The fixed-size Luria–Delbruck model with a nonzero death rate

Natalia L. Komarova a,*, Lin Wu b,c, Pierre Baldi b,c

a Department of Mathematics, University of California, Irvine, CA 92697, United States
b Department of Computer Science, University of California, Irvine, CA 92697, United States
c Institute for Genomics and Bioinformatics, University of California, Irvine, CA 92697, United States

Received 22 January 2007; received in revised form 6 April 2007; accepted 20 April 2007
Available online 3 May 2007

Abstract

What is the expected number of mutants in a stochastically growing colony once it reaches a given size, N? This is a variant of the famous Luria–Delbruck model which studies the distribution of mutants after a given time-lapse. Instead of fixing the time-lapse, we assume that the colony size is a measurable quantity, which is the case in many in-vivo oncological and other applications. We study the mean number of mutants for an arbitrary cell death rate, and give partial results for the variance. For a restricted set of parameters we provide analytical results; we also design a very efficient computational method to calculate the mean, which works for most of the parameter values, and any colony size, no matter how large. We find that a cellular population with a higher death rate will contain a larger number of mutants than a population of equal size with a smaller death rate. Also, a very large population will contain a larger percentage of mutants; that is, irreversible mutations act like a force of selection, even though here the mutants are assumed to have no selective advantage. Finally, we investigate the applicability of the traditional, ‘fixed-time’ approach and find that it approximates the ‘fixed-size’ problem whenever stochastic effects are negligible.

© 2007 Elsevier Inc. All rights reserved.

Keywords: Stochastic modeling; Cancer; Drug resistance; Mutations; Partial differential equations; Diffusion approximation; Boundary layer

* Corresponding author. Tel.: +1 949 824 1268; fax: +1 949 824 7993.
E-mail address: komarova@math.uci.edu (N.L. Komarova).
1. Introduction

The fluctuation analysis of mutations in bacterial cultures was first developed by Luria and Delbruck in 1943 [1]. Since then the distribution of the number of mutants in growing populations of cells has been studied by many authors (see the review by Zheng [2] and references therein). Even though the fluctuation analysis of Luria and Delbruck was originally designed for bacterial populations, it has since been widely applied to cancer genetics [3–5]. There are many formulations of the problem, which emphasize different aspects of the mutation process. In particular, the following formulation has been extensively studied [6–11]. Suppose a colony of cells starts growing at time \( t = 0 \). A cell division may result in a mutation, whereby a mutant daughter cell is created. There is only one type of mutants, and mutant cells divide faithfully. What is the probability distribution of the number of mutants after a given time, \( t \)? Most authors considered the problem with a zero death rate. In their recent paper, Dewanji et al. [12] found the probability distribution of the number of mutants for the problem with a general, nonzero death rate. Similar methods were also used by Iwasa et al. [13].

Even though the above formulation has received much attention, the following problem appears to be more relevant in many clinical situations. Suppose a colony of cells initially contains a known number of wild-type cells and mutants, and the population grows stochastically. When (if) the colony reaches a given size, \( N \), the number of mutants is counted. What is the expected number of mutants, given that the colony indeed grows to size \( N \)? What is the variance? We will refer to this formulation as the ‘fixed-size problem’, in order to distinguish it from the standard ‘fixed-time’ problem. The only study we are aware of that pertains to the fixed-size problem is the paper by Angerer [14]. There, a closed-form expression for the probability distribution of the number mutants in a colony of size \( N \) was found, under the restrictive assumption of the zero death rate of cells.

The relevance of the fixed-size formulation is apparent if we consider, for instance, the generation of resistant mutants by cancerous cells. The question of interest is to determine the probability of resistance in a tumor at the start of treatment. In clinical situations, it is the size of the tumor that is a known quantity, and not its age. Therefore the fixed-size formulation must be used to model this process. Other examples come from studying cancer generation and progression through a sequence of selection barriers.

In the present paper we solve the fixed-size problem for the expected number of mutants in the general case of nonzero death rates, and obtain partial results for the variance. A comparison of the results for the fixed-size and the fixed-time formulations shows clearly that the two models can generate very different results. Therefore, one has to be careful when approximating the results in a fixed-size situation with a fixed-time calculation.

We find that the fixed-time problem is a ‘diffusion-free’ approximation of the fixed-size problem. This approximation works well for small death rates and for large initial numbers of cells. If the initial population of cells is small, or the death rate is large, then stochastic effects are strong, and the fixed-time approximation fails.

Several new methods are developed in this paper to solve the fixed-size problem. In particular, we find a special, ‘critical’ value of the death rate, \( d \), which allows for an exact analytical solution. The critical value is defined by \( d = l(1 - u) \), where \( l \) is the division rate and \( u \) is the mutation rate.
For other parameter regimes, as long as the death rate is not too high ($|l - d| \gg 1/N$), we design a very efficient ‘boundary layer method’ of solution. When this inequality fails, we describe other methods of solution.

