**evosim**: fast and scalable stochastic simulations of evolutionary dynamics

Dalit Engelhardt*1,2,3,4 and Thomas O. McDonald1,2,3,4

1 Department of Data Science, Dana-Farber Cancer Institute, Boston, MA 02215, USA
2 Center for Cancer Evolution, Dana-Farber Cancer Institute, Boston, MA 02215, USA
3 Department of Biostatistics, Harvard T. H. Chan School of Public Health, Boston, MA 02115, USA
4 Department of Stem Cell and Regenerative Biology, Harvard University, Cambridge MA 02138, USA

Abstract

The simulation of clonal dynamics with branching processes can provide valuable insights into disease progression and treatment optimization, but exact simulation of branching processes via the Stochastic Simulation Algorithm (SSA) is computationally prohibitive at the large population sizes associated with therapeutically-relevant scenarios. **evosim** is a versatile and flexible Python implementation of a fast and unbiased tau-leaping algorithm for the simulation of birth-death-mutation branching processes that is scalable to any population size. Package functionalities support the incorporation and tracking of a sequence of evolutionary changes such as therapeutic interventions as well as the analysis of population diversity. We show that runtimes scale logarithmically with population size, by contrast to the linear scaling of the SSA, and simulations exhibit strong agreement with SSA simulation results. These findings are also supported by mathematical results (Supplementary information).

Availability: Package, documentation, and tutorials / usage examples are available on GitHub (https://github.com/daliten/evosim). Mathematical details of the algorithm and the pseudocode are provided in the included Supplementary information.

1 Introduction

The dynamical behavior of cell populations is controlled by inherently stochastic processes that determine a wide range of biological outcomes such as gene regulation, evolutionary progression, cellular differentiation, and drug response. Among these processes, births, deaths, and mutations all involve variable degrees of biological randomness that can lead to diverse outcomes under different environmental pressures. The ability to obtain accurate predictions of the interplay between these processes and conditions to which a population of cells is subjected is key to developing better therapeutic interventions, predicting the effects of changing environments on ecosystem dynamics, and understanding the progression of diseases driven by genomic alterations and cell proliferation.

Birth-death-mutation processes, in which a given cell can undergo a division into identical progeny cells, undergo a division resulting in (a) mutant progeny cell(s), or die, are commonly modeled by stochastic branching process models. Gillespie’s Stochastic Simulation Algorithm (SSA) [1, 2] provides an exact simulation of these processes; however, the SSA becomes computationally expensive at large population sizes due to its linear scaling with system size, making the simulation of populations greater than $10^6$ cells typically intractable. This poses significant challenges for biomedical modeling efforts: with a 1 cm$^3$ tumor volume containing an ap-

* dengelhardt@ds.dfci.harvard.edu
Figure 2: Comparison of evosim performance using its tau-leaping algorithm versus the SSA for a three-type simulation starting from a single cell. (a) Projections of 5000 simulations for each pairwise combination of cell types shows strong agreement in the clonal distributions between the leaping algorithm and the SSA; (b) Execution times for 30 simulations (same setup as (a)) up to different total population sizes show a logarithmic increase with total population size when tau-leaping is used and an exponential increase with total population size for the SSA. Parameters used: $\beta_1 = \beta_2 = 0.004$ cells/day, $\beta_3 = 2 \beta_1$, $\delta_i = 0.1 \beta_1$ for $i = 1, 2, 3$, $\mu_{1 \rightarrow 2} = 10^{-3}$, $\mu_{1 \rightarrow 3} = \mu_{2 \rightarrow 3} = 10^{-4}$. Initial population sizes were set at $n_1 = 1$, $n_2 = n_3 = 0$. Simulations were set to terminate after 8.5 years of simulation time. For plotting on a logarithmic scale, $\epsilon = 0.01$ was added to results with zero cells.

proximately estimated size of $\sim 10^8 - 10^9$ cells [3], simulating the evolution and progression of such tumors with the SSA is computationally prohibitive. Even at smaller population sizes ($\sim 10^5 - 10^6$), studies involving the optimization of treatment or other evolutionary pressures typically require a large number of simulations, which is precluded by the long execution times of the SSA.

As a result, many approximation approaches to the SSA have been developed over the years. Tau-leaping [4], in which multiple events are sampled over the course of a single, longer time interval (leap) underlies a large number of such approaches. Leaping necessarily leads to an efficiency-accuracy tradeoff, and various tau-leaping methods have been developed with the aim of optimizing this tradeoff as well as preventing the population from becoming negative over a single extended leap. These include different approaches to step size selection ([5, 6]), binomial [7, 8]) and multinomial sampling [9], and estimation bias correction [10], among others.

Here, we present a packaged novel tau-leaping algorithm that combines (1) optimized step-size selection, (2) unbiased estimation via event rate rescalings, and (3) distinct ordering in simulation stepping in order to optimize both accuracy and efficiency at any population size and (polyclonal) structure. We show that this approach exhibits high agreement with SSA results while retaining speed and scalability. This approach applies to continuous-time Markov birth-death-mutation branching processes in which a division can result in either two identical progeny cells or one identical and one mutant progeny cell (Fig. [1]), where cell “types” are defined by their respective rates of birth, death, and mutation to other types. This approach is valid for first-order reaction kinetics in which reactions within species (e.g. birth and death processes) occur at higher rates than interspecies transitions (mutations). Beyond this assumption, our approach is parameter-agnostic in its accuracy and therefore also addresses accuracy gaps in normal-approximation based approaches [11] that can arise when mutation rates are low.

2 Description

The evosim package is a wrapper for this algorithm and consists of two modes, Simulator and EnsembleSimulator, which can be used for recording a single run (of possibly multiple sequential segments) and an ensemble of simulation runs, respectively. The two simulation modes are designed to permit users to specify a sequence of different conditions experienced by the system over any stretch of time, such as different or alternating treatments that affect the birth and death rates of cells or a serial passaging experiment that dilutes the population when a certain confluence level is reached. An evosim simulation run takes as inputs the number of (discrete) cell types (phenotypes, genotypes, or species) that exist or may emerge in the system; the current pop-
ulation distribution, i.e. cell numbers for each of these cell types (which may also be zero for “potential” cell types that could arise by mutation from existing ones); the per-cell birth/division rates and death rates for each cell type; and all probabilities of (inherited) mutations between any two cell types. A cell “type” is uniquely defined by its per-cell birth rate, per-cell death rate, and the probability of mutation from this cell type to all other cell types in the system.

A full simulation may proceed as either a single segment or a sequence of such segments, with different or same values for any of the parameters above, as tracking is maintained until a subsequent reset. This can be used to modify the conditions of the simulation as it proceeds, e.g. introduce a drug through a modification of the birth or death rates. evosim comes equipped with tracking and plotting capabilities, a function for subsampling from the entire population (e.g. as in an in vitro dilution or surgical resection), as well as functions for computing a number of diversity indices (e.g. Shannon and Gini indices) and for converting the time units on the output; this is particularly useful when the user is interested in long-term evolution (e.g. years) but event rates like births and deaths are specified in short-time units (e.g. days). Optionally, a resource capacity and/or population threshold for terminating the simulation may be specified. Plotting is enabled both for clonal trajectories (i.e. different cell types shown in different colors over time) as well as for total population size over time. Usage examples and tutorials for both the Simulator and EnsembleSimulator modes are provided with the package.

