{3}{4}\big) \tag{26}$$

with two stable points at $x_1 = \frac{1}{4}$ and $x_3 = \frac{3}{4}$ and one unstable point at $x_2 = \frac{1}{2}$ for the system, and thus $\Phi = -\int \tilde{\phi}$ is a double-well potential. Since $\tilde{\phi}$ is antisymmetric about the line $x = \frac{1}{2}$ the potential can be expressed as

$$\Phi(x) = \tfrac{1}{6}(2x-1)^4 - \tfrac{1}{12}(2x-1)^2 + C,$$

which is symmetric about the line $x = \frac{1}{2}$, and thus Φ has equally deep wells $\Phi(\frac{1}{4}) = \Phi(\frac{3}{4})$.

This system bears resemblance to the so-called Schlögl model [26], which consists of four reactions $A + 2X \rightleftharpoons 3X$, $B \rightleftharpoons X$, with the resulting drift for X cubic. In [28] the authors formulate the Kolmogorov forward equation (chemical master equation) to analyze the stochastic model for this reaction system.

For this example we take the simplest splitting/resampling mechanism [Example (Bern) in Section 2.2] in which at each split an error in the molecular count of A from the parent to the daughter cell is at most 1. Its rate is $\gamma(x,N) = \gamma(N)x/N(1-x/N)$ and its probabilities are $p_{x,x+1} = p_{x,x-1} = 1/2$ for $x \neq 0, N$, and $p_{0,0} = p_{N,N} = 1$. Note that here the factor $\gamma(N)$ will depend on N, but is state independent. This mechanism can also be represented in terms of reactions as

$$A + B \overset{N^{-2}\gamma(N)}{\longrightarrow} 2A, \qquad A + B \overset{N^{-2}\gamma(N)}{\longrightarrow} 2B. \tag{27}$$

We stress that this representation (27) of the resampling in terms of reactions is done merely to illustrate the mechanism in a similar way to the reactions, and is not to be confused with an actual set of biological reactions as in (22)–(25). This can be done in the particular case of Moran-type resampling, since the rates of this mechanism depend on the product of both the count of A and of B. This is a consequence of the fact that each resampling event picks either one molecule of A or one molecule of B with probabilities relative to their proportions in the cell, and replaces it in the daughter cell with a random choice of either A or B with equal probability.

As shown in Section 2.2, if we choose the splitting parameter to be $\gamma(N) = \frac{1}{2}\varepsilon^2 N^2$ for some small constant $\varepsilon^2 > 0$, then $\gamma(x, N)$ satisfies all the conditions of assumption (7*), and the limiting process $\tilde{X}_\varepsilon$ satisfies the stochastic differential equation with drift (26) and diffusion coefficient $\varepsilon^2 \frac{1}{2} x(1-x)$

$$
\begin{aligned}
d\tilde{X}_\varepsilon(t) &= \tfrac{1}{3}\big(3 - 22\tilde{X}_\varepsilon(t) + 48\tilde{X}_\varepsilon^2(t) - 32\tilde{X}_\varepsilon(t)^3\big)\,dt \\
&\quad + \varepsilon\sqrt{\tilde{X}_\varepsilon(t)\big(1-\tilde{X}_\varepsilon(t)\big)}\,dB(t).
\end{aligned} \tag{28}
$$

To find the value of the quasipotential for this problem we find the transformation of the potential via $\alpha(y) = \tilde{\phi}(g(y))/\tilde{\sigma}(g(y))$, where g is the solution to $g'(y) = \tilde{\sigma}(g(y)) = \sqrt{g(y)(1-g(y))}$, given explicitly by

$$
g(y) = \cos^2\left(\frac{1}{2}\left(y - \frac{\pi}{2}\right)\right) = \cos^2\left(\frac{y}{2} - \frac{\pi}{4}\right), \qquad y \in \left[-\frac{\pi}{2}, \frac{\pi}{2}\right].
$$

We chose the constant of integration so that $g(0) = \frac{1}{2}$, and $g(-y) = 1 - g(y)$. The inverse of g is given by

$$
h(x) = g^{-1}(x) = 2\arctan\left(-\sqrt{\frac{1}{x} - 1}\right) + \frac{\pi}{2}, \qquad x \in [0, 1];
$$

hence, the transformed equilibrium points $y_i = h(x_i)$ are

$$
y_1 = 2\arctan(-\sqrt{3}) + \frac{\pi}{2} = -\frac{\pi}{6}, \qquad y_3 = 2\arctan\left(-\sqrt{\frac{1}{3}}\right) + \frac{\pi}{2} = \frac{\pi}{6}
$$

and

$$
y_2 = 2\arctan(-1) + \frac{\pi}{2} = 0.
$$

Note as well that the wells of the transformed potential are of equal depth, which follows from the fact that α is an odd function

$$
\alpha(-y) = \frac{\tilde{\phi}(g(-y))}{\tilde{\sigma}(g(-y))} = \frac{\tilde{\phi}(1-g(y))}{\tilde{\sigma}(1-g(y))} = \frac{-\tilde{\phi}(g(y))}{\tilde{\sigma}(g(y))} = -\alpha(y),
$$

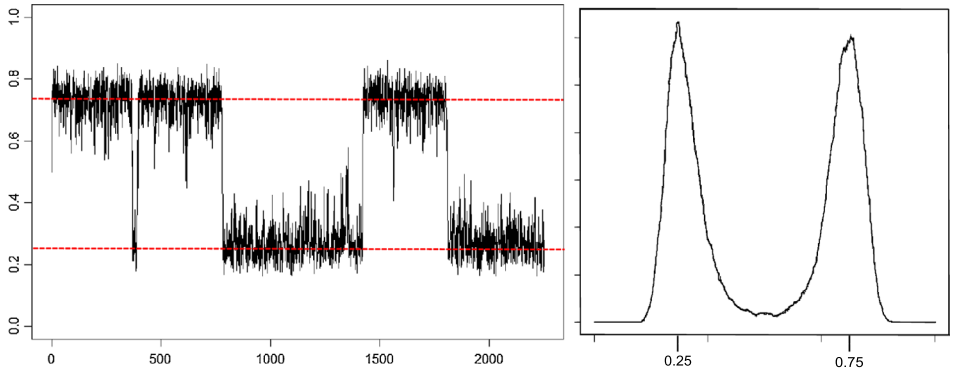

FIG. 1. *Sample path $X_N(t)$ (left: x-axis $= t$, y-axis $= X_N(t) = N^{-1}X(t)$) and its occupation density (right: x-axis $=$ state space of $X_N \subset [0,1]$, y-axis $=$ proportion of time X_N spends in each state by time $t = 2500$) from the system (22)–(25) with birth–death splitting, under standard mass-action scaling for reactions and $\gamma(\varepsilon, N) = \frac{1}{2}\varepsilon^2 N^2$ (parameters $N = 1500$, $\varepsilon^2 = 0.02$). Dashed red lines indicate quasi-equilibria at $1/4$ and $3/4$.*

and thus $A = -\int \alpha(y)\,dy$ is an even function. Since $y_1 = -y_3$, $y_2 = 0$, and $A(y_1) = A(y_3) = 0$, we have $A(y_2) = \int_0^{\pi/6} \alpha(y)\,dy$ with a rather complicated expression

$$A(y_2) = \int_0^{\pi/6} \frac{\tilde{\phi}(\cos^2((y/2) - (\pi/4)))}{\tilde{\sigma}(\cos^2((y/2) - (\pi/4)))}\,dy = \frac{1}{3}\int_{1/2}^{3/4} \frac{3 - 22x + 48x^2 - 32x^3}{x(1-x)}\,dx$$
$$\doteq 0.0913.$$

By Proposition 3.1 on a time-scale of $O(e^{-\varepsilon^{-2}2A(y_2)})$, the process exists a neighborhood of the stable equilibria $x_1 = \frac{1}{4}, x_3 = \frac{3}{4}$. Symmetry of A around $y_2 = 0$ implies that we are in the bistable case (ii) of Proposition 3.2, and the occupation measure of the process $\tilde{X}_\varepsilon$ converges to the occupation measure of a two-state Markov chain, which transitions between states $\{\frac{1}{4}, \frac{3}{4}\}$ with equal rates. Figure 1 shows an exact simulation of a sample path of the rescaled process $X_N = N^{-1}X_A$ with choice of parameters $N = 1500$, $\varepsilon^2 = 0.02$; since $\varepsilon^2 \gg 1/N$, we expect the ε-perturbation of the limiting diffusion to be driving the switching. Indeed, the process appears to be spending most of its time in neighborhoods $B_{0.1}(x_1) \cup B_{0.1}(x_3)$, switching between them at the approximate rate $R = e^{-\varepsilon^{-2}2A(y_2)} \doteq 0.0001083$.