We show that the number of mutants depends on: (a) the magnitude of the product, $uN$, of mutation rate, $u$, and the population size, $N$, and (b) the value of the death rate, $d$, with respect to the division rate, $l$. Starting from one wild-type cell, the average fraction of mutants at size $N$ is given roughly by $Nulog N$ for $d = 0$. The fraction of mutants increases monotonically with $d$. As $d$ approaches $l$, the following behavior is observed. In the small mutation rate limit ($uN \ll 1$), the fraction of mutants approaches $Nu/2$ as $d \to l$. More precisely, it reaches this magnitude for $l - d \ll 1/N$. In the large population size limit ($uN \gg 1$), the fraction of mutants approaches $1 - 1/(Nu)$ for $d = l(1 - u)$, and it further increases to $1$ as $d \to l$.

In short, the cellular population with a higher death rate will contain a larger number of mutants, everything else being equal. Also, a larger population will contain a larger percentage of mutants.

The paper is organized as follows. In Section 2 we present some biological background to motivate our analysis. Then we formulate both fixed-time and fixed-size problems mathematically, and show that the two formulations lead to different results for the number of mutants. In Section 3 we derive the equations for the moments, and introduce a continuous (PDE) approximation and the diffusion-free approximation for the discrete fixed-size problem. Also, we present an analytical solution of the problem for the critical regime. In Section 4 we present complete results for the mean number of mutants in all regimes. In particular, we derive a very efficient boundary layer method which is in an excellent agreement with stochastic simulations. Section 5 is reserved for a summary and discussion. The bulk of the paper is devoted to finding the mean of the number of mutants. Details of the calculations, as well as results for the variance are given in the Appendix.

2. Fixed-time and fixed-size problems

2.1. Biological motivation

In this section we present two oncological examples to motivate the analysis of the ‘fixed-size’ problem.

2.1.1. Drug resistance in cancer

Drug resistance in cancer is one of the main problems in cancer treatment. There are many reasons why cancer may become resistant to therapy, one of them being the alteration of a target molecule [15,16]. A typical example is the new and very effective generation of cancer drugs, the so-called small molecule inhibitors, which work by binding to a particular target protein found only in cancer cells, and killing the cancer cells. If the target protein is altered, then the drug becomes ineffective.

The target protein may be altered by mutations, which are very common in cancers. While the slightly altered version of the gene may still function in the cell, it can no longer be inhibited by that particular drug. An example of this process is drug resistance in Chronic Myeloid Leukemia
(CML) patients treated with the kinase inhibitor Imatinib [17,18]. Recent research has identified the specific mutations in the target gene that render the protein resistant to the drug [19–21]. The quantity we would like to estimate ultimately is the probability of drug resistance in patients at the time of detection, or, more generally, at the time when treatment begins. A related mathematical problem is finding the probability distribution of resistant mutants in a colony of a given size. The main motivation is the fact that cancer remains undetected until it grows above a certain threshold size, where, for instance, symptoms may become apparent. Also, the proper ‘measurable’ quantity in clinical situations at the time of the start of treatment is the colony size, and not the tumor age (this is unknown). In in vitro and in vivo experiments the time since the start of growth could be a known quantity, and there the ‘fixed-time’ formulation of the problem may be appropriate. In cancer therapy however we must solve a ‘fixed-size’ problem.

2.1.2. Tumor progression

A second example is not concerned with cancer treatment per se, but rather with cancer progression, diagnostics and predicting the future course of the disease. Under the theory of multi-stage carcinogenesis, a tumor grows as a sequence of one or more waves of clonal expansion interspersed with selection barriers, where the growth stops until a further molecular transformation can take the growth process to its next level [22,23]. For instance, early pre-angiogenic tumors can only grow up to a size of about 2 mm [24]. To grow beyond this size, a new mechanism of de-novo vessel formation, or angiogenesis, must be involved [25]. Because of the existence of a sequence of selective barriers, many tumors correspond to dead ends [26–28]. They may grow to a certain size but never go beyond that and thus remain benign, or at least controllable. Other tumors overcome selection barriers by mutations and may become malignant.

The ability for continuing growth is defined by the presence of certain mutants in the tumor. For instance, the presence of angiogenic mutations in an early lesion will make tumor progression possible beyond the 2 mm threshold [25]. Likewise, the presence of metastatic mutations in an early tumor will lead to metastases later on in the disease progression [29]. It is fair to say that the future of the tumor is largely defined by its present. More precisely, the composition of the tumor colony, and in particular, the presence of certain malignant mutants, determines the course of the disease later on. Therefore, a model which finds the probability distribution of mutants in a tumor of a given size, would be extremely helpful for tumor prognoses.