Over the simulation time specified by the user, the simulation proceeds by taking adaptive time leaps that depend on the instantaneous population distribution (subpopulations of all cell types) and the per-cell birth and death rates of all nonzero-population cell types. The adaptive leaping scheme is based on the observation that small and large populations evolve on different time scales: the expected time to next event (e.g. division) in a small population of cells is longer than the expected time in a large population of cells with the same per-cell division rate. By assigning an “effective” time scale to the simulation (see mathematical details in the Supplemental Information), the simulation proceeds with logarithmic time complexity $O(\log N)$ in total population size $N$ rather than linear complexity $O(N)$, thus permitting scalability (Fig. 2). Since, as with any tau-leaping scheme, these leaps are larger than the fundamental time scale of the system, a naive application would lead to a biased estimation of the system dynamics. By rescaling the birth and death rates through a derived transformation (see the SI), this bias is removed from the simulation, resulting in very high agreement between the tau leaping simulation results and those of the SSA. Additionally, within each leap populations are sampled in the order of their effective time scales (see pseudocode and explanation in the SI) in order to ensure highly accurate modeling of the effect of mutations on population growth at any population scale without simulation efficiency loss. We note that due to a bias-variance tradeoff, unbiased tau-leaping results in a slight loss of variance (see SI for mathematical results); however, this is a second-order effect that in practice carries minor effect (Fig. 2).

3 Results

A branching process of single cell type with no death takes an average of 212 milliseconds up to a population size of $10^9$ and 300 ms up to $10^{12}$ cells on a personal 2.7 GHz Intel Core i7 CPU with 16 GB RAM running a Jupyter notebook server in OS X. Including a mutation probability per division of 0.01 (two-type process, each type starting at 1 cell) results in an average runtime of 602 ms up to $10^9$ cells and 879 ms up to $10^{12}$ cells. Simulation results confirming high agreement between leaping simulations and the SSA are shown in Fig. 2 and in the SI.

Acknowledgements

We are grateful to Franziska Michor and the Michor Lab for helpful discussions.

Funding

This work was supported by the Helen Gurley Brown Presidential Initiative at Dana-Farber Cancer Institute and the Center for Cancer Evolution at Dana-Farber Cancer Institute. Research reported in this publication was supported in part by The National Institute of General Medical Sciences of the National Institutes of Health under grant number R01GM144962.

References


Rate-Rescaled Tau-Leaping for Accurate and Scalable Simulation of Continuous-Time Branching Processes: Supplemental Information

Dalit Engelhardt\textsuperscript{1,2,3,4} and Thomas O. McDonald\textsuperscript{1,2,3,4}

\textsuperscript{1}Department of Data Science, Dana-Farber Cancer Institute, Boston, MA
\textsuperscript{2}Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA
\textsuperscript{3}The Center for Cancer Evolution, Dana-Farber Cancer Institute, Boston, MA
\textsuperscript{4}Department of Stem Cell and Regenerative Biology, Harvard University, Cambridge, MA

1 Introduction

We present here the mathematical and computational details of the tau-leaping algorithm implemented in the evosim package. In Section 2 we set out notation and define the concept of characteristic-time leaping, a simulation algorithm that incorporates tau-leaping but sets the leap size to the expected time to next event. In Section 3 we present our method (extended-time leaping) for extending these leaps and accordingly rescaling the sampling probabilities to retain simulation accuracy. In Section 4 we derive the efficiency gain of extended-time leaping over the SSA for the single-species case. Supporting simulation results and pseudocode are included.

2 Preliminaries

2.1 Notation and basic tau-leaping

We consider \(C\) interacting cell types, i.e. such that transitions between different cell types may occur. The state of the system at time \(t\) is described by \(x(t) = (x_1(t), ..., x_C(t))\) where \(x_j(t)\) is the population at time \(t\) of the \(j\)th species. The set of self-interactions and inter-species transitions of species \(j\) are described by size-dependent rates \(\{r_j(x_j)\}\), \(j \in \{1, ..., C\}\).

Gillespie’s Stochastic Simulation Algorithm (SSA) \cite{Gillespie1976} is based on the assumption that the time \(\tau\) to the next event in this system is an exponentially-distributed random variable with mean \(\frac{1}{\sum_i r_i(x_i)}\). The SSA thus proceeds by generating two random numbers \(\chi_1, \chi_2 \in [0, 1]\); the time to next reaction is computed as \(\tau = \log(1/\chi_1) / \sum_i r_i(x_i)\), and the next event is chosen in accordance with the reaction probability interval in which \(\chi_2\) is found, where the event probability for event type \(j\) is given by \(r_j(x_j) / \sum_i r_i(x_i)\). The populations are then updated according to the chosen reaction. Since this procedure quickly becomes very computationally expensive at larger system sizes, a number of methods have been developed to increase simulation efficiency. The focus of this work will be on tau-leaping, whose basic form assumes that an interval \(\tau\) can be selected such that the change in \(x(t)\) over \(\tau\) does not lead to appreciable changes in any of the rate functions. We denote by \(K_j(\tau; x, t)\) the number of events, i.e. transitions or self-interactions, \(E_{jk}\) (where \(k\) indexes different events pertaining to species \(j\), e.g. birth, death, or mutation to a different type), taking place within \([t, t + \tau]\) in species \(j\). We define \(v_{ji} \) as the change in \(x_j\) due to an event \(E_{ik}\), where \(i_k\) indexes the set of events \(E_i\) involving species \(i\). The state of the system at \(t + \tau\) is updated as

\[
x_j(t + \tau) = x_j(t) + \sum_{i=1}^{n} \sum_{i_k \in E_i} K_{ik}(\tau; x, t) v_{ji_k},
\] (2.1)
where $K_i \sim (\tau; x, t)$ are random numbers drawn from Poisson distributions with means $\lambda = r_i \tau$.

We will restrict ourselves to rate functions $r_j$ whose size-dependence is given by $r_j(x_j) = g_j x_j$, where $g_j$ is a fixed constant that defines the event rate per individual in cell type $j$ (e.g. the rate at which a single cell divides or dies). Since we consider here branching processes, the system will consist of cell types $X_j$, $j \in (1, ..., C)$ undergoing births, deaths, and mutations via the following reactions

$$X_j \xrightarrow{\beta_j} 2X_j \quad \text{(birth/division)}$$
$$X_j \xrightarrow{\delta_j} \emptyset \quad \text{(death)}$$
$$X_j \xrightarrow{\beta_j \mu_{j,k}} X_j + X_k \quad \text{(mutation)}$$

occurring with respective rates $\beta_j$ (the rate at which a single cell of type $j$ divides), $\delta_j$ (the rate at which a single cell of type $j$ dies), and $\beta_j \mu_{j,k}$ (the rate at which a single cell of type $j$ divides into a $j$ cell and a mutant $k$ cell), where $\mu_{j,k}$ is the probability of mutation from type $j$ to type $k$. As noted in the main text, we will assume throughout that $\mu_{j,k} \ll 1$ for all $j, k$ (this is equivalent to the assumption that mutations occur at much lower rates than birth and death rates). When referring generically to any of these rates, we will use the notation $g_j$, where

$$g_j \in \{\beta_j, \delta_j, \beta_j \mu_{j,k} \text{ for } k \in (1, ..., C), k \neq j\}.$$ 

For simplicity of notation in subsequent sections, we will use in parts of the derivation the notation $g_{j,\pm}$ for $\beta_j (g_{j,+})$ and $\delta_j (g_{j,-})$.