If we take $\varepsilon^2 \ll 1/N$, then transitions between stable equilibria are based only on the scaled rates for the reaction system (22)–(25),

$$\tilde{r}_+(x) = 1 - x + \tfrac{32}{3}x^2(1-x) \quad \text{and} \quad \tilde{r}_-(x) = x + \tfrac{16}{3}x(1-x).$$

By Proposition 3.3 the values of the quasipotential for the birth–death Markov process are

$$\iota_{x_1,x_2} = \int_{x_1}^{x_2} \ln\left(\frac{\tilde{r}_-(x)}{\tilde{r}_+(x)}\right)dx = \int_{1/4}^{1/2} \ln\left(\frac{x + (16/3)x(1-x)}{1 - x + (32/3)x^2(1-x)}\right)dx \doteq 0.006713$$

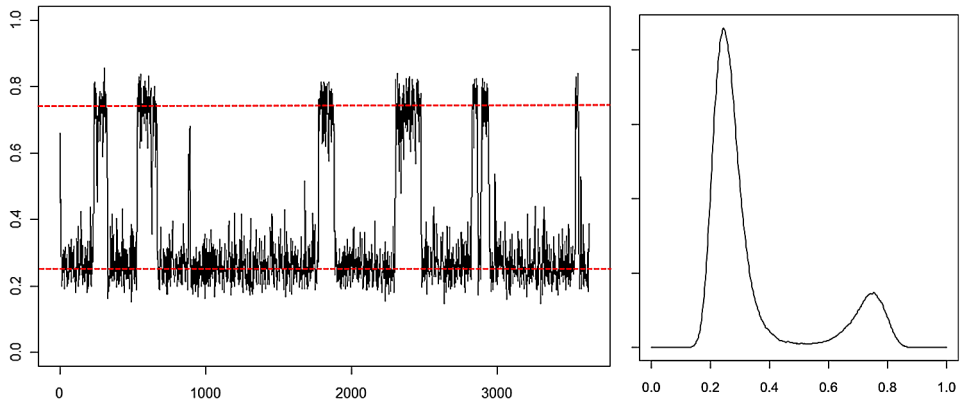

FIG. 2. *Sample path* $X_N(t)$ [*left*: *x-axis* $= t$, *y-axis* $= X_N(t)$] *and its occupation density* (*right*: *x-axis* = *state space of* $X_N \subset [0,1]$, *y-axis* = *proportion of time* X_N *spends in each state by time* $t = 4000$) *for the system* (22)–(25) *with birth–death splitting, under standard mass-action scaling for reactions and* $\gamma(\varepsilon, N) = \frac{1}{2}\varepsilon^2 N^2$ (*parameters* $N = 500$, $\varepsilon^2 = 2 \times 10^{-4}$). *Dashed red lines indicate quasi-equilibria at* $1/4$ *and* $3/4$ *as above.*

and

$$\iota_{x_3,x_2} = \int_{x_3}^{x_2} \ln\left(\frac{\tilde{r}_-(x)}{\tilde{r}_+(x)}\right) dx \doteq 0.005534.$$

Note that here the values for the quasipotential are no longer equal, and the process will take longer to get out of the neighborhood of the equilibrium $x_1 = \frac{1}{4}$. Figure 2 shows a simulation of a sample path of the rescaled process X_N for $\gamma(N) = \varepsilon^2 N$ with the choice of parameters $N = 500$, but $\varepsilon^2 = 2 \times 10^{-4}$. In this case $1/N \gg \varepsilon^2$ and we expect the transitions to be due to noise from the reactions arising from finite-N effects. Based on the above calculation we expect the process to be switching away from $B_{0.1}(x_1)$ at rate $R = e^{-N\iota_{x_1,x_2}} = e^{-3.356629} \doteq 0.035$ and away from $B_{0.1}(x_3)$ at a rate $R' = e^{-N\iota_{x_3,x_2}} = e^{-2.769957} \doteq 0.062$; indeed, the time spent near x_1 is appreciably larger than the time spent near x_3.

We make a particular note that the reaction system considered here is very sensitive to the exact values given for the reaction constants; a small change in these would preserve the double-well potential, but would lead to nonequal depth of the two wells for the quasipotential, and hence instead of a limiting bistable behavior would lead to a limiting metastable behavior as in case (i) of Proposition 3.2. In the next section we discuss the conditions on the scaling of the reaction and splitting/resampling which yield behavior that can also be described as bistable, but where the underlying mechanism is qualitatively different and the restrictions on the reaction system are negligible.

4. Bistable behavior from fast splitting. We next consider the case $\varepsilon_A \approx \infty$, and assume that time has been rescaled so that $\tilde{c}_{\sigma^2} = \lim_{N\to\infty} c_{\sigma^2}(N) \in (0, \infty)$

and $\tilde{c}_\mu = \lim_{N\to\infty} c_\mu(N) \approx 0$. This is a more unconventional scaling, in which the noise (from balanced reactions and splitting) overwhelms the contribution due to the drift (from biased reactions).

One way to model this with a diffusion would be to introduce a time-scale separation with an additional small parameter ε in the scaling of all reactions rather than in the rate of splitting. Suppose all reaction constants scale as $\kappa_\zeta^{ab} = \varepsilon\tilde{\kappa}_\zeta^{ab} N^{1-(a+b)}$, while the rate of splitting $\gamma(x, N)$ satisfies assumption (7*). For any fixed $\varepsilon > 0$, the resulting limit of the rescaled process X_N would be

$$d\tilde{X}^\varepsilon(t) = \varepsilon\tilde{\phi}\big(\tilde{X}^\varepsilon(t)\big)\,dt + \tilde{\gamma}\tilde{\sigma}\big(\tilde{X}^\varepsilon(t)\big)\,dB(t), \qquad \tilde{X}^\varepsilon \in [0,1], \tag{29}$$

where B is a standard Brownian motion, and we have the case of a diffusion with a small drift. Note that although $\tilde{\sigma}^2(0) = \tilde{\sigma}^2(1) = 0$ [by assumption (7)], the boundaries $\{0, 1\}$ are not absorbing, since there is at least one biased reaction that allows escape from either boundary $\tilde{\phi}(0) > 0, \tilde{\phi}(1) < 0$ [by assumption (2)]. Other than at the boundaries the contribution of the drift is essentially negligible, and $\tilde{X}^\varepsilon$ is approximately a martingale. Most attempts to escape a boundary are followed by the return to the same boundary point; only some end up at the opposite one. In the limit as $\varepsilon \to 0$, the rate of escapes from the boundaries for $\tilde{X}^\varepsilon$ vanishes, and there is no switching.

However, under the right conditions, the limit of the original rescaled process will spend almost all of its time at one boundary or the other, switching between the two on a reasonable time-scale, creating again a bistable system. How the effect of the attempts to escape the boundary appears in the limit depends on the rate of the attempts, and the time spent between the boundaries. In order to make a precise statement we need to examine the behavior of the rescaled process X_N directly and specify a general set of conditions for a Markov jump process to exhibit this type of switching behavior.

4.1. *Stochastic switching.* The unscaled process X_A is a Markov chain on $\{0, 1, \ldots, N\}$ with transitions that are due to the reactions $(a, b, \zeta) \in \mathcal{I}$, as well as the splitting mechanism with distribution $p_{x,y}$, $(x, y) \in \{0, 1, \ldots, N\}^2$. The rates of these transitions from $X_A = x$ are equal to $\sum_{(a,b,\zeta)\in\mathcal{I}} \lambda_\zeta^{ab}(x) = \sum_{(a,b,\zeta)\in\mathcal{I}} \kappa_\zeta^{ab}(N)(x)_a(N-x)_b$ from the reaction system and $\gamma(x, N)$ from the splitting, respectively. We denote the total combined rate of X_A from $i \in \{0, \ldots, n\}$ to $j \in \{0, \ldots, n\}$ by

$$r_{ij} = \sum_{(a,b,j-i)\in\mathcal{I}} \kappa_{j-i}^{ab}(N) i_a (N-i)_b + \gamma(i, N) p_{i,j}.$$

Transitions due to splitting can have jumps whose size can in principle be as large as $N - 1$ (such as those of the Wright–Fisher splitting process example in Section 2.2), although with very small probability. However, a splitting mechanism is absorbing at $\{0, N\}$, $p_{0,0} = p_{N,N} = 1$, and the rates of jumps off the boundaries

$x \in \{0, N\}$ are created by reactions using only molecules of B (for $x = 0$), or using only molecules of A (for $x = N$), with rates

$$r_{0j} = \sum_{(0,b,j)\in\mathcal{I}} \kappa_j^{0j}(N)(N)_b, \qquad r_{Nj} = \sum_{(a,0,N-j)\in\mathcal{I}} \kappa_{N-j}^{a0}(N)(N)_a.$$

By assumption (2) in Section 2.1, there exist $j, j' \in \{1, \dots, N-1\}$ such that $r_{0j}, r_{Nj'} \neq 0$. The leading powers of N, $\max_{(0,b,j)\in\mathcal{I}}\{b\} > 0$ and $\max_{(a,0,N-j)\in\mathcal{I}}\{a\} > 0$, respectively, will determine the rate at which attempts to counteract absorption at the boundaries happen, and in particular, this implies that $r_{0j}, r_{Nj} \to \infty$ as $N \to \infty$ [allowing for upcoming condition (30)].