2.1.3. Mutations

The conventional assumption for mutations is as follows: a cell reproduces, and with a small probability one of the daughter cells contains a mutation, while the other daughter cell does not [30–35]. Whether this is a well-justified model depends on a particular kind of mutations being described.

All mutations can be classified roughly into two types [36,37]: small scale mutations (such as point mutations, insertions and deletions), and large-scale mutations (such as gene amplifications or large-scale deletions, chromosome gains or losses, translocations, inversions, and recombinations). If we are interested in the statistics of a particular small-scale mutation, then a good model for this is a replication event whereby only one of the daughter cells contains a mutation. This is because a mutation of this kind is a random copying event which happens during DNA synthesis [38], and the probability that the same error occurs in copying both strands in the replication fork...
is extremely small (and can be neglected). The same type of reasoning also applies to some large-scale mutational events such as gene amplifications.

When we talk about other large-scale mutational events, an error in one of the daughter cells could be coupled with an error in the other. For example, a chromosome loss in one daughter cell can be accompanied by a chromosome gain in the other daughter cell. In such cases, both daughter cells will be mutant. However, if we are only keeping track of a particular type of mutations (e.g. losses of a given chromosome) then of the two daughter cells only one will contain the alteration of interest.

With some other types of mutations, hypothetically we cannot exclude the possibility that both daughter cells contain the same mutation as a result of one cell division of a wild-type cell. Also, a genetic alteration can be created independently of a cell division, as a result of chemicals, radiation and other stresses. In this paper we do not consider such scenarios.

2.1.4. The fitness of mutant cells

In both biological examples described above, the mutant cells of interest may in principle be advantageous or disadvantageous with respect to the ‘wild-type’ tumor cells. However, a good null-assumption is that the mutants are neutral. For example, the degree of binding of a certain protein to the drug Imatinib is thought to be irrelevant to colony growth before the treatment is applied.\textsuperscript{1} Similarly, the genes crucial for the metastases (or angiogenesis) at late stages of tumor progression may be contained in an early lesion, where they have no specific function. Such mutants can be considered neutral.

Therefore, in the present paper we concentrate on studying neutral mutations. Disadvantageous and advantageous mutations will be studied elsewhere.

2.2. Formulation of the process

Let us suppose that cells can divide faithfully, divide with a mutation, or die. A faithful division of a cell results in two daughter cells identical to the mother cell. A division with a mutation results in one cell identical to the mother cell and one mutant cell. In principle one can study a process where there is more than one type of mutation. In this case a helpful representation is a selection–mutation diagram, which specifies for all cell types, what types they can transform into by single mutations [39].

In this paper we restrict ourselves to the simplest possible case, where only one type of mutation can occur: \( A \to B \), where ‘A’ is the wild type, and ‘B’ is the mutant type. Moreover, we will concentrate on the problem where the fitness of type ‘B’ is the same as that of type ‘A’.

The population of cells can change by means of a cell reproduction or a cell death. We assume that only one such event can happen at a time. We introduce three parameters, \( l \), \( d \), and \( u \), such that \( 0 \leq l, d, u \leq 1 \), and \( l + d = 1 \). Let us pick a cell at random from the colony (independent of its type). This cell can either reproduce with probability \( l \), or die with probability \( d \). If the chosen cell

\textsuperscript{1} More precisely, two different types of resistance against Imatinib have been studied extensively in the literature. One type of resistant mutants, which contains multiple copies of the oncogene, is thought to be negatively selected in the absence of the drug. However, this type of resistance is by far the less common, compared to resistance by a point mutation in the oncogene. The more common point mutations are thought to be neutral.
is of type ‘A’, and it reproduces, then the reproduction is faithful with probability 1 − u. With probability u, it results in one type ‘A’ and one type ‘B’ cell. Type ‘B’ always reproduces faithfully. We call l the reproduction rate, d the death rate, and u the mutation rate. Rates l and d are the same for both types of cell.

So far nothing is known about the timing of the process. There are multiple ways of modeling this, for instance, the standard continuous-time birth–death process, where the timing of changes in the system follows an exponential distribution. Let us pick two non-negative constants L and D, such that

\[ l = \frac{L}{L + D}, \quad d = \frac{D}{L + D}. \]

The probability for one cell to divide within an infinitesimal time-interval, \( \Delta t \), can be defined as \( L \Delta t \), and the probability to die as \( D \Delta t \). The quantity \( L + D \) determines the time-scale of the cellular dynamics. Another way to handle the timing of events in the discrete-time stochastic process is described in Section 2.4.