### 2.2 Characteristic-time leaping

The probability of an event $E$ with rate $r_j(x_j) = g_j x_j$ occurring at some point over the interval $\tau$ is given by

$$P(\tau) = 1 - \exp\left[-g_j x_j \tau\right].$$

When $\tau = 1/(g_j x_j)$, in expectation a single event $E$ will occur over the interval $\tau$. We accordingly define the characteristic time $\tau_{\text{char}, j}$ of cell type $j$ at time $t$ as

$$\tau_{\text{char}, j}(t) = \min\left(\frac{1}{g_{j,+} x_j(t)}, \frac{1}{g_{j,-} x_j(t)}\right). \quad (2.2)$$

This is the expected time to the first event (single member addition or removal) in the population of cell type $j$. Simulating the dynamics of cell type $j$ by tau-leaping on intervals of length $\tau_{\text{char}, j}(t)$ is a close approximation to exact simulation, as only one event is expected in each interval, i.e.

$$\langle x_j(t + \tau_{\text{char}, j}) \rangle = x_j(t) + (g_{j,+} - g_{j,-}) x_j(t) \tau_{\text{char}, j} = \left\{\begin{array}{ll}
    x_j(t) + 1 - \frac{g_{j,-}}{g_{j,+}}, & g_{j,+} \geq g_{j,-} \\
    x_j(t) + \frac{g_{j,+}}{g_{j,-}} - 1, & g_{j,+} < g_{j,-}
\end{array}\right. \quad (2.3)$$

where we have incorporated the expectation values of Poisson draws $\lambda_j = g_j x_j \tau$ (see Eq. 2.1). For simplicity, recalling that by assumption $\mu_{j,k} \ll 1 \forall j, k$, we have neglected the contribution of $\mu_{j,i}$ to the rates, which would lead to an adjustment $g_{j,+} \rightarrow g_{j,+} + \left(1 - \sum_{i \neq j} \mu_{j,i}\right)$. If a simulation is carried out over $m$ such steps, then in expectation (where at each step we consider the expectation)

$$\langle x_{j,m} \rangle = \left\{\begin{array}{ll}
    x_{j,0} + m \left(1 - \frac{g_{j,-}}{g_{j,+}}\right), & g_{j,+} \geq g_{j,-} \\
    x_{j,0} + m \left(\frac{g_{j,+}}{g_{j,-}} - 1\right), & g_{j,+} < g_{j,-}
\end{array}\right. \quad (2.4)$$

where $x_{j,m}$ is the population of cell type $j$ at the beginning of the $m$-th step, so that $x_{j,0} = x_j(t)$ for a simulation interval staring at time $t$. We will refer to each step $m \rightarrow m + 1$ of such a simulation as

---

1The discussion here does not in principle preclude rate functions with a dependence on $x_j$, e.g. as in the case of logistic growth; here, however, we restrict ourselves to the linear case.
characteristic-time leaping. We emphasize that the expectation value of the change in the population of $j$ is constant in each such step, as indicated in (2.4), but that the time interval corresponding to each such step ($\tau_{\text{char},j}$) will vary with the population size (2.2) and hence may change with each step. Written differently, over an interval of length $T$ we have in expectation that

$$
\langle x_j (t + T) \rangle = \begin{cases} 
x_j (t) + M (T) \left( 1 - \frac{g_{j,+}}{g_{j,-}} \right), & g_{j,+} \geq g_{j,-} \\
x_j (t) - M (T) \left( 1 - \frac{g_{j,+}}{g_{j,-}} \right), & g_{j,+} < g_{j,-}
\end{cases}
$$

(2.5)

where $M (T)$ is the number of $\tau_{\text{char},j}$-length simulations (each of potentially different length), i.e. the number of Poisson sampling draws for each event, that cover the interval $T$. In expectation, $M$ is the solution to

$$
T (M) = \sum_{m=0}^{M-1} \frac{1}{\max \{ \langle g_{j,\pm} \rangle \langle x_{j,m} \rangle \}} = \sum_{m=0}^{M-1} \left\{ \left[ \frac{g_{j,+} \left( x_{j,0} + m \left( 1 - \frac{g_{j,-}}{g_{j,+}} \right) \right)}{g_{j,-} \left( x_{j,0} + m \left( 1 + \frac{g_{j,-}}{g_{j,+}} \right) \right)} \right]^{-1}, \ g_{j,+} \geq g_{j,-} \\
\left[ \frac{g_{j,-} \left( x_{j,0} + m \left( 1 - \frac{g_{j,+}}{g_{j,-}} \right) \right)}{g_{j,+} \left( x_{j,0} + m \left( 1 + \frac{g_{j,+}}{g_{j,-}} \right) \right)} \right]^{-1}, \ g_{j,+} < g_{j,-}
\right\}

(2.6)

where (see Eq. 2.7)

$$
\tau_{\text{char},j} (t_m) \equiv \min \left( \frac{1}{g_{j,+} x_{j,m}}, \frac{1}{g_{j,-} x_{j,m}} \right).
$$

(2.7)

In closed form (2.6) is given by

$$
T (M) = \begin{cases} 
\psi \left( M + \frac{z_j^{(i)}}{\tau} \right) - \psi \left( \frac{z_j^{(i)}}{\tau} \right), & g_{j,+} \geq g_{j,-} \\
\frac{G_+ g_{j,+}}{G_- g_{j,-}} \psi \left( M + \frac{z_j^{(i)}}{\tau} \right) - \psi \left( \frac{G_+ g_{j,+}}{G_- g_{j,-}} \right), & g_{j,+} < g_{j,-}
\end{cases}
$$

(2.8)

where $\psi (z) = \Gamma' (z) / \Gamma (z)$ is the digamma function and, for notational simplicity, we have defined

$$
\begin{align*}
G_+ &\equiv 1 - \frac{g_{j,+}}{g_{j,-}} \\
G_- &\equiv \frac{g_{j,+}}{g_{j,-}} - 1.
\end{align*}
$$

2.3 Multinomial sampling

As with the SSA, characteristic-time leaping scales with population size and becomes computationally expensive at large populations due to the large number of sampling draws required. We wish to establish whether leaping on longer intervals can sufficiently well approximate characteristic-time leaping; however, the possibility of negative population counts is a general feature of tau-leaping while sampling from a Poisson distribution since at larger values of $M$ the population may become negative in a single leap. One approach for preventing negative population counts during tau-leaping is to substitute Poisson sampling for binomial sampling [3, 4]. In binomial tau-leaping the number of events is sampled from a binomial distribution with mean $n_j p_j = \lambda_j = g_j x_j \tau$, where $\lambda_j$ is the Poisson parameter, under the assumption that this provides a good approximation in the regime of large $n_j$ and small $p_j$, where $n_j$ is the current population size of species $j$ and $p_j$ is the probability of having the event of interest (e.g. a cell division) occur in any member of species $j$. Here, since a cell type can have multiple rates $g_j$, we will consider the multinomial distribution. We will first establish the validity of multinomial sampling in the case of characteristic-time leaping – where the characteristic-time Poisson sampling described above is the baseline for comparison – before proceeding to the case of extended leap times. We will first focus on the rates $g_{j,\pm}$ that add or remove members of cell type $j$ and return to inter-type transition rates (i.e. mutations) in Sec 3.2.