Define an *excursion* of X_A to be any segment $X_A(t), t \in [t_1, t_2)$ such that $X_A(t_1-), X_A(t_2) \in \{0, N\}$ and $X_A(t) \notin \{0, N\}$ for $t \in [t_1, t_2)$. Call an excursion on $[t_1, t_2)$ "successful" if $X_A(t_1-) \neq X_A(t_2)$, and "unsuccessful" otherwise. For $0 \le j \le N$, let $\tau_j := \inf\{t \ge 0 : X_A(t) = j\}$ be the first hitting time of state j, and let $\tau_{0,N} = \tau_0 \wedge \tau_N$ denote the first hitting time of either boundary state. Let

$$e_{j0} = \mathbf{E}\big[\tau_{0,N} | X_A(0) = j, X_A(\tau_{0,N}) = 0\big],$$
$$e_{jN} = \mathbf{E}\big[\tau_{0,N} | X_A(0) = j, X_A(\tau_{0,N}) = N\big]$$

be the expected hitting time of the two boundaries from j and π_{jN} be the probability that an excursion from j hits the N boundary first

$$\pi_{jN} = \mathbf{P}\big[X_A(\tau_{0,N}) = N | X_A(0) = j\big],$$

and thus $\pi_{j0} = 1 - \pi_{jN}$ is the probability it first hits the 0 boundary. The values of $\{e_{j\cdot}, \pi_{jN}\}_{j\in\{1,\dots,N-1\}}$ can be determined by setting up and solving the appropriate linear functionals of the generator for the Markov process X_A; explicit expressions, however, may be hard to come by for general processes.

Excursions of X_A depend on transitions from both reactions and the splitting mechanism. However, if the noise overwhelms the drift, then at each step in the interior transition rates are dominated by those from the balanced reactions and the splitting mechanism. In particular, this will imply that in the interior X_A behaves approximately like a martingale, and will allow us to approximate the probability of switching from one boundary point to the other in terms of the relative rates of biased reactions versus balanced reactions and splitting. We will estimate e_{j0}, e_{jN}, π_{jN} in an example to come, and exhibit more explicit conditions than the ones below in the case when the reactions and splitting yield a birth–death process for X_A.

We first state general conditions under which the rescaled process $X_N = X_A/N$ can be approximated by a simple Markov jump process. Suppose that there exists two scaling parameters: the order of magnitude of the rate of reactions on the boundary $\omega_N \to \infty$, and a time scaling parameter $\beta_N > 0$ for the rescaled process

X_N, such that

$$\frac{1}{\omega_N}\sum_j r_{0j} \to \tilde{r}_+, \qquad \frac{1}{\omega_N}\sum_j r_{Nj} \to \tilde{r}_-, \tag{30}$$

$$\beta_N \sum_j r_{0j}\pi_{jN} \to \tilde{r}_{01}, \qquad \beta_N \sum_j r_{Nj}\pi_{j0} \to \tilde{r}_{10}, \tag{31}$$

$$\begin{aligned} &\frac{1}{\beta_N\omega_N}\sum_j r_{0j}e_{jN}, \qquad \sum_j r_{0j}e_{j0}, \\ &\frac{1}{\beta_N\omega_N}\sum_j r_{Nj}e_{j0}, \qquad \sum_j r_{Nj}e_{jN} \to 0, \end{aligned} \tag{32}$$

with $\tilde{r}_+, \tilde{r}_-, \tilde{r}_{01}, \tilde{r}_{10} \in (0,\infty)$. Since $r_{0j}, r_{Nj} \to \infty$, there is no need to change the time-scale for the process. These conditions imply that there are many excursions in any finite time interval $[0,t]$, only a small fraction of which are successful, and during which the total time spent is very small. Consequently, the rescaled process will spend most of its time on one boundary until the first time a successful excursion takes it to the other boundary. Let $\tilde{T}_N^0 = \inf\{t \ge 0 : X_A(t) = 0\}$, and

$$T_N^i = \inf\bigl\{t > \tilde{T}_N^{i-1} : X_A(t) = N\bigr\}, \qquad \tilde{T}_N^i = \inf\bigl\{t > T_N^i : X_A(t) = 0\bigr\}, \qquad i = 1, 2\ldots$$

be a sequence of times at which X_A first reaches a boundary different from the one where it was most recently. Also, define the measure-valued process $(\nu_t^N)_{t\ge 0}$ for some $\rho_N > 0$ such that $\frac{\rho_N}{\beta_N} \to 0$ by

$$\nu_t^N(f) = \frac{1}{\rho_N}\int_{\beta_N t}^{\beta_N t+\rho_N} f\bigl(X_N(s)\bigr)\,ds$$

for any (bounded continuous) function f on $\{0, \frac{1}{N}, \ldots, 1\}$; this (ν_t^N) approximates the law of the location of the rescaled process $X_N(t) = X_A(t)/N$ on a short time interval of length ρ_N.

PROPOSITION 4.1. *If X_A satisfies* (30)–(32), *then*

$$\lim_{N\to\infty} \mathbf{P}\bigl[T_N^i - \tilde{T}_N^{i-1} > t\beta_N\bigr] = e^{-\tilde{r}_{01}t}, \qquad \lim_{N\to\infty} \mathbf{P}\bigl[T_N^{i+1} - \tilde{T}_N^i > t\beta_N\bigr] = e^{-\tilde{r}_{10}t} \qquad \forall t > 0,$$

and we have convergence in law on the space of cadlag paths (*with the Skorokhod topology*) $(\nu_t^N)_{t\ge 0} \Rightarrow (\nu_t)_{t\ge 0}$ *to a jump process*

$$\nu_t = \begin{cases} \delta_0, & T_{2i} \le t < T_{2i+1}, \\ \delta_1, & T_{2i+1} \le t < T_{2i+2}, \end{cases} \qquad i = 0, 1, 2, \ldots,$$

where $\{T_{2i+1} - T_{2i}\}_{i\ge 0}$ and $\{T_{2i+2} - T_{2i+1}\}_{i\ge 0}$ are two independent sequences of i.i.d. exponential variables with rates $\tilde{r}_{01}$ and $\tilde{r}_{10}$, respectively.

The rescaled process X_N can therefore be approximated by a jump Markov process $(J(t))_{t\geq 0}$ on $\{0,1\}$ with transition rates $\tilde{r}_{01}$ from $0\to 1$, and $\tilde{r}_{10}$ from $1\to 0$ in the following sense: the occupation times of X_N on $\{0,1\}$ converge to the respective occupation times of J, and the times of successful excursions of X_N from $0\to 1$ and from $1\to 0$ converge to the respective transitions of J. We cannot expect a stronger kind of convergence than stated, since, for example, convergence in law of X_N to J in the Skorokhod topology is precluded by the fact that for arbitrarily large N, there remain unsuccessful excursions of X_N that stray from their originating boundary by a distance which is bounded away from 0.

A different set of conditions from those in (32) for the length of excursions away from the boundaries, where in the limit we get four nonzero limiting constants $\tilde{e}_{01},\tilde{e}_{00},\tilde{e}_{10},\tilde{e}_{11}$, would imply convergence to a limiting process which spends a nontrivial fraction of time away from the boundary. The limiting process would behave similar to a diffusion with "sticky" boundaries; see [16], Section 15.8C.