Ultimately we are interested in studying the probability distribution and the moments of the number of mutants in such a colony of cells. In the next sections we compare the two formulations of this problem. One studies the number of mutations at a given time elapsed since the beginning of the growth. The other looks at the number of mutants conditioned on the total colony size.

2.3. The fixed-time problem

Here we review the results pertaining to the more traditional model, which we term the fixed-time problem. This problem is formulated by asking the following question: starting from a given initial condition, what is the probability distribution of mutations in a colony of dividing cells, after a certain time, t? In Ref. [12], a very general model was proposed which allows to calculate the probability distribution and moments of the number of mutants, \( Y(t) \), in a deterministically growing colony, at a given time, t. The production of mutants is described by a doubly-stochastic process. Let us assume that the birth and death rates of mutants are constant coefficients \( L \) and \( D \), and the normal cells grow exponentially according to \( M_0 e^{ct} \), with \( c = \frac{L}{1 - C_0 u} \). Following Ref. [12], we denote the probability to have \( i \) mutants at time \( t \) as \( P_i(t) \). We have the following recursive formula:

\[
P_0(t) = M_0 \exp \left\{ -Lu \int_0^t \frac{L - D}{L - De^{(L-D)(s-t)}} \, ds \right\},
\]

\[
P_m(t) = \sum_{i=0}^{m-1} \frac{m-i}{m} P_i(t) p_{m-i}(t),
\]

where \( p_k(t) \) is given by

\[
p_k(t) = M_0 (L - D)^2 u \int_0^t \left( \frac{L(1 - e^{(L-D)(s-t)})}{L - De^{(L-D)(s-t)}} \right)^k \frac{e^{cs} \, ds}{Le^{(L-D)(t-s)} + De^{(L-D)(s-t)} - (L + D)}.\]

This is an iterative formula for calculating the probability distribution of the number of mutants. The formula for the mean number of mutants is as follows:

\[
E(Y) = M_0 e^{(L-D)t} (1 - e^{-Lu}).
\]
2.4. The fixed-size problem

Now let us formulate the fixed-size problem. We start from a fixed number of cells of fixed types, let the colony grow stochastically and stop the process when (if) the colony reaches the size \( N \), where \( N \) is a fixed number. The objective in the fixed-size problem is to calculate the resulting probability distribution of the number of mutant cells.

There are two main differences in the formulations if the fixed-size and the fixed-time models. (1) The stopping condition: the mutants are counted at a certain point of time in the fixed-time model, and at a certain colony size in the fixed-size model, and (2) The growth of the wild-type cells is deterministic in the fixed-time model and it is stochastic in the fixed-size model. The second difference is not a necessity, and alternatives are discussed in Section 5.2. The first difference has important consequences which are studied in this paper.

In the fixed-size model we are interested in the mutant distribution in a colony of a given size, and not in the time-evolution of the process. In this context, the exact timing of the colony’s growth is unimportant. For instance, we can use the following very simple description. Envisage a discrete-time stochastic process, where at each moment of time a randomly chosen cell undergoes one of the three processes: faithful reproduction, reproduction with a mutation, or death. This choice of time-evolution corresponds to adopting a ‘nonlinear’ clock which advances one time-step only if an event occurs in the system. For as long as no processes occur, the clock stands still. In such a model no assumptions are made about the real, ‘physical’ timing of the process, except that only one event can occur at a time.

Note that while the continuous-time model of Section 2.2 provides a more biologically realistic description of the timing of the process, both the discrete-time and the continuous-time processes will result in exactly the same probability distribution when the colony reaches a given size, \( N \). Since our objective is to study just this distribution, here we adopt the simpler, discrete-time formulation.

Let us denote by \( i \) the number of wild-type cells and by \( j \) the number of mutants. The dynamics of the system can be envisaged as a two-dimensional Markov walk on the simplex, \( i + j \leq N, i, j \geq 0 \), see Fig. 1. We assume that all states \((i, j)\) such that \( i + j = N \) are absorbing (see Section 2.6 for the validity of this assumption). This is equivalent to stopping the process as soon as the colony reaches size \( N \). The transition probabilities can be characterized as up-, down-, left- and right-probabilities, \( p(i, j \rightarrow i + 1, j) = P^+_{ij} \), \( p(i, j \rightarrow i - 1, j) = P^-_{ij} \), \( p(i, j \rightarrow i, j + 1) = P_{ij}^\uparrow \) and \( p(i, j \rightarrow i, j - 1) = P_{ij}^\downarrow \), Fig. 1. These are given by the following:

\[
\begin{align*}
    p^\uparrow_{ij} &= \frac{li}{i+j}, \quad p^\downarrow_{ij} = \frac{di}{i+j}, \quad p^-_{ij} = \frac{l(iu+j)}{i+j}, \quad p^+_{ij} = \frac{dj}{i+j}
\end{align*}
\]

for \( i, j > 0, i + j < N \). The rest of the elements in the transition matrix are identically zero. Obviously, the state \((0,0)\) is absorbing. Also, we choose the normalization of probabilities such that \( P^\uparrow_{ij} + P^\downarrow_{ij} + P^\downarrow_{ij} + P^\downarrow_{ij} = 1 \). This means that the clock only shifts one unit if a birth or a death occurs.