When sampling from the multinomial distribution we consider the set of probabilities $p_{\pm} (\tau) = g_{j,\pm} \tau$ as well as the probability that no change has occurred, $p_{\text{null}} (\tau) = 1 - p_+ (\tau) - p_- (\tau)$, so that the expected size of the population after a sampling is given by

$$
x_j \left( 2p_+ - p_- + 1 - p_+ - p_- \right) = x_j \left( 1 + p_+ - p_- \right),
$$
where the first two terms represent, respectively, cell division (hence the factor of 2) and cell death. Then, substituting \( p_\pm = g_{j,\pm} \tau \), we observe that

\[
\langle x_j (t + \tau_{\text{char},j}) \rangle = x_j (t) + (g_{j,+} - g_{j,-}) x_j (t) \tau_{\text{char},j} = \begin{cases} x_j (t) + 1 - \frac{g_{j,-}}{g_{j,+}}, & g_{j,+} \geq g_{j,-} \\ x_j (t) - 1 + \frac{g_{j,+}}{g_{j,-}}, & g_{j,+} < g_{j,-} \end{cases}
\]  

(2.9)

identically to that obtained through Poisson sampling above \( \frac{2}{3} \), showing that multinomial sampling results in an unbiased estimate of the population size (as compared to Poisson) in characteristic time simulations. We can show \( \frac{A.3}{} \) that the variance is given by

\[
\sigma^2 (t + \tau) = x_j (t) \left[ p_+ (\tau) + p_- (\tau) - (p_+ (\tau) - p_- (\tau))^2 \right]
\]

(2.10)
or, once again substituting \( p_\pm = g_{j,\pm} \tau \),

\[
\sigma^2_{\text{multinomial}} (t + \tau_{\text{char},j}) = \begin{cases} 1 + \frac{g_{j,-}}{g_{j,+}} - \left( 1 - \frac{g_{j,-}}{g_{j,+}} \right)^2 \frac{1}{x_j (t)}, & g_{j,+} \geq g_{j,-} \\ - \frac{g_{j,-}}{g_{j,+}} + 1 - \left( \frac{g_{j,+}}{g_{j,-}} - 1 \right)^2 \frac{1}{x_j (t)}, & g_{j,+} < g_{j,-} \end{cases}
\]

(3.1)
The variance in Poisson sampling will be the sum of variances \( \lambda_j \) and \( \lambda_j^-' \),

\[
\sigma^2_{\text{Poission}} (t + \tau_{\text{char},j}) = (g_{j,+} + g_{j,-}) x_j (t) \tau_{\text{char},j} = \begin{cases} 1 + \frac{g_{j,-}}{g_{j,+}}, & g_{j,+} \geq g_{j,-} \\ \frac{g_{j,+}}{g_{j,-}} + 1, & g_{j,+} < g_{j,-} \end{cases}
\]

(2.11)

so that the fractional deviation in the variance is of \( \mathcal{O} \left( x_j^{-1} \right) \):

\[
\eta_{\text{var}} (t + \tau_{\text{char},j}) = \frac{1}{x_j (t)} \left\{ \frac{\left( 1 - \frac{g_{j,-}}{g_{j,+}} \right)^2}{1 + \frac{g_{j,-}}{g_{j,+}}}, \ g_{j,+} \geq g_{j,-} \right\} \left\{ \frac{\left( 1 - \frac{g_{j,+}}{g_{j,-}} \right)^2}{1 + \frac{g_{j,+}}{g_{j,-}}}, \ g_{j,+} < g_{j,-} \right\}
\]

(2.12)

and hence at large \( x_j \), as expected, multinomial sampling provides an excellent approximation to Poisson sampling. We will define \( x_{\text{threshold}} \) at the threshold value at which we consider \( \frac{2}{12} \) to be acceptably low (e.g. \( x_j \sim 100 - 1000 \)).

### 3 Extended-time leaping: estimation accuracy

#### 3.1 Self-interactions

Let \( \mathbb{C} = \{1, ..., C\} \) be the set of all cell types and define

\[
\theta_j = \begin{cases} x_j, & x_j < x_{\text{threshold}} \\ x_{\text{threshold}}, & x_j \geq x_{\text{threshold}} \end{cases}
\]

(3.1)

and the \textit{effective} characteristic time

\[
\tau_{\text{eff},j} (t) = \frac{1}{\theta_j (t) \max \{g_{j,+}, g_{j,-}\}}
\]

(3.2)

while recalling that the true characteristic time is given by

\[
\tau_{\text{char},j} (t) = \frac{1}{x_j (t) \max \{g_{j,+}, g_{j,-}\}}
\]

(3.3)

Then we set the simulation time step of the system at time \( t \) to

\[
\tau_{\text{step}} (t) = \min \{ \tau_{\text{eff},j} (t) \ \forall \ j \in \mathbb{C} \}
\]

(3.4)
We will refer to simulations that involve a single draw from a multinomial distribution \( \text{Mult}(n_j, p_j) \), where \( p_j = (p_{j,+}, p_{j,-}, 1 - p_{j,+} - p_{j,-}) \) over the time interval \( T = \tau_{\text{step}}(t) \) as extended-time leaping.

We define the following transformations of the multinomial sampling parameters:

\[
\begin{cases}
  n_j = x_j(t), p_{j,\pm} = g_{j,\pm}\tau_{\text{step}} & \text{if } \tau_{\text{char}} \geq \tau_{\text{step}} \\
  n_j = x_j(t), p_{j,\pm} = \gamma(g_{j,\pm}\tau_{\text{step}})\tau_{\text{step}} & \text{if } \tau_{\text{char}} < \tau_{\text{step}}
\end{cases}
\]

(3.5)

where

\[
\gamma(g_{j,\pm}, \tau_{\text{step}}) = \begin{cases}
  \left[ \frac{M(g_{j,\pm}, \tau_{\text{step}})}{x_j(t)\tau_{\text{step}}} \left(1 - \frac{g_{j,-}}{g_{j,+}}\right) + g_{j,-}\right] & , g_{j,+} \geq g_{j,-} \\
  g_{j,+} & , g_{j,+} < g_{j,-}
\end{cases}
\]

(3.6)

and

\[
\gamma(g_{j,-}, \tau_{\text{step}}) = \begin{cases}
  \left[ \frac{M(g_{j,-}, \tau_{\text{step}})}{x_j(t)\tau_{\text{step}}} \left(1 - \frac{g_{j,+}}{g_{j,-}}\right) + g_{j,+}\right] & , g_{j,+} \geq g_{j,-} \\
  g_{j,+} & , g_{j,+} < g_{j,-}
\end{cases}
\]

(3.7)

such that \( M(g_{j,\pm}, \tau_{\text{step}}) \) is the solution to (2.8) with \( T = \tau_{\text{step}} \).