PROOF OF PROPOSITION 4.1. For each $i\geq 0$, define a sequence of times after $\tilde{T}_N^i$ at which excursions from 0 start $\tilde{\sigma}_N^{i,i'}$ and end $\tilde{\tau}_N^{i,i'}$, by letting $\tilde{\tau}_N^{i,0}=\tilde{T}_N^i$, and for $i'=1,2,\ldots$

$$\tilde{\sigma}_N^{i,i'}=\inf\{\tilde{\tau}_N^{i,i'-1}<t:X_A(t)\neq 0,X_A(t-)=0\},$$
$$\tilde{\tau}_N^{i,i'}=\inf\{\tilde{\tau}_N^{i,i'-1}<t:X_A(t)=0,X_A(t-)\neq 0\}$$

and let $s(i)=\inf\{i'\geq 1:\tilde{\tau}_N^{i,i'}>T_N^{i+1}\}$ be the index of the first excursion from 0 that is successful, hence $\tilde{\tau}_N^{i,s(i)}=\tilde{T}_N^{i+1}$. Note that $X_A(t)=0,\forall t\in[\tilde{\tau}_N^{i,i'-1},\tilde{\sigma}_N^{i,i'})$ and that $\sum_{i'\leq s(i)}(\tilde{\sigma}_N^{i,i'}-\tilde{\tau}_N^{i,i'-1})$ is the time spent at 0 between successful excursions, while $X_A\neq 0$ for $t\in[\tilde{\sigma}_N^{i,i'},\tilde{\tau}_N^{i,i'})$, and thus $\sum_{i'<s(i)}(\tilde{\tau}_N^{i,i'}-\tilde{\sigma}_N^{i,i'})$ is the time spent on unsuccessful excursions.

Consider the time interval $[\tilde{T}_N^i,T_N^{i+1}]-\bigcup_{i'<s(i)}[\tilde{\sigma}_N^{i,i'},\tilde{\tau}_N^{i,i'})$ from which subintervals for unsuccessful excursions are excised. Excursions from 0 are started at overall rate $\sum_{j'}r_{0j'}$, and since excursions whose first step is to j are successful with probability π_{jN}, successful excursions are started at rate $\sum_j r_{0j}\pi_{jN}$. So

$$W_N^i:=\tilde{\sigma}_N^{i,s(i)}-\tilde{T}_N^i-\sum_{i'<s(i)}(\tilde{\tau}_N^{i,i'}-\tilde{\sigma}_N^{i,i'})\sim\text{exponential}\Big(\sum_j r_{0j}\pi_{jN}\Big)$$

and

$$s(i)\sim\text{geometric}\Big(\frac{\sum_j r_{0j}\pi_{jN}}{\sum_j r_{0j}}\Big).$$

Also, for any $i'<s(i)$, the unsuccessful excursion times $\tilde{\tau}_N^{i,i'}-\tilde{\sigma}_N^{i,i'}$ are independent and identically distributed with

$$\mathbf{E}[\tilde{\tau}_N^{i,i'}-\tilde{\sigma}_N^{i,i'}]=\sum_j\frac{r_{0j}}{\sum_{j'}r_{0j'}}\mathbf{E}[\tau_{0,N}|X_A(0)=j,X_A(\tau_{0,N})=0],$$

while $T_N^{i+1} - \tilde{\sigma}_N^{i,s(i)}$ is a subinterval for a successful excursion with

$$\mathbf{E}\big[T_N^{i+1} - \tilde{\sigma}_N^{i,s(i)}\big] = \sum_j \frac{r_{0j}}{\sum_{j'} r_{0j'}} \mathbf{E}\big[\tau_{0,N} | X_A(0) = j,\, X_A(\tau_{0,N}) = N\big].$$

Let

$$U_N^i := \sum_{i' < s(i)} \big(\tilde{\tau}_N^{i,i'} - \tilde{\sigma}_N^{i,i'}\big) \quad \text{and} \quad S_N^i := T_N^{i+1} - \tilde{\sigma}_N^{i,s(i)},$$

so that $T_N^{i+1} - \tilde{T}_N^i = W_N^i + U_N^i + S_N^i$. Assumption (31) implies $W_N^i/\beta_N \Rightarrow$ exponential$(\tilde{r}_{01})$ as $N \to \infty$. We next show convergence for both $U_N^i \to 0$ and $S_N^i \to 0$ in probability as $N \to \infty$, which will imply that $(T_N^{i+1} - \tilde{T}_N^i)/\beta_N \Rightarrow$ exponential$(\tilde{r}_{01})$.

We first note that

$$\begin{aligned}\mathbf{E}\big[S_N^i\big] &= \mathbf{E}\big[T_N^{i+1} - \tilde{\sigma}_N^{i,s(i)}\big] = \sum_j \frac{r_{0j}}{\sum_{j'} r_{0j'}} e_{jN} \\ &= \frac{1}{\sum_{j'}(r_{0j'}/\omega_N)} \frac{1}{\beta_N \omega_N} \sum_j r_{0j} e_{jN} \cdot \beta_N;\end{aligned}$$

therefore, $\mathbf{E}[S_N^i/\beta_N] \to 0$, since the first fraction converges to $1/\tilde{r}_+$, and the second to 0, by (30) and (32), respectively. Similarly, for each unsuccessful excursion $1 \le i' < s(i)$

$$\mathbf{E}\big[\tilde{\tau}_N^{i,i'} - \tilde{\sigma}_N^{i,i'}\big] = \frac{1}{\omega_N} \frac{1}{\sum_{j'}(r_{0j'}/\omega_N)} \sum_j r_{0j} e_{j0},$$

and since $s(i)$ is geometric,

$$\mathbf{E}\big[s(i)\big] = \frac{\sum_j r_{0j}}{\sum_j r_{0j}\pi_{jN}} = \omega_N \frac{\sum_j r_{0j}/\omega_N}{\sum_j r_{0j}\pi_{jN}}.$$

We have

$$\mathbf{E}\big[U_N^i\big] = \mathbf{E}\Big[\sum_{i' < s(i)} \big(\tilde{\tau}_N^{i,i'} - \tilde{\sigma}_N^{i,i'}\big)\Big] \le \mathbf{E}[s(i)]\mathbf{E}\big[\tilde{\tau}_N^{i,i'} - \tilde{\sigma}_N^{i,i'}\big] = \frac{\sum_j r_{0j} e_{j0}}{\beta_N \sum_j r_{0j}\pi_{jN}} \cdot \beta_N$$

and so $\mathbf{E}[U_N^i/\beta_N] \to 0$, since by (31) the denominator converges to $\tilde{r}_{01}$, and by (32) the numerator goes to 0. Hence for any $\delta > 0$ we have $\mathbf{P}[S_N^i > \delta] \le \frac{\mathbf{E}[S_N^i]}{\delta} \to 0$ and $\mathbf{P}[U_N^i > \delta] \le \frac{\mathbf{E}[U_N^i]}{\delta} \to 0$.

A completely analogous proof shows that $(\tilde{T}_N^i - T_N^i)\beta_N \Rightarrow$ exponential$(\tilde{r}_{10})$, and the claim about the probability measure ν_t is immediate from the fact that $\mathbf{E}[U_N^i + S_N^i] \to 0$. $\square$

To verify condition (30) one only needs to use the rates of biased reactions on the boundary. For (31), note the fact that if not for biased reactions, the process would be a martingale; if the rates of the biased reactions are overpowered by those of the balanced reactions and splitting [as quantified in (31)], then the process is approximately a martingale. Conditions in (32) predominantly depend on how fast the rates of the balanced reactions and splitting are, as they determine the length of excursions of the process.

These conditions are the easiest to verify when the reactions as well as splitting/resampling mechanism make only unit net changes at each step, so that X_A is a birth–death process with $r_{ij} = 0$ if $|i - j| > 1$. In this case one can specify more precise conditions on the rates r_{ij} that will ensure that (30)–(32) hold. We consider the case when the time is already rescaled, that is, $\beta_N = 1$, and the rate of reactions on the boundaries is $\omega_N = N$. We use the following notation for birth and death rates:

$$r_+(i) := r_{i(i+1)}, \qquad r_-(i) := r_{i(i-1)}, \qquad \varepsilon_N(i) = \frac{r_-(i)}{r_+(i)} - 1,$$

with $\varepsilon_N(i)$ quantifying the strength of the bias at state i [we stress its dependence on N via transition rates $r_\pm(i)$].