We wrote a computer code which simulates a random walk on a grid with location-dependent transition probabilities, Eq. (4). At time \( t = 0 \), the simulation starts with 1 wild-type cell and 0 cancer cells, \((i, j) = (1,0)\). At each time \( t \), a random number between 0 and 1 is generated. Eq. (4) are used to divide the interval \([0,1]\) into 4 sub-intervals. We then check which of the 4
sub-intervals the random number falls into; this defines the direction of the jump. The simulation is stopped when either the colony is extinct (state (0,0)) or it reaches the population size \( N \) (state \((i, j)\) with \(i + j = N\)). The number of mutants is then recorded, and the simulation is repeated.

2.5. Comparison between the fixed-time and the fixed-size models

It seems tempting to use the result for the fixed-time problem, formula (3), to estimate the expected number of mutants in a colony of a given size. Starting from \( M_0 \) wild-type cells, the total number of cells at time \( t \) is a random variable, consisting of the number of wild-type cells, \( X(t) = M_0e^{L(1 - u) - Du} \), and the number of mutants, \( Y(t) \). If the mutants produced by the colony grew deterministically at the same rate as the wild type cells, then the total number of cells at time \( t \) would be \( N = M_0e^{L - Du} \). Using this relation we obtain \( t = \log(N/M_0)/(L - D) \), the characteristic time when the size of the colony equals \( N \). Let us rewrite the formula for the mean number of mutants, Eq. (3), in terms of the number of cells:

\[
E(Y) = N \left( 1 - \frac{N}{M_0} \right)^{-\frac{L}{D}}.
\]

(5)

This method of approximating the fixed-size problem by the fixed-time problem has been used by many authors, including [40–43,13].

In Fig. 2 we present a comparison of formula (5) (solid lines), with the values obtained by stochastic simulations of the fixed-size problem (squares). The figure shows the results for the mean number of mutants as a function of the death rate. The relative difference is presented in Fig. 3 (triangles). We can see that the fixed-time results for the mean deviate significantly from the fixed-size results, and the difference increases with \( d \) (the relative error may grow beyond 80%). In the rest of the paper we derive numerical and analytical methods for calculating the mean number of mutants in the fixed-size problem.
2.6. An extended formulation of the fixed-size problem

When a tumor is detected, or when the treatment starts, the size of the colony is the measurable quantity. However, the observer does not know whether the given size has been reached by the colony for the first time. With a nonzero death rate, the colony may fluctuate around a given size $N$ for a long time and accumulate more mutations in the process.

Our formulation of the fixed-time problem does not take into account the possibility of a return to size $N$. We impose absorbing boundary conditions at size $N$ and stop the simulation as soon as this size is reached. Therefore, it may happen that our predictions for the number of mutants are an underestimation.

In order to check this, we have numerically solved the following problem. Suppose the absorbing boundary in the fixed-size problem is imposed at the value $i + j = M$, where $M \geq N$. We run the simulation until the colony becomes extinct or gets absorbed at size $M$. Suppose in the course of a simulation, the colony reached size $N$ exactly $n$ times. Then, for each time the colony size was...
N, we can count the number of mutants. These quantities are denoted \( m_i, i = 1, 2, \ldots, n \). The following quantities are of interest:

- The number of mutants at the first encounter, \( m_1 \),
- The mean number of mutants at size \( N \): \( \bar{m} = \sum_{i=1}^{n} m_i / n \), or
- The number of mutants at the last encounter, \( m_n \).

Note that \( m_1 \) is identical to the measured ‘number of mutants at size \( N \)’ in the first, simplest formulation. Also, if \( M = N \), then all the three quantities above are identical to each other.

In the extended formulation of the fixed-size problem we use \( \bar{m} \) and \( m_n \), and calculated the average values of these over many runs. We then compared the results with the simplest formulation (quantity \( m_1 \)), as well as with the results for the fixed-time calculation. In Fig. 3 we plot the quantity

\[
\frac{\langle X \rangle - \langle m_1 \rangle}{\langle m_1 \rangle},
\]

where \( \langle . \rangle \) denotes the average over many runs, and \( X \) is \( \bar{m} \) or \( m_n \), with \( M = 1.5N \) and \( M = 10N \). Also, for comparison we show the relative error of the fixed-time calculation,

\[
\frac{E(Y) - \langle m_1 \rangle}{\langle m_1 \rangle},
\]

where \( E(Y) \) is given by formula (5). The corresponding data are marked by triangles. We can see that even though the average number of mutants for \( M > N \) is slightly bigger than that for \( M = N \), the relative error is still very small, especially compared to the difference between the fixed-time and the fixed-size calculations.