Claim 1. Under the transformations of Eqns. 3.3, 3.6 and 3.7, extended-time leaping of leap size \( T = \tau_{\text{step}} \) does not lead to any appreciable degradation in the estimation accuracy of the mean and variance of the resulting population distributions for any cell type \( j \in \mathcal{C} \) as compared with characteristic-time leaping (with characteristic time \( \tau_{\text{char},j} \) over the interval \( \tau_{\text{step}} \)).

Proof. If for cell type \( j \) \( \tau_{\text{char},j} > \tau_{\text{step}} \), multinomial sampling at more frequent intervals simply represents a closer approximation to the continuous-time simulation limit, and therefore no loss of accuracy is incurred. If \( \tau_{\text{char},j} < \tau_{\text{step}} \), we have that under multinomial sampling the expected value will be given by

\[
\langle x_j(t + \tau_{\text{step}}) \rangle = n_j (1 + p_{j,+} - p_{j,-})
\]

so that, substituting in the probabilities \( p_{j,\pm} \), we have

\[
\langle x_j(t + \tau_{\text{step}}) \rangle = \begin{cases}
  x_j(t) + M \left(1 - \frac{g_{j,-}}{g_{j,+}}\right), & g_{j,+} \geq g_{j,-} \\
  x_j(t) - M \left(1 - \frac{g_{j,+}}{g_{j,-}}\right), & g_{j,+} < g_{j,-}
\end{cases}
\]

(3.8)

where \( M = M(g_{j,\pm}, \tau_{\text{step}}) \). From (2.5), \( \tau_{\text{step}} \), we see that the expected value (3.8) after an extended \( \tau_{\text{step}} \) leap is the same as that obtained over the requisite number of characteristic-time Poisson sampling steps.

The variance is given by (2.10):

\[
\sigma_j^2(t + \tau_{\text{step}}) = \begin{cases}
  M \left(1 - \frac{g_{j,-}}{g_{j,+}}\right) + 2g_{j,-}\tau_{\text{step}}x_j(t) - \frac{1}{x_j(t)} M^2 \left(1 - \frac{g_{j,-}}{g_{j,+}}\right)^2, & g_{j,+} \geq g_{j,-} \\
  M \left(1 - \frac{g_{j,+}}{g_{j,-}}\right) + 2g_{j,+}\tau_{\text{step}}x_j(t) - \frac{1}{x_j(t)} M^2 \left(1 - \frac{g_{j,+}}{g_{j,-}}\right)^2, & g_{j,+} < g_{j,-}
\end{cases}
\]

(3.9)

To see how (3.9) compares to the variance from characteristic-time leaping, we note that when \( x_j(t) \) is large, i.e. \( x_j \geq x_{\text{thresh}} \), as we assume here, we can asymptotically expand the digamma functions \( \psi(z) \sim \log z \) in (2.8) to obtain the solutions

\[
M \approx x_j(t) (\exp[\tau_{\text{step}}(g_{j,+} - g_{j,-})] - 1) \left( \frac{1}{g_{j,+}} \right)^{-1}, \quad g_{j,+} \geq g_{j,-} \\
M \approx x_j(t) (\exp[\tau_{\text{step}}(g_{j,+} - g_{j,-})] - 1) \left( \frac{g_{j,+}}{g_{j,-}} - 1 \right)^{-1}, \quad g_{j,+} < g_{j,-}
\]

(3.10)

Next, we let \( \alpha \) denote the species for which \( \tau_{\text{eff},\alpha} = \tau_{\text{step}} \) (\( \alpha \) may or may not be the same as \( j \)) and further expand (3.10) in \( x_{\text{thresh}} \) (recalling the expression for \( \tau_{\text{step}} \), Eqns. 3.2 and 3.4) as

\[
M \approx x_j(t) \left( \frac{g_{j,+}}{g_{\alpha}} \frac{1}{x_{\text{thresh}}} + \left( \frac{g_{j,+}}{g_{\alpha}} \right)^2 \frac{1}{x_{\text{thresh}}} \left(1 - \frac{g_{j,-}}{g_{j,+}}\right) + O \left( x_{\text{thresh}}^{-3}\right), \quad g_{j,+} \geq g_{j,-} \\
M \approx x_j(t) \left( \frac{g_{j,-}}{g_{\alpha}} \frac{1}{x_{\text{thresh}}} + \left( \frac{g_{j,-}}{g_{\alpha}} \right)^2 \frac{1}{x_{\text{thresh}}} \left(\frac{g_{j,+}}{g_{j,-}} - 1\right) + O \left( x_{\text{thresh}}^{-3}\right), \quad g_{j,+} < g_{j,-}.
\]

(3.11)
Substituting into (3.9), we obtain

\[
\sigma_j^2 (t + \tau_{\text{step}}) = \left\{ \begin{array}{ll}
\left( \frac{g_{j,+}}{g_{\alpha}} \right. & \left( \frac{g_{j,-}}{g_{\alpha}} \right) x_{\text{thresh}} x_j(t) - \frac{1}{2} \left( \frac{g_{j,+}-g_{j,-}}{g_{\alpha}} \right) x_j(t) + O \left( \frac{x_j(t)}{x_{\text{thresh}}} \right), & g_{j,+} \geq g_{j,-} \\
\left( g_{j,+} + g_{j,-} \right) x_{\text{thresh}} x_j(t) - \frac{1}{2} \left( g_{j,+} - g_{j,-} \right)^2 x_j(t) + O \left( \frac{x_j(t)}{x_{\text{thresh}}} \right), & g_{j,+} < g_{j,-}
\end{array} \right.
\]  

(3.12)

On the other hand, \( M (\tau_{\text{step}}) \) steps of characteristic-time Poisson simulations result in a cumulative variance (see Eq. 2.11) of

\[
\sigma^2_{\text{Poisson}} (t + \tau_{\text{step}}) = M \sigma^2_{\text{Poisson}} (t + \tau_{\text{char,j}}) = M (\tau_{\text{step}}) \left\{ \begin{array}{ll}
1 + \frac{g_{j,+}}{g_{j,-}}, & g_{j,+} \geq g_{j,-} \\
\frac{g_{j,+}}{g_{j,-}} + 1, & g_{j,+} < g_{j,-}
\end{array} \right.
\]

which, substituting \( M \) from the expressions (3.11), becomes

\[
M \sigma^2_{\text{Poisson}} (t + \tau_{\text{char,j}}) = \left( \frac{g_{j,+}}{g_{\alpha}} + \frac{g_{j,-}}{g_{\alpha}} \right) x_{\text{thresh}} \frac{x_j(t)}{2} - \frac{1}{2} \frac{g_{j,+}^2 - g_{j,-}^2}{g_{\alpha}^2} \frac{x_j(t)}{x_{\text{thresh}}}.
\]

(3.13)

We see that to leading order the variance obtained in \( M (\tau_{\text{step}}) \) Poisson characteristic-time simulations (3.13) is the same as that of a single leap \( \tau_{\text{step}} \) with multinomial sampling (3.12), and the fractional deviation in variances is of \( O \left( \frac{x_{\text{thresh}}}{\tau_{\text{thresh}}} \right) \). This deviation is on the same order as that obtained in (2.12) for characteristic-time leaping, which we considered to be acceptably low for a fixed choice of \( x_{\text{thresh}} \sim 100 \), with extended-time leaping slightly overestimating the variance (by, to leading order, \( g_{j,+} / (g_{j,+}^2 x_{\text{thresh}}) \)) in the population of \( j \) if \( g_{j,+} > g_{j,-} \) or if \( g_{j,-} > 2g_{j,+} \) and slightly underestimating the variance if \( g_{j,+} < g_{j,-} < 2g_{j,+} \).