PROPOSITION 4.2. *If X_A is a birth–death chain whose rates satisfy*

$$\frac{r_+(0)}{N} \to \tilde{r}_+ \in (0, \infty), \qquad \frac{r_-(N)}{N} \to \tilde{r}_- \in (0, \infty), \tag{33}$$

$$\sum_{i=1}^{N-1} |\varepsilon_N(i)| \to 0 \quad \textit{and} \tag{34}$$

$$\sum_{i=1}^{N-1} \frac{N-i}{r_+(i)} \to 0, \qquad \sum_{i=1}^{N-1} \frac{i}{r_-(i)} \to 0, \tag{35}$$

then conditions (30)–(32) *hold with $\omega_N = N$, $\beta_N = 1$ and $\tilde{r}_{01} = \tilde{r}_+$, $\tilde{r}_{10} = \tilde{r}_-$.*

Analogous to the general case, (33) depends only on the rates of biased reactions on the boundaries, (34) reflects the fact that off of the boundaries the drift of the biased reactions is much weaker than the noise of the balanced reactions and splitting and (35) is an estimate on the speed of the balanced reactions and splitting.

PROOF OF PROPOSITION 4.2. (30) is immediate from (33) and $\omega_N = N$. To verify (31) we solve for π_{jN}, $j \in \{1, \dots, N-1\}$.

LEMMA 4.1. *If* (34) *holds, then $N\pi_{1N} \to 1$ and $N\pi_{(N-1)0} \to 1$.*

PROOF. Let φ be such that $\varphi(X_A)$ is a martingale, that is, let $\varphi(x) = \mathbf{E}[\varphi(X_A(\tau_{0,N}))|X(0) = x]$ for $x \in \{1, \dots, N-1\}$ and $\varphi(0) = 0, \varphi(1) = 1$. Standard result for birth–death processes, using a recursive equation for $\psi(x) = \varphi(x) - \varphi(x-1)$, gives

$$\varphi(x) = \sum_{i=1}^{x} \psi(i) = \sum_{i=1}^{x} \prod_{j=1}^{i-1} \frac{r_-(j)}{r_+(j)}.$$

By the optional stopping theorem for the stopping time $\tau_{0,N}$,

$$\varphi(i) = \mathbf{E}\big[\varphi\big(X_A(\tau_{0,N})\big)|X(0) = i\big] = \pi_{i0}\varphi(0) + \pi_{iN}\varphi(N),$$

so $\pi_{iN} = (\varphi(i) - \varphi(0))/(\varphi(N) - \varphi(0)) = \varphi(i)/\varphi(N)$, and

$$\pi_{1N} = \frac{1}{\varphi(N)} = \left(\sum_{i=1}^{N} \prod_{j=1}^{i-1} \frac{r_-(j)}{r_+(j)}\right)^{-1} = \left(\sum_{i=1}^{N} \prod_{j=1}^{i-1} (1 + \varepsilon_N(j))\right)^{-1} = \frac{1}{Nc(N)},$$

where $c(N) = \frac{1}{N} \sum_{i=1}^{N} \prod_{j=1}^{i-1} (1 + \varepsilon_N(j))$.

Condition (34) implies that $\sup_{1 \le j \le N-1} \{|\varepsilon_N(j)|\} \to 0$, so let N_0 be such that $\forall N > N_0$ and $\forall j \in \{1, \dots, N-1\}$, $|\varepsilon_N(j)| < 1/3$. Since $\forall x \in [0, 1/3)$, $1 - x \ge e^{-x-x^2}$, and $\forall x \in \mathbb{R}$, $1 + x \le e^x$, we have that uniformly for all $1 \le a, b \le N-1$, where $N > N_0$

$$\prod_{j=a}^{b} (1 + \varepsilon_N(j)) \le \prod_{j=1}^{N-1} (1 + |\varepsilon_N(j)|) \le \prod_{j=1}^{N-1} e^{|\varepsilon_N(j)|} = \exp\left(\sum_{j=1}^{N-1} |\varepsilon_N(j)|\right) \tag{36}$$

and

$$\begin{aligned} \prod_{j=a}^{b} (1 + \varepsilon_N(j)) &\ge \prod_{j=a}^{b} (1 - |\varepsilon_N(j)|) \ge \prod_{j=1}^{N-1} (1 - |\varepsilon_N(j)|) \\ &\ge \prod_{j=1}^{N-1} \exp\big(-|\varepsilon_N(j)| - |\varepsilon_N(j)|^2\big) \\ &= \exp\left(-\sum_{j=1}^{N-1} \big(|\varepsilon_N(j)| + |\varepsilon_N(j)|^2\big)\right) \end{aligned} \tag{37}$$

hence $N\pi_{1N} = 1/c(N) \to 1$.

To get $N\pi_{(N-1)0} \to 1$, if we flip the state space by letting $\check{\imath} = N - i$, then the new boundaries are $\check{0} = N$ and $\check{N} = 0$, and we get a birth–death process $\check{X}_A$ whose rates are precisely the flip of those for X_A. That is, the rates of $\check{X}_A$ are $\check{r}_+(\check{\imath}) = r_-(N - i)$, $\check{r}_-(\check{\imath}) = r_+(N - i)$, and their ratio is

$$1 + \check{\varepsilon}_N(\check{\imath}) = \frac{\check{r}_-(\check{\imath})}{\check{r}_+(\check{\imath})} = \frac{r_+(N-i)}{r_-(N-i)},$$

giving the same product of ratios as for the original process.

$$\prod_{\check{\imath}=1}^{N-1}\bigl(1+\check{\varepsilon}_N(\check{\imath})\bigr)=\prod_{j=1}^{N-1}\bigl(1+\varepsilon_N(j)\bigr).$$

Hence, the exact argument above now applied to $\check{X}_A$ gives $N\check{\pi}_{\check{1}\check{N}}=N\pi_{(N-1)0}\to 1$ as well. □

Once we have the result of Lemma 4.1, it is immediate that $N\pi_{1N}\to 1$, $N\pi_{(N-1)0}$ imply (31) with $\tilde{r}_{01}=\tilde{r}_+$, $\tilde{r}_{10}=\tilde{r}_-$.

To verify (32) we next solve for e_j, $j\in\{1,\dots,N-1\}$, where $e_j=\mathbf{E}[\tau_{0,N}\mid X_A(0)=j]$ for $j\in\{1,\dots,N-1\}$, and $e_0=e_N=0$.

LEMMA 4.2. *If* (34) *and* (35) *hold, then* $Ne_1\to 0$ *and* $Ne_{N-1}\to 0$.

PROOF. The expected time of an excursion satisfies the recursion

$$e_i=\frac{1}{r_-(i)+r_+(i)}+\frac{r_-(i)}{r_-(i)+r_+(i)}e_{i-1}+\frac{r_+(i)}{r_-(i)+r_+(i)}e_{i+1},$$

which gives

$$r_+(i)(e_{i+1}-e_i)-r_-(i)(e_i-e_{i-1})=-1;$$

letting $f(i)=e_i-e_{i-1}$ gives the recursive equation

$$f(i+1)=-\frac{1}{r_+(i)}+\frac{r_-(i)}{r_+(i)}f(i)=-\frac{1}{r_+(i)}+\bigl(1+\varepsilon_N(i)\bigr)f(i).$$

Note that except for the $-\frac{1}{r_+(i)}$ term, this is reminiscent of the recursion for $\psi(i)=\frac{r_-(i)}{r_+(i)}\psi(i-1)$. Hence

$$f(k)=f(1)\prod_{j=1}^{k-1}\bigl(1+\varepsilon_N(j)\bigr)-\sum_{i=1}^{k-1}\frac{1}{r_+(i)}\prod_{j=i+1}^{k-1}\bigl(1+\varepsilon_N(j)\bigr).$$

To find $f(1)=e_1-e_0=e_1$ we impose the condition $\sum_{i=1}^N f(i)=e_N-e_0=0$ and get

$$e_1=\left(\sum_{k=1}^{N}\sum_{i=1}^{k-1}\frac{1}{r_+(i)}\prod_{j=i+1}^{k-1}\bigl(1+\varepsilon_N(j)\bigr)\right)\Bigg/\left(\sum_{k=1}^{N}\prod_{j=1}^{k-1}\bigl(1+\varepsilon_N(j)\bigr)\right).$$

Let $\eta_N=\sup_{1\le a,b\le N-1}|\prod_a^b(1+\varepsilon_N(j))-1|$. Then (34) implies $\eta_N\to 0$ for $N>N_0$ via (36) and (37). We have

$$\begin{aligned}e_1&\le\left(\sum_{k=1}^{N}\sum_{i=1}^{k-1}\frac{1}{r_+(i)}(1+\eta_N)\right)\Bigg/\left(\sum_{k=1}^{N}(1-\eta_N)\right)\\&=\frac{1+\eta_N}{(1-\eta_N)N}\sum_{k=1}^{N}\sum_{i=1}^{k-1}\frac{1}{r_+(i)}=\frac{1+\eta_N}{(1-\eta_N)N}\sum_{i=1}^{N-1}\frac{N-i}{r_+(i)},\end{aligned}$$

and thus (35) implies $Ne_1 \to 0$.