One could argue that even the \( M = 10N \) computation is not quite accurate because, as \( d \to l \), the colony can depart from \( i + j = N \) to sizes much beyond \( 10N \) and still come back. However, such theoretical cases are implausible. If we assume that detection occurs at size \( N \), then it is hardly possible that the tumor has been at size more that \( 10N \) before detection has occurred. The same holds for the prognosis problem: a 10-fold difference in the population size would have been noticed, which would redefine the clinical value of \( N \).

Therefore, in the rest of the paper we will analyze the simplest fixed-size model (with \( M = N \)) and keep in mind that for very large death rates, the obtained number of mutation is a slight underestimation, which is still significantly better than the prediction given by the fixed time model.

3. Analytical methods for the fixed-size problem

3.1. Notations

Let us denote as \( h_{ij}^{\text{mut},k} \) the probability to be absorbed at the state \((N - k, k)\) starting from \((i, j)\), see Fig. 1. The superscript ‘mut’ reminds us that it is the number of mutants we are concerned with. Similarly, \( h_{ij}^{\text{wt},k} \) is the probability to have \( k \) wild-type cells at size \( N \) starting from \((i, j)\).

We will reserve the notation \( E(\text{mut};i, j) \) for the expected number of mutants starting from \( j \) wild-type and \( j \) mutant cells, given that colony reaches size \( N \). In particular, we will use
\[ E(\text{mut}) \equiv E(\text{mut}; 1, 0) \] for the expected number of mutants if the colony starts from one wild-type cell. Similarly, we will use the notation \( E(\text{wt}; i, j) \) or \( E(\text{wt}) \) for the expected number of wild-type cells in a colony of \( N \) cells, starting from \( i \) wild-type and \( j \) mutant cells.

### 3.2. Previous results for the zero death rate

A general formula for the probability distribution \( h_{ij}^{\text{mut},k} \) in the special case of \( d = 0 \) was obtained by Ref. [14]. In particular, the probability to have \( k \) mutant cells in a colony of size \( N \), starting from one wild-type cell, is given by

\[
h_{1,0}^{\text{mut},k} = \sum_{i=1}^{N-k} (-1)^{N-i} \binom{N-k-1}{i-1} \binom{i(1-u) - 1}{N-1}.
\]  

(7)

The following useful quantity was defined [14],

\[
E^c(N) = \sum_{k=1}^{N} k(k+1) \ldots (k+z-1) h_{1,0}^{\text{mut},N-k}.
\]

(8)

The mean number of mutants is given by \( E(\text{mut}) = N - E^1(N) \):

\[
E(\text{mut}) = N - \frac{\Gamma(N+1-u)}{\Gamma(2-u)(N-1)!}.
\]

(9)

Using this expression, we have approximately, for \( N \) large,

\[
E(\text{mut}) \approx N - \frac{N^{1-u}}{\Gamma(2-u)} [1 + O(1/N)].
\]

(10)

For small values of \( u \) we have

\[
E(\text{mut}) \approx Nu(\log N + \gamma - 1)[1 + O(u \log N + 1/N)],
\]

where \( \gamma \) is Euler’s number \( \gamma \approx 0.5772 \). This formula suggests that the proportion of mutants, \( E(\text{mut})/N \), grows with \( N \) logarithmically, and in larger colonies the frequency of mutants is larger. Formula (11) breaks down if \( \log Nu \) is not small. A biologically realistic range corresponding to this regime is when, for instance, \( N \approx 10^{14} \) and \( u \approx 10^{-2} \), if we have genomic instability. For very large values of \( N \) we have from Eq. (10),

\[
\lim_{N \to \infty} E(\text{mut})/N = 1.
\]

### 3.3. Equations for the moments

The method of Ref. [14] does not apply if the death rate is nonzero. We have to come up with a set of different tools. Let us formulate the problem for the mutant probability distribution given the colony size. We have the following recursive equations for the quantity \( h_{ij}^{\text{mut},k} \), the probability to have \( k \) mutants at size \( N \) starting from state \((i, j)\):

\[
h_{ij}^{\text{mut},k} = P_{ji} h_{i+1,j}^{\text{mut},k} + P_{ij} h_{i-1,j}^{\text{mut},k} + P_{j} h_{i,j+1}^{\text{mut},k} + P_{j+1} h_{i,j-1}^{\text{mut},k}, \quad 0 \leq i + j \leq N - 1.
\]