Fig. 1 shows a comparison of single cell type simulation outcomes for extended-time leaping and the SSA for two different leaping threshold levels \( x_{\text{thresh}} = 10^2, 10^3 \) over three different simulation time intervals, with identical initialization parameters used for all simulations. Strong overlap in population sizes is observed in all cases. SSA runtimes are shown to increase linearly with population size whereas extended-time leaping runtimes only increase logarithmically with population size.

Finally, we note that the transformations (3.7) and (3.6) amount to transformations of the demographic rates. I.e. we have shown that extending the simulation interval while implementing an appropriate rescaling of the relevant event rates results in a similar distribution of outcomes (unbiased estimate, with an acceptably low deviation in variances). Substituting in the expression for \( M (\tau_{\text{char,j}}) \), we can approximate the rate transformations as

\[
g_{j,+} \rightarrow \left\{ \begin{array}{ll}
\exp \left[ \frac{\tau_{\text{min}} (g_{j,+} - g_{j,-})}{\tau_{\text{min}}} \right] + g_{j,-}, & g_{j,+} \geq g_{j,-} \\
g_{j,+}, & g_{j,+} < g_{j,-}
\end{array} \right.
\]

(3.14)

\[
g_{j,-} \rightarrow \left\{ \begin{array}{ll}
g_{j,-} \left[ 1 - \exp \left( \frac{\tau_{\text{min}} (g_{j,+} - g_{j,-})}{\tau_{\text{min}}} \right) \right], & g_{j,+} \geq g_{j,-} \\
g_{j,+} + \frac{g_{j,+}}{g_{\alpha}} - \frac{\tau_{\text{min}}}{2} (g_{+} - g_{-})^2 + O \left( \frac{\tau_{\text{min}} (g_{+} - g_{-})^3}{\tau_{\text{min}} (g_{+} - g_{-})^3} \right), & g_{j,+} < g_{j,-}
\end{array} \right.
\]

(3.15)

If we expand to second order, we see that the rate transformations are given by

\[
g_{+} \rightarrow g_{+} + \frac{\tau_{\text{min}}}{2} (g_{+} - g_{-})^2 + O \left( \frac{\tau_{\text{min}} (g_{+} - g_{-})^3}{\tau_{\text{min}} (g_{+} - g_{-})^3} \right)
\]

\[
g_{-} \rightarrow g_{-} - \frac{\tau_{\text{min}}}{2} (g_{+} - g_{-})^2 + O \left( \frac{\tau_{\text{min}} (g_{+} - g_{-})^3}{\tau_{\text{min}} (g_{+} - g_{-})^3} \right)
\]

so that extended-time leaping retains simulation accuracy while simulating in larger steps by increasing birth rates \( g_{j,+} \) when a subpopulation is growing \( g_{j,+} > g_{j,-} \) or decreasing death rates \( g_{j,-} \) when a subpopulation is shrinking \( g_{j,+} < g_{j,-} \).

3.2 Inter-species transitions

While rate rescaling in \( g_{\pm} \), as described in the previous section, ensures an unbiased estimate of the growing or declining species subpopulation, we must further ensure that flux to other subpopulations is also estimated correctly under extended-time leaping. For a transition (mutation) probability \( \mu \equiv \mu_{j,k} \) from subpopulation...
Figure 1: Single-species simulations show strong agreement and a high efficiency increase factor between extended-time leaping (orange) and the SSA (blue). In each panel the first three plots show the population size density distribution over 5000 SSA and 5000 extended-time leaping simulations for the specified number of simulated years and $x_{\text{thresh}}$, and the rightmost plot shows the respective (leaping and SSA) distributions of execution times versus population size. Simulation parameters were set at $g_+ = 0.004 \text{ day}^{-1}$, $g_- = 0.1 g_+$ and initiated with a single cell in all cases. All plots are shown on a log-log scale.

To subpopulation $k$, the expected number $\Omega_{j \to k, \text{ETL}}$ of new members of $k$ as a result of transitions from $j$ after time interval $\tau_{\text{min}}$, starting at time $t_0$, is given by $\mu x_j(t) p_{j,+}(\tau_{\text{step}})$. From (3.5) and (3.6), this is given by

$$\mu x_j(t) \tau_{\text{step}} \left\{ \frac{M(g_{j,+} - g_{j,-})}{x_j(t) \tau_{\text{step}}} \right\}^{\frac{1 - g_{j,+}}{g_{j,+}}, \ g_{j,+} \geq g_{j,-}, \ g_{j,+} < g_{j,-}} (3.16)$$

On the other hand, after $M$ steps of characteristic-time leaping, in expectation the total flux to $k$, $\Omega_{j \to k, \text{CTL}}$, is given by

$$\Omega_{j \to k, \text{CTL}} = M \mu \left\{ \frac{1}{g_{j,+}}, \ g_{j,+} \geq g_{j,-}, \ g_{j,+} < g_{j,-} \right\} (3.17)$$

We can therefore show that a rescaling of the mutation rate $\mu \rightarrow \mu'$ such that

$$\mu' = \mu \left\{ \frac{1 - g_{j,-}}{g_{j,+}} + \frac{g_{j,-} - x_j(t) \tau_{\text{step}}}{M} \right\}^{-1}, \ g_{j,+} \geq g_{j,-}, \ g_{j,+} < g_{j,-} (3.18)$$

will result in equality of expected fluxes under extended-time leaping.
As we loop through sampling the different subpopulations (see Algorithm 1), mutations are added to the subpopulations into which they occur. Since intervals are larger in the leaping regime than the characteristic time scales of system, there is a potential risk of underestimating the effects of mutations on smaller populations: once a cell has been born in a certain cell type, that cell can then divide and thus continue to expand this subpopulation; however, if such mutations are only accounted for after many characteristic time steps, the subpopulation size will be underestimated. To counteract this effect, at each simulation step all cell types are ordered by their effective characteristic times, and sampling is performed in this order. As a result, populations with larger time scales that evolve more slowly receive their proper mutation influx within their characteristic time step, hence avoiding the underestimation problem (see Algorithm 1).