To obtain $Ne_{N-1} \to 0$ we can flip the process and consider $\check{X}_A$ with the flipped rates as in the proof of Lemma 4.1. Now, in addition to (34), we also require the flip version of the first condition in (35),

$$\sum_{\check{\imath}=1}^{N-1} \frac{\check{N}-\check{\imath}}{\check{r}_+(\check{\imath})} = \sum_{i=1}^{N-1} \frac{N-i}{r_-(N-i)} = \sum_{i=1}^{N-1} \frac{i}{r_-(i)} \to 0,$$

which are guaranteed by the second condition in (35). □

Once we have the results of both Lemmas 4.1 and 4.2, we can deduce that $Ne_1 \to 0$ and $Ne_{N-1} \to 0$, which imply (32). Namely, from $e_1 = e_{1N}\pi_{1N} + e_{10}\pi_{10}$,

$$e_{1N} \le \frac{e_1}{\pi_{1N}} = \frac{Ne_1}{N\pi_{1N}} \to 0$$

since Lemma 4.1 ensures convergence of the denominator to 1 and Lemma 4.2 of the numerator to 0. Similarly

$$Ne_{10} \le \frac{Ne_1}{\pi_{10}} \le \frac{Ne_1}{r_-(0)/(r_+(0)+r_-(0))} \to 0$$

since π_{10} contains the positive probability (independent of N) of an immediate return to 0. □

For a reaction system and splitting with unit net changes only, since splitting is unbiased we have $p_{i,i+1} = p_{i,i-1} = \frac{1}{2}$, for $i \neq 0, N$, and the contribution to $r_+(i)$ and $r_-(i)$ from splitting is $\frac{1}{2}\gamma(i,N)$. Let us write $\gamma(i,N) = \gamma(N)p_i$ where $\gamma(N)$ depends on N only (i.e., is state independent) and $p_i = O(1)$. Then, in any state, the contribution of the splitting is of $O(\gamma(N))$, while the contribution of the reaction system is of $O(N)$ due to the standard scaling of reaction rates. Hence, we have the following result.

THEOREM 4.1. *If the reaction system has increments of size* $\{1,-1\}$ *only; contains reactions* $aA \to (a-1)A + B$, $bB \to A + (b-1)B$ *some* $a, b > 0$; *has rates with standard scaling* $\kappa_\zeta^{ab}(N) = \tilde{\kappa}_\zeta^{ab} N^{1-(a+b)}$; *and if the splitting mechanism has increments of size* $\{1,-1\}$, $p_{0,0} = p_{N,N} = 1$, *with rate is* $\gamma(i,N) = \gamma(N)p_i$ *where* $\gamma(N)$ *and* $p_i = O(1)$ *satisfy*

$$(38)\quad \frac{N}{\gamma(N)} \sum_{i=1}^{N-1} \frac{1}{p_i} \to 0, \qquad \frac{1}{\gamma(N)} \sum_{i=1}^{N-1} \frac{i}{p_i} \to 0, \qquad \frac{1}{\gamma(N)} \sum_{i=1}^{N-1} \frac{N-i}{p_i} \to 0;$$

then the results of Proposition 4.2 apply with $\beta_N = 1$ *and* $\tilde{r}_{01} = \sum_{(0,b,1)\in\mathcal{I}} \tilde{\kappa}_1^{0b}$, $\tilde{r}_{10} = \sum_{(a,0,-1)\in\mathcal{I}} \tilde{\kappa}_{-1}^{a0}$.

PROOF. The transition rates for X_A are given by

$$r_+(i) = \frac{1}{2}\gamma(N)p_i + N\sum_{(a,b,1)\in\mathcal{I}} \tilde{\kappa}_1^{ab}(i/N)_{a,N}(1-i/N)_{b,N}, \qquad i = 0,\dots,N-1,$$

$$r_-(i) = \frac{1}{2}\gamma(N)p_i + N\sum_{(a,b,-1)\in\mathcal{I}} \tilde{\kappa}_{-1}^{ab}(i/N)_{a,N}(1-i/N)_{b,N}, \qquad i = 1,\dots,N.$$

On the boundary the rates are

$$r_+(0) = N\sum_{(0,b,1)\in\mathcal{I}} \tilde{\kappa}_1^{0b}, \qquad r_-(N) = N\sum_{(a,0,-1)\in\mathcal{I}} \tilde{\kappa}_{-1}^{a0},$$

and (33) holds with $\tilde{r}_+ = \sum_{(0,b,1)\in\mathcal{I}} \tilde{\kappa}_1^{0b}$, $\tilde{r}_- = \sum_{(a,0,-1)\in\mathcal{I}} \tilde{\kappa}_{-1}^{a0}$. Also,

$$\begin{aligned}
\varepsilon_N(i) &= \frac{(1/2)\gamma(N)p_i + N\sum_{(a,b,-1)\in\mathcal{I}} \tilde{\kappa}_{-1}^{ab}(i/N)_{a,N}(1-i/N)_{b,N}}{(1/2)\gamma(N)p_i + N\sum_{(a,b,1)\in\mathcal{I}} \tilde{\kappa}_1^{ab}(i/N)_{a,N}(1-i/N)_{b,N}} - 1 \\
&= \Bigg(N\Bigg(\sum_{(a,b,1)\in\mathcal{I}} \tilde{\kappa}_{-1}^{ab}(i/N)_{a,N}(1-i/N)_{b,N} \\
&\qquad - \sum_{(a,b,1)\in\mathcal{I}} \tilde{\kappa}_1^{ab}(i/N)_{a,N}(1-i/N)_{b,N}\Bigg)\Bigg) \\
&\qquad \Bigg/\Bigg((1/2)\gamma(N)p_i + N\sum_{(a,b,1)\in\mathcal{I}} \tilde{\kappa}_1^{ab}(i/N)_{a,N}(1-i/N)_{b,N}\Bigg) \\
&\le \frac{2NR\max_{(a,b,\zeta)\in\mathcal{I}} \tilde{\kappa}_\zeta^{ab}}{\gamma(N)p_i}
\end{aligned}$$

since $\tilde{\kappa}_\zeta^{ab} \ge 0$, where $R < \infty$ is the number of reactions in the system. Therefore

$$\sum_{i=1}^{N-1} |\varepsilon_N(i)| \le 2R \max_{(a,b,\zeta)\in\mathcal{I}} \{\tilde{\kappa}_\zeta^{ab}\} \frac{N}{\gamma(N)} \sum_{i=1}^{N-1} \frac{1}{p_i},$$

and the first condition in (38) ensures that $\sum_{i=1}^{N-1} |\varepsilon_N(i)| \to 0$ and (34) holds. On the other hand,

$$\begin{aligned}
\sum_{i=1}^{N-1} \frac{i}{r_-(i)} &= \sum_{i=1}^{N-1} \frac{i}{(1/2)\gamma(N)p_i + \sum_{(a,b,-1)\in\mathcal{I}} \tilde{\kappa}_{-1}^{ab}(i/N)^{a-1}(1-i/N)^b} \\
&\le \frac{2}{\gamma(N)} \sum_{i=1}^{N-1} \frac{i}{p_i}
\end{aligned}$$

and

$$\sum_{i=1}^{N-1} \frac{N-i}{r_+(i)} \leq \frac{2}{\gamma(N)} \sum_{i=1}^{N-1} \frac{N-i}{p_i}$$

so the last two conditions in (38) ensure that (35) is satisfied as well. □

4.2. *Example*: *Bistable behavior from fast splitting*. We revisit the same example of the reaction system we analyzed in Section 3.3:

$$A \xrightarrow{\kappa_{-1}^{10}} B, \tag{39}$$

$$B \xrightarrow{\kappa_{1}^{01}} A, \tag{40}$$

$$A + B \xrightarrow{\kappa_{-1}^{11}} 2B, \tag{41}$$

$$2A + B \xrightarrow{\kappa_{1}^{21}} 3A \tag{42}$$

with the standard mass-action scaling, $\kappa_\zeta^{ab} = N^{-(a+b)+1}\tilde{\kappa}_\zeta^{ab}$. In this system the only reactions which counteract the absorption on the boundaries are the first two unimolecular reactions. Also, note that all system reactions change the molecular count of A only by increments of size 1.