(12)
Inserting the values for the transition probabilities from Eq. (4), we get:

\[(i + j) h_{ij}^{\text{mut},k} = l(1 - u)ih_{i+1,j}^{\text{mut},k} + dih_{i-1,j}^{\text{mut},k} + l(iu + j)h_{i,j+1}^{\text{mut},k} + djh_{i,j-1}^{\text{mut},k}, \quad 0 \leq i + j \leq N - 1. \tag{13}\]

The boundary conditions are

\[h_{00}^{\text{mut},k} = 0, \quad h_{m,N-m}^{\text{mut},k} = \delta_{k,N-m},\]

where \(\delta_{k,N-m}\) is the Kronecker symbol. The total probability to get absorbed in any of the states \((N - k, k)\) is given by \(\sum_{k=0}^{N} h_{ij}^{\text{mut},k}\) which in turn is equal to the probability of non-extinction starting from the total of \(i + j\) cells. Because both types of cells have identical division and death rates, the probability of non-extinction is easily calculated from the classical 1D gambler’s ruin problem, see e.g. [44,45]:

\[P_{\text{non-ext}}^{i+j} = \sum_{k=0}^{N} h_{ij}^{\text{mut},k} = \frac{1 - (d/l)^{i+j}}{1 - (d/l)^{N}}. \tag{14}\]

The probability distribution of the number of mutants, starting from \(i\) wild-type cells and \(j\) mutants, is given by the vector \(\{h_{ij}^{\text{mut},k}\}\), obtained from the solution of a linear system, Eq. (13). The conditional probability distribution of the number of mutants, given that the colony reaches size \(N\) and does not go extinct, is given by \((P_{\text{non-ext}}^{i+j})^{-1}\{h_{ij}^{\text{mut},k}\}\).

Similarly, we can define the probability to have \(k\) wild-type cells at size \(N\), \(h_{ij}^{\text{wt},k}\). This quantity satisfies the same Eq. (13) with different boundary conditions:

\[h_{00}^{\text{wt},k} = 0, \quad h_{m,N-m}^{\text{wt},k} = \delta_{k,m}.\]

Next, let us derive an equation for the expected number of mutants at size \(N\), starting from point \((i, j)\). We introduce the variable \(z_{ij}^{\text{mut}} = \sum_{k=0}^{N} h_{ij,k}^{\text{mut},k}\), the expected number of mutants as the system either reaches size \(N\) or goes extinct. In order to obtain the quantity of interest, the expected number of mutants in a colony of size \(N\), we need to normalize this by the probability of non-extinction. For the expected number of mutants, we have

\[E(\text{mut}; i, j) = \frac{z_{ij}^{\text{mut}}}{P_{\text{non-ext}}^{i+j}}, \quad E(\text{mut}) = \frac{z_{0,0}^{\text{mut}}}{P_{\text{1}}^{\text{non-ext}}}.\]

In order to find the quantities \(z_{ij}^{\text{mut}}\), we multiply Eq. (13) by \(k\) and sum over \(0 \leq k \leq N\). This yields a set of equations for \(z_{ij}\):

\[(i + j)z_{ij}^{\text{mut}} = l(1 - u)iz_{i+1,j}^{\text{mut}} + diz_{i-1,j}^{\text{mut}} + l(iu + j)z_{i,j+1}^{\text{mut}} + djz_{i,j-1}^{\text{mut}}, \quad 0 \leq i + j < N, \tag{15}\]

\[z_{0,j}^{\text{mut}} = N \frac{1 - (d/l)^{j}}{1 - (d/l)^{N}}, \quad z_{N-j,j}^{\text{mut}} = j, \quad 0 \leq j \leq N. \tag{16}\]

The first of the two boundary conditions means that, starting from \(j\) mutant cells and no wild-type cells, the system will either end up at state \((0, 0)\) or, with probability \(P_{\text{non-ext}}^{j}\), it will reach size \(N\), such that there will be exactly \(N\) mutants (and no wild-type cells, as no back mutations are considered). The right hand side of the first boundary condition is thus

\[0 \times (1 - P_{\text{1}}^{\text{non-ext}}) + N \times P_{\text{1}}^{\text{non-ext}}.\]
We will also consider the expected number of wild-type cells as $t \to \infty$, which we denote as $z_{ij}^{\text{wt}}$. Again, the expected number of wild-type cells given that the colony has reached size $N$ is found from

$$E(\text{wt}) = \frac{z_{1,0}^{\text{wt}}}{P_1^{\text{non-ex}}},$$

The quantity $z_{ij}^{\text{wt}}$ satisfies the same Equations as (15), with different boundary conditions:

$$
(i + j)z_{ij}^{\text{wt}} = l(1 - u)iz_{i+1,j}^{\text{wt}} + diz_{i-1,j}^{\text{wt}} + l(iu + j)z_{i,j+1}^{\text{wt}} + djz_{i,j-1}^{\text{wt}}, \quad 0 \leq i + j < N, \quad (17)
$$

$$
z_{0,j}^{\text{wt}} = 0, \quad z_{N-j,j}^{\text{wt}} = N - j, \quad 0 \leq j \leq N. \quad (18)
$$

For simplicity, we will omit the superscript from the variable $z_{ij}$ whenever the context is clear.