Algorithm 1: Extended-time leaping for subpopulations \( j \in \text{sim\_pops} \) over time interval \( \text{sim\_time} \)

\[
t \leftarrow 0;
\]
\[
\text{while } t < \text{sim\_time} \text{ do}
\]
\[
\text{effective\_times} \leftarrow \left\{ (\theta_j(t) \max\{g_{j,+}, g_{j,-}\})^{-1} \right\}_j \in \text{sim\_pops} \quad (3.2); \\
\tau_{\text{step}} \leftarrow \min \{\text{effective\_times}\} \quad (3.4); \\
\text{for subpopulation } j \in \text{sim\_pops} \text{ (s.t. } x_j > 0) \text{, in increasing order of effective\_times, do}
\]
\[
\text{retrieve current subpopulation size } x_j, \text{ rates } g_{j,+} \text{ and } g_{j,-}, \text{ and mutation probabilities } \vec{\mu}_{j \rightarrow i} \text{ for set } U_j \text{ of possible mutants with populations } u_i; \\
\tau_{\text{char}} \leftarrow (x_j(t) \max\{g_{j,+}, g_{j,-}\})^{-1} \quad (3.3) \text{ if } \tau_{\text{char}} < \tau_{\text{step}} \text{ then}
\]
\[
\text{transform } g_{j,\pm} \text{ according to } (3.14) \text{ and } (3.15) \text{ transform } \vec{\mu}_{j \rightarrow i} \text{ according to } (3.18)
\]
\[
p_{j,+} \leftarrow g_{j,+} \left(1 - \sum_{i \in U_j} \mu_{j \rightarrow i}\right) \tau; \\
p_{j,-} \leftarrow g_{j,-} \tau; \\
\vec{p}_{j \rightarrow i} \leftarrow \{g_{j,+} \mu_{j \rightarrow i} \tau \forall \ i \in U_j \}; \\
x = (x_{j,0}, x_{j,1}, x_{j,2}, u_1, ..., u_m) \leftarrow \text{sample from the multinomial distribution with probabilities } (1 - p_{j,+} - p_{j,-}, p_{j,+}, p_{j,-}, p_{j \rightarrow 1}, ..., p_{j \rightarrow m}), \text{ where } m = |U_j|; \\
x_j \leftarrow \max (0, x_{j,0} + 2x_{j,1}); \\
\text{for } i \in U_j \text{ do}
\]
\[
\left\{ x_i \leftarrow x_i + u_i \text{ } \forall \ i \in U_j; \\
\right.
\]
\[
t \leftarrow t + \tau_{\text{step}} \\
\text{end}
\]
\[
\text{end}
\]
\[
\text{end}
\]

Fig. 2 shows a comparison of how well extended-time leaping performs against the SSA in two-species simulation when rescalings of the sampling probabilities are implemented versus when extended-time leaping is carried naively without any rescalings. Both simulation setups shown in Fig. 2 exhibit high agreement with the SSA when rescaling is implemented and decreased agreement otherwise.

4 Extended-time leaping: time complexity

4.1 The single-species case

The time complexity of a population growth simulation is defined by the number of steps – each consisting of a multinomial sampling – needed to grow its population from an initial size \( n_0 \) to a final specified size \( N \). We first consider the case of a single cell type with \( g_+ > g_- \). When \( n_0, N < x_{\text{thresh}} \), then the simulation is
in the characteristic-leaping regime for its entire duration. In this case, from (2.5),
\[
M(N, n_0, r_g) = \frac{N - n_0}{1 - g_{j,-}/g_{j,+}} \tag{4.1}
\]
steps are taken to reach size \( N \). i.e. characteristic-time leaping is, as expected, \( O(N) \) complexity. If no extended-time leaping is implemented, the simulation is of linear time complexity for arbitrarily large \( N \).

With extended leaping, \( \tau_{\text{step}} = 1/(x_{\text{thresh}} g_+) \), and every leap \( \ell \) covers a number \( M_\ell \) of characteristic-time steps (see Eqn. 2.6):
\[
\frac{1}{x_{\text{thresh}}} = \sum_{m=0}^{M_\ell-1} \left[ x_{j,\ell-1} + m \left( 1 - \frac{g_{j,-}}{g_{j,+}} \right) \right]^{-1}
\]
where \( x_{j,\ell} \) is the population size at the end of leap \( \ell - 1 \), i.e., recalling (2.8),
\[
\psi \left( M_\ell + \frac{x_{j,\ell-1}}{1 - g_{j,-}/g_{j,+}} \right) - \psi \left( \frac{x_{j,\ell-1}}{1 - g_{j,-}/g_{j,+}} \right) = \frac{1 - g_{j,-}}{g_{j,+}} x_{\text{thresh}},
\]
or, again using the digamma asymptotic expansion,

\[
M_\ell \approx x_{j,\ell-1} \left( \exp \left[ \frac{1}{x_{\text{thresh}}} \left( 1 - \frac{g_{j,-}}{g_{j,+}} \right) \right] - 1 \right)
\]

(4.2)

where, as we have previously established,

\[
\langle x_{j,\ell} \rangle = x_{j,\ell-1} + M_\ell \left( 1 - \frac{g_{j,-}}{g_{j,+}} \right)
\]

which we can therefore write as

\[
\langle x_{j,\ell} \rangle = x_{j,\ell-1} \exp \left[ \frac{1}{x_{\text{thresh}}} \left( 1 - \frac{g_{j,-}}{g_{j,+}} \right) \right],
\]

and which, in expectation, has the explicit solution

\[
\langle x_{j,\ell} \rangle = x_{j,0} \exp \left[ \frac{\ell}{x_{\text{thresh}}} \left( 1 - \frac{g_{j,-}}{g_{j,+}} \right) \right].
\]

(4.3)

Recall that \( M \) constitutes the number of multinomial samplings under characteristic-time leaping. We denote by \( L \) the number of multinomial samplings under extended-time leaping, i.e. \( L \) represents the number of extended-time leaps (of length \( r_{\text{min}} \)). When leaping starts at population \( n_0 \), then, from (4.3),

\[
L(N, n_0, g_+, g_-) = x_{\text{thresh}} \frac{\log (N) - \log (n_0)}{1 - g_- / g_+}
\]

(4.4)

samplings will be needed by the time the population reaches size \( N \), i.e. extended-time leaping is \( O(\log N) \) complexity.

In a one-species SSA simulation (see Algorithm 2), on the other hand, at the \( m \)-th simulation step the following may occur with probabilities \( p_{\pm,0} \):

\[
x_{m+1} = \begin{cases} 
x_m + 1, & p_+ = \frac{g_{+} x_m}{(g_{+} + g_{-}) x_m} = \frac{1}{1 + g_- / g_+} \\
x_m - 1, & p_- = \frac{g_{-} x_m}{(g_{+} + g_{-}) x_m} = \frac{1}{1 + g_+ / g_-} \\
x_m, & p_0 = 1 - \frac{1}{1 + g_- / g_+} - \frac{1}{1 + g_+ / g_-}
\end{cases}
\]

so that

\[
\langle x_{m+1} \rangle = \frac{x_m + 1}{1 + g_- / g_+} + \frac{x_m - 1}{1 + g_+ / g_-} + x_m \left( 1 - \frac{1}{1 + g_- / g_+} - \frac{1}{1 + g_+ / g_-} \right)
\]

\[
= x_m + \frac{g_+ - g_-}{g_+ + g_-}
\]

In expectation, then, after \( M \) steps, starting from population \( n_0 \), we have that

\[
\langle x_m \rangle = n_0 + M \left( \frac{g_+ - g_-}{g_+ + g_-} \right).
\]

For a final population \( N \), we can solve for \( M \) as

\[
M = (N - n_0) \left( \frac{1 + g_- / g_+}{1 - g_- / g_+} \right)
\]

(4.5)

so that the SSA operates with linear, \( O(N) \) complexity.