We chose the same simple splitting mechanism as before, since conditions (34), (35) and (38) are much easier to verify than conditions (31), (32). Recall that, if we were to assume $\gamma(N) = \frac{1}{2}\varepsilon^2 N^2$ for some small $\varepsilon^2 > 0$, then the limiting process for X_N would be the diffusion process $\tilde{X}_\varepsilon$ in (28); the splitting noise is even less present if we were to assume $\gamma(N) = \frac{1}{2}N$, as shown in Section 3.3. In contrast, if we assume the rate $\gamma(N)$ grows fast enough so that $N^2 \ln N/\gamma(N) \to 0$, then we can show that the conditions in Proposition 4.2 are satisfied, and the behavior of the limiting process for X_N is described by a different two-state jump Markov process.

There are only two reactions in (39)–(42) active on the boundaries, so $\tilde{r}_{01} = \tilde{\kappa}_1^{01}$ and $\tilde{r}_{10} = \tilde{\kappa}_{-1}^{10}$. To verify (38), note that we have $\gamma(i, N) = \gamma(N)p_i$ with $p_i = \frac{i}{N}(1 - \frac{i}{N}) = i(N-i)/N^2$, so

$$\frac{N}{\gamma(N)} \sum_{i=1}^{N-1} \frac{1}{p_i} = \frac{N^3}{\gamma(N)} \sum_{i=1}^{N-1} \frac{1}{i(N-i)} = \frac{N^2}{\gamma(N)} \sum_{i=1}^{N-1} \left(\frac{1}{i} + \frac{1}{N-i} \right) = \frac{2N^2 h_N}{\gamma(N)}$$

using partial fractions $\frac{1}{i(N-i)} = \frac{1}{N}(\frac{1}{i} + \frac{1}{(N-i)})$, where h_N is the Nth harmonic sum. Also

$$\frac{1}{\gamma(N)} \sum_{i=1}^{N-1} \frac{i}{p_i} = \frac{N^2}{\gamma(N)} \sum_{i=1}^{N-1} \frac{1}{N-i} = \frac{N^2 h_N}{\gamma(N)}$$

and

$$\frac{1}{\gamma(N)}\sum_{i=1}^{N-1}\frac{N-i}{p_i}=\frac{1}{\gamma(N)}\sum_{i=1}^{N-1}\frac{1}{i}=\frac{N^2h_N}{\gamma(N)}$$

as well. Hence, $N^2\ln N/\gamma(N)\to 0$ ensures that all conditions in (38) hold.

This example shows that for any reaction system with unit increments whose drift has a double well potential, and for this particular choice of the splitting mechanism, we can identify orders of magnitude for $\gamma(N)$ that lead to different limiting behaviors:

• If $\gamma(N)\ll N$, bistability is caused by large deviations of the Markov jump process, and the rescaled process transitions between neighborhoods of the drift equilibirum points on a time-scale of order $e^{N(\gamma(N))^{-1}\iota_{x_i,x_2}}$, with $N(\gamma(N))^{-1}\to\infty$.

• If $\gamma(N)\sim\varepsilon^2N^2$, $\varepsilon^2>0$ a constant, bistability is caused by large deviations of a diffusion with a small perturbation coefficient, with transitions between neighborhoods of the drift equilibirum points on a time-scale of order $e^{\varepsilon^{-2}I_{x_i,x_2}}$.

• If $\gamma(N)\gg N^2\ln N$, bistability is caused by excessive noise, and switching between the boundary points occurs on a time-scale of order 1.

Note that the order of magnitude N^2 only represents the scale on which we have assumed that the variance of the splitting mechanism is in the diffusive case [see assumption (7*) in Section 2.2]. Also note that existence of two stable states in the deterministic model for the reaction system is not needed for the result of this section. We chose the same reaction system in order to make the comparison with the results in Section 3 and emphasize the difference between the effects of "slow" and "fast" splitting on the same reaction system.

Figure 3 shows an exact simulation of a sample path of the rescaled process $X_N=N^{-1}X_A$ for a relatively short period of time, spending most of its

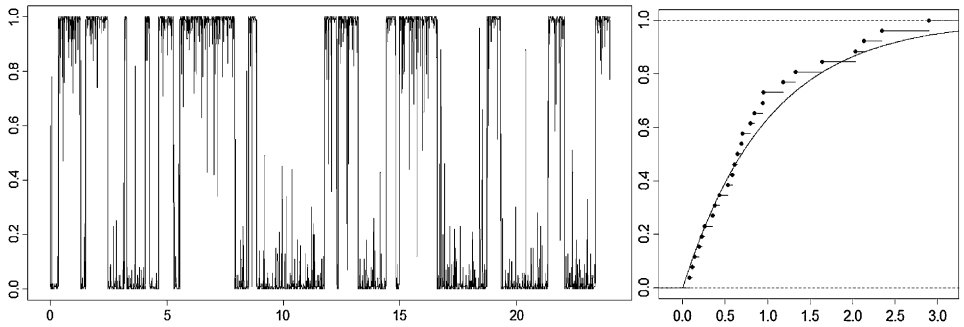

FIG. 3. *Sample path $X_N(t)$ (left: x-axis $=t$, y-axis $=X_N(t)=N^{-1}X(t)\subset[0,1]$) of the system (22)–(25) with (27) splitting, under standard mass-action scaling for reactions and $\gamma(N)=\frac{1}{2}N^3$ (parameter $N=200$); and the distribution of switching times plotted (dots) in terms of quantiles (right: x-axis $=t$, y-axis $=$ fraction of switching times of length $\le t$). Solid line $(1-e^{-t})$ indicates the quantiles of the exponential (mean 1) distribution for comparison.*

time at boundaries $\{0\} \cup \{1\}$, switching between them at approximately rates $\tilde{r}_{01} = \tilde{\kappa}_1^{01} = 1, \tilde{r}_{10} = \tilde{\kappa}_{-1}^{10} = 1$; see Section 3.3 for coefficient values. Switching between states occurs at a time-scale $\beta_N = 1$, and since $\tilde{r}_{01} = \tilde{\kappa}_1^{01} = 1, \tilde{r}_{10} = \tilde{\kappa}_{-1}^{10} = 1$ the distribution of switching times should approximately be an exponential distribution (mean 1) distribution. This is shown in the quantile plot in Figure 3, where the fraction of switching times of length $\leq t$ is plotted against the same fraction $1 - e^{-t}$ for the exponential (mean 1) distribution.

5. Discussion. We showed that there are two different types of stochastic bistable behavior in which the system spends most of its time at or near one of two states and switches between them. For one of these types of bistability, because the magnitude of noise is high, it can occur even in a system whose deterministic model would not allow for a possibility of bistability at all. The detreministic system can have unique stable points, as, for example, in the neutral Wright–Fisher model with mutation. For the other type of bistability, where the noise is relatively low, one needs the reaction system to have two deterministic stable points, as, for example, in the Schlögl model. The important point is what constitutes "high" and "low" levels of noise: the determining quantity $\varepsilon_A(N)$ (11) depends on the relative size in terms of N of the variance to the average change in the system, where N is a scaling parameter for the size of the system. We referred to $\varepsilon_A(N) \approx 0$ as "slow" splitting, and to $\varepsilon_A(N) \approx \infty$ as "fast" splitting, interpreted relative to the reaction dynamics.

We discussed the differences in the qualitative signatures of bistability in the two cases:

- In case of "slow" splitting, the states where the process spends most of its time are determined by the drift of the deterministic model for the reaction system; in contrast, in case of "fast" splitting, they are simply the two extremes for the size of the system.
- In case of "slow" splitting, the rate of switching is determined by the relative magnitude of the splitting variance to the reaction drift and by the size of the potential barrier in the deterministic model for the system; on the other hand, in case of "fast" splitting, the rates of switching are determined only by the standardized rates of the reactions that are realizable from one of the extremes for the system size.
- In case of "slow" splitting, the time-scale β_ε or β_{ε_N} on which the switching happens is exponential in (some increasing function of) the size of the system; in contrast, in case of "fast" splitting, the time-scale β_N is at most polynomial.

We also showed that the observables of bistability (switching states and rates) are not sensitive to precise specification of the reaction system, as they depend only on: equilibrium points, size of potential barrier in "slow" splitting, and drift values at boundaries in "fast" splitting. However, bistability is very sensitive to

the distributional form of the splitting/resampling mechanism: the variance of its distribution determines the potential barrier in "slow" splitting, and the harmonic sum of its transition probabilities determines the threshold for appearance of "fast" splitting.