The efficiency gain to extended-time leaping with threshold \( x_{\text{thresh}} \) from \( x = x_{\text{thresh}} \) onwards is given by

\[
\frac{M(N_0, N, g_+, g_-)}{L(N, x_{\text{thresh}}, g_+, g_-)} = \frac{(N / x_{\text{thresh}} - n_0 / x_{\text{thresh}})}{\log (N / x_{\text{thresh}})} \left( 1 + \frac{g_-}{g_+} \right).
\]

(4.6)

Fig. 3 shows a comparison of speed and number of simulation steps for a SSA simulation versus extended-time leaping set at multiple thresholds. Simulation results confirm the logarithmic complexity of extended-time leaping versus the linear complexity of the SSA. We show in Fig. 4 as well as in the main text that these relationships continue to hold for larger numbers of cell types.
Figure 3: Extended-time leaping leads to significant gains in simulation speed. Shown above is total simulation time versus the population size at simulation termination for single-species growth simulations. All simulations were started at population size $x = 100$ and terminated at a sequence of increasing times (30 through 170 months, in 20 month intervals, with 100 simulation runs at each termination point). The fits shown in dashed lines represent logarithmic fits (Eqn. 4.4) for population average versus execution time (note that Eqn. 4.4 is representative of number of steps, of which runtime is a proxy) at each termination point for $x_{\text{thresh}} = 100$ ($r^2 = 0.982$), $x_{\text{thresh}} = 600$ ($r^2 = 0.987$), and $x_{\text{thresh}} = 1000$ ($r^2 = 0.995$), and a linear fit (Eqn. 4.5) for the runtime and population averages for the SSA ($r^2 = 0.9997$). Simulation parameters used were the same as in Fig. 1.

Figure 4: Execution times scale linearly with population size for the SSA and logarithmically with population size for extended-time leaping. Figures (a) and (b) show, respectively, the execution times versus total population size for the configurations shown in (a) and (b) of Fig. 2 at different leaping thresholds. Simulations were carried out in each case for 5, 6, 7, and 8 years of simulation time, with 100 simulations at each time point. Fits shown are linear for the SSA and logarithmic for leaping. The SSA is shown to result in equivalent or faster execution times in small populations but takes prohibitively long times to simulate at larger population sizes. The particular simulation parameters and leaping threshold used determine the crossover point, typically at $10^3 - 10^4$ for $x_{\text{thresh}} = 10^2$. 
Algorithm 2: SSA simulation of duration time \textbf{SSA.time} for set of subpopulations \{\textbf{X}_{SSA}\} comprising the full system (\textbf{SSA.type} = \textit{pure}) or a subsystem (\textbf{SSA.type} = \textit{hybrid} starting at simulation time \textit{t})

\begin{verbatim}
current_time \leftarrow t;
while current_time < t + SSA.time do
    retrieve from mutant_birth_times_and_number all transitions \{\mu\} recorded as occurring prior to current_time;
    update population sizes with the relevant new transitions and clear these transitions from mutant_birth_times_and_number;
    update if necessary the set of populations \{\textbf{X}_{SSA}\} to be simulated with the SSA;
for \textit{j} in \{\textbf{X}_{SSA}\} do
    calculate transition rates for \textit{j}:
    \( r_{j,+} \leftarrow g_{j,+} \left( 1 - \sum_{i \in U_j} \mu_{j \rightarrow i} \right) x_j; \)
    \( r_{j,-} \leftarrow r_{j,-} x_j; \)
    \( \vec{r}_{j \rightarrow i} \leftarrow \{ g_{j,+} \mu_{j \rightarrow i} x_j \ \forall \ i \in U_j \}; \)
end
r_0 = \sum_k r_k;
a_k = r_k / r_0 \text{ with } r_k \text{ ordered from smallest to largest such that } k \text{ represents an individual transition;}
drawn two pseudorandom numbers \( r_1, r_2 \) from the uniform distribution on the unit interval;
\( \Delta t \leftarrow \log \left( 1 / r_1 \right) / r_0; \)
time_remaining \leftarrow SSA.time + t - current_time;
if \( \Delta t > \text{time_remaining} \) then
    current_time \leftarrow current_time + time_remaining;
    break
end
current_time \leftarrow current_time + \Delta t;
select transition \( \lambda \) to occur based on the interval of \( a_\lambda \) coinciding with \( r_2 \), i.e. select transition \( \lambda \) s.t. \( a_{\lambda+1} < r_2 \leq a_\lambda \);
Add (remove) 1 cell to (from) the appropriate population based on the selected transition;
if \textbf{SSA.type} = \textit{pure} and \( \exists \ \textit{j} \in \{\textbf{X}_{SSA}\} \ s.t. \ x_j \geq x_{\text{thresh}} \) then
    break
end
return current_time
\end{verbatim}

A Appendix

The sampling variance at any particular step is given by

\[
\sigma^2 (x_{\text{no change}} + 2x_{\text{divided}}) = E (x_{\text{no change}}^2 + 4x_{\text{no change}}x_{\text{divided}} + 4x_{\text{divided}}^2) \\
- (E (x_{\text{no change}} + 2x_{\text{divided}}))^2 \\
= E (x_{\text{no change}}^2 + 4E (x_{\text{no change}}x_{\text{divided}}) + 4E (x_{\text{divided}}^2) \\
- E^2 (x_{\text{no change}}) - 4E (x_{\text{no change}}) E (x_{\text{divided}}) - 4E^2 (x_{\text{divided}})
\]

where

\[
E (x_{\text{no change}}) = np_{\text{no change}} \\
E (x_{\text{divided}}) = np_{\text{birth}} \\
E (x_{\text{no change}}^2) = np_{\text{no change}} (1 - p_{\text{no change}}) + n^2 p_{\text{no change}}^2 \\
E (x_{\text{divided}}^2) = np_{\text{birth}} (1 - p_{\text{birth}}) + n^2 p_{\text{birth}}^2 \\
E (x_{\text{no change}}x_{\text{divided}}) = n (n - 1) p_{\text{no change}} p_{\text{birth}}
\]

which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.
where in the last line we have used that $E(X_i X_j) = \text{cov}(X_i, X_j) + E(X_i) E(X_j)$. Setting $p_{\text{no\,change}} = 1 - (p_{\text{birth}} + p_{\text{death}})$, we have that

$$\sigma^2(t + \Delta t) = n(t) \left[ p_{\text{birth}}(\Delta t) + p_{\text{death}}(\Delta t) - (p_{\text{birth}}(\Delta t) - p_{\text{death}}(\Delta t))^2 \right]. \quad (A.3)$$

References