In the context of cellular systems of biochemical reactions, the problem of determining the partitioning errors due to cell division is experimentally extremely challenging (Huh and Paulsson [14, 15]). The measurements for single cells rely on count estimates for related species rather than the molecular species of interest. In addition, in order to estimate the magnitude of intracellular noise, one has to separate the intrinsic from the extrinsic sources of randomness. How random is cell division, and how it compares in magnitude to the biochemical noise is a question that is very much open. However, since our analysis only depends on a few general features of the splitting mechanism (unbiasedness and time-scale of the rate), it is also possible that stochastic bistability is achieved by a set of auxiliary reactions, instead of splitting, acting on a different time-scale from the rest of the system. For example, the protein bursting mechanism may act as the driver of stochastic bistability (Zong et al. [29], Kaufman et al. [17]).

One can try to rely on the qualitative signatures of bistability in order to assess which of the two types of bistability we discussed is relevant in a specific cellular biochemical systems. When the switching times are orders of magnitude greater than the molecular count of the switching species, as in the lysogenic switch of *E. coli*, the "slow" splitting may be the more likely mechanism. This evaluation is sensitive to the choice of time units, given which both the splitting and reaction rates should be of reasonable orders of magnitude in terms of the molecular count. It is natural to chose units of time corresponding to cell-doubling or cell-division time (the splitting rate is then of order 1—and the range of splitting rates in our model, in any of the different cases, is at most linear). In an experimental analysis of this system, Zong et al. [29] observed that the switching times of the cell are exponential in the number of protein burst events, and correspond to a calculation of the rare event probability of the bursts, as can be interpreted by large deviations in our "low" auxiliary noise ("slow" splitting) type of bistability. In contrast, when the switching times are relatively short, as in the gene expression switch in *S. cerevisiae*, the "fast" splitting is the probable mechanism. In the engineered chemical reaction network version of this system, Kaufmann et al. [17] show that increasing the protein burst size (increasing the auxiliary noise) leads to more highly correlated switching behavior in different cell lineages, as could be inferred from properties of our "high" auxiliary noise ("fast" splitting) bistability type.

Acknowledgment. The authors would like to thank Jonathan Mattingly, whose suggestion during the BIRS workshop on "Multi-scale Stochastic Modeling of Cell Dynamics" began this investigation.

REFERENCES

[1] ACAR, M., METTETAL, J. T. and VAN OUDENAARDEN, A. (2008). Stochastic switching as a survival strategy in fluctuating environments. *Nat. Genet.* **40** 471–475.

[2] AVERY, S. V. (2006). Microbial cell individuality and the underlying sources of heterogeneity. *Nat. Rev. Microbiol.* **4** 577–587.

[3] BALL, K., KURTZ, T. G., POPOVIC, L. and REMPALA, G. (2006). Asymptotic analysis of multiscale approximations to reaction networks. *Ann. Appl. Probab.* **16** 1925–1961. MR2288709

[4] BISHOP, L. M. and QIAN, H. (2010). Stochastic bistability and bifurcation in a mesoscopic signaling system with autocatalytic kinase. *Biophysical Journal* **98** 111.

[5] BREHM-STECHER, B. F. and JOHNSON, E. A. (2004). Single-cell microbiology: Tools, technologies, and applications. *Microbiol. Mol. Biol. Rev.* **68** 538–559.

[6] CHVÁTAL, V. (1979). The tail of the hypergeometric distribution. *Discrete Math.* **25** 285–287. MR0534946

[7] DEMBO, A. and ZEITOUNI, O. (1998). *Large Deviations Techniques and Applications*, 2nd ed. *Applications of Mathematics* (*New York*) **38**. Springer, New York. MR1619036

[8] DURRETT, R. (1996). *Stochastic Calculus*: *A Practical Introduction*. CRC Press, Boca Raton, FL. MR1398879

[9] DURRETT, R. (2008). *Probability Models for DNA Sequence Evolution*, 2nd ed. Springer, New York. MR2439767

[10] ELOWITZ, M. B., LEVINE, A. J., SIGGIA, E. D. and SWAIN, P. S. (2002). Stochastic gene expression in a single cell. *Science Signalling* **297** 5584.

[11] ETHIER, S. N. and KURTZ, T. G. (1986). *Markov Processes*: *Characterization and Convergence*. Wiley, New York. MR0838085

[12] FREIDLIN, M. I. and WENTZELL, A. D. (1998). *Random Perturbations of Dynamical Systems*, 2nd ed. *Grundlehren der Mathematischen Wissenschaften* **260**. Springer, New York. MR1652127

[13] GALVES, A., OLIVIERI, E. and VARES, M. E. (1987). Metastability for a class of dynamical systems subject to small random perturbations. *Ann. Probab.* **15** 1288–1305. MR0905332

[14] HUH, D. and PAULSSON, J. (2011). Non-genetic heterogeneity from stochastic partitioning at cell division. *Nat. Genet.* **43** 95–100.

[15] HUH, D. and PAULSSON, J. (2011). Random partitioning of molecules at cell division. *Proc. Natl. Acad. Sci. USA* **108** 15004–15009.

[16] KARLIN, S. and TAYLOR, H. M. (1981). *A Second Course in Stochastic Processes*. Academic Press, New York. MR0611513

[17] KAUFMANN, B. B., YANG, Q., METTETAL, J. T. and VAN OUDENAARDEN, A. (2007). Heritable stochastic switching revealed by single cell genealogy. *PLOS Biology* **5** 1973–1980.

[18] KURTZ, T. G. (1970). Solutions of ordinary differential equations as limits of pure jump Markov processes. *J. Appl. Probab.* **7** 49–58.

[19] KURTZ, T. G. (1971). Limit theorems for sequences of jump Markov processes approximating ordinary differential processes. *J. Appl. Probab.* **8** 344–356. MR0287609

[20] KURTZ, T. G. (1977/78). Strong approximation theorems for density dependent Markov chains. *Stochastic Process. Appl.* **6** 223–240. MR0464414

[21] KURTZ, T. G. (1981). *Approximation of Population Processes*. *CBMS-NSF Regional Conference Series in Applied Mathematics* **36**. SIAM, Philadelphia, PA. MR0610982

[22] MCADAMS, H. H. and ARKIN, A. (1999). It's a noisy business! Genetic regulation at the nanomolar scale. *Trends Genet.* **15** 65–69.

[23] OLIVIERI, E. and VARES, M. E. (2005). *Large Deviations and Metastability*. *Encyclopedia of Mathematics and Its Applications* **100**. Cambridge Univ. Press, Cambridge. MR2123364

[24] ROBERT, L., PAUL, G., CHEN, Y., TADDEI, F., BAIGL, D. and LINDNER, A. B. (2010). Pre-dispositions and epigenetic inheritance in the Escherichia coli lactose operon bistable switch. *Molecular Systems Biology* **6** 1.

[25] SAMOILOV, M., PLYASUNOV, S. and ARKIN, A. (2005). Stochastic amplification and signaling in enzymatic futile cycles through noise-induced bistability with oscillations. *Proc. Natl. Acad. Sci. USA* **102** 2310–2315.

[26] SCHLÖGL, F. (1972). Chemical reaction models for non-equilibrium phase transitions. *Zeitschrift Für Physik A* **253** 147–161.

[27] SHWARTZ, A. and WEISS, A. (1995). *Large Deviations for Performance Analysis*: *Queues*, *Communications*, *and Computing*. Chapman & Hall, London. MR1335456

[28] VELELLA, M. and QIAN, H. (2009). Stochastic dynamical and non-equilibrium thermodynamics of a bistable chemical system: The Schlögl model revisited. *Journal of the Royal Society Interface* **6** 925–940.

[29] ZONG, C., SO, L. H., SEPÚLVEDA, L. A., SKINNER, S. O. and GOLDING, I. (2010). Lysogen stability is determined by the frequency of activity bursts from the fate-determining gene. *Mol. Syst. Biol.* **6** 440.

DEPARTMENT OF MATHEMATICS
ROSE-HULMAN INSTITUTE OF TECHNOLOGY
5500 WABASH AVE.
TERRE HAUTE, INDIANA 47803
USA
E-MAIL: mcsweene@rose-hulman.edu

DEPARTMENT OF MATHEMATICS AND STATISTICS
CONCORDIA UNIVERSITY
1455 DE MAISONNEUVE BLVD. WEST
MONTREAL, QC H3G 1M8
CANADA
E-MAIL: lpopovic@mathstat.concordia.ca