The definition of $z_{ij}$ and the derivation of the corresponding equations, generalize in an obvious way to any function of the number of mutants and wild-type cells. The master equation is always the same, and the boundary condition along the line $i + j = N$ reflects the nature of the variable of interest.

Note that even though the equations that our variables satisfy are nothing but a linear sparse system of algebraic equations, the implementation is not easy. The number of equations is given roughly by $N^2/2$, and since realistic colony sizes can be as large as $10^{13}$, a direct numerical solution of system (13) is hard or impossible for large values of $N$.

### 3.4. The partial differential equation approach

Let us consider system (17) and (18). It is possible to obtain a partial differential equation for $z_{ij}$ which in some regimes will approximate the system of interest. We set $x = i/N$, $y = j/N$, expand the system in the Taylor series in terms of the step-size, $1/N$, and keep all the terms up to the second order (a standard procedure described e.g. in [46,47]). The resulting PDE is

$$z_x(P_\perp - P_\perp) + z_y(P_\perp - P_\perp) + \frac{1}{2N} [z_{xx}(P_\perp + P_\perp) + z_{yy}(P_\perp + P_\perp)] = 0.$$

Here, all the rates are functions of $x$ and $y$. Rewritten in terms of the parameters of the system, the equation for the expected number of the wild-type cells becomes:

$$z_x[l(1 - u) - d] + z_y[l(xu + y) - dy] + \frac{1}{2N} (z_{xx}[l(1 - u) + d]$$

$$+ z_{yy}[l(xu + y) + dy]) = 0, \quad 0 < x + y < 1, \quad (19)$$

$$z^{\text{wt}}(x, 1-x) = Nx, \quad 0 \leq x \leq 1, \quad z^{\text{wt}}(0,0) = 0. \quad (20)$$

Solutions of Eq. (19) can be viewed as stationary solutions of an advection–diffusion equation. We will sometimes refer to the second derivative terms as diffusion terms, even though there is no time-dependence in our system. The first order terms can be seen as advection terms.

Eq. (19) is a continuous approximation to the original, discrete system. This approach is reminiscent of the commonly used diffusion approximation (also called the Fokker–Planck equation) to Kolmogorov forward and backward equations. The applicability conditions of
Eq. (19) can be written using the arguments of Refs. [46] or [47] behind the derivation of the Fokker–Planck equation. In our derivation of Eq. (19), we tacitly assumed that the dependence of the function $z_{ij} \equiv z(x, y)$ on its variables is sufficiently slow, that is, the change in the interval of length $1/N$ is small. However, this assumption does not always hold. Let us set $d = 0$ and consider the behavior near the origin $(0, 0)$. We have $z_{00}^{ul} = 0$ and the quantity $z_{1,0}^{ul} \approx N(1 - u \log N) \sim N$, formula (11). Therefore, we have a jump of size $N$ between the value of the function at points $x = 0$ and $x = 1/N$. The gradient becomes a lot smaller, of the order of $1$ over one step, as we move away from the origin. Also, we expect that for larger values of the death rate, the gradient at the origin becomes less steep. This is because as $d$ grows, the probability of extinction also increases which lowers the expectation $z_{ij}$ for low values of $i + j$, the initial number of cells. More precisely, the gradient around zero becomes $\sim |l - d|/l$, which is small when $d$ is close to $l$.

In order to illustrate these points, we plotted the numerical solution\(^2\) of PDE (19) (Fig. 4, solid lines) against the solution of the original discrete problem (15)–(18) (squares connected by dashed lines). We use the parameters $N = 1000$ and $u = 10^{-3}$. The functions $z_{00}^{mut}$ and $z_{00}^{nt}$ are plotted against the initial number of wild type cells, $i$. In this particular case, for $d = 0$ we have $z_{1,0}^{nt} \approx 993.5$, leading to a large gradient, Fig. 4(b). For smaller $|l - d|$ (cases (c) and (d), with $d = 0.4$), the gradient is smaller, and agreement between the continuous and discrete solution is a lot better than for $d = 0$ (cases (a) and (b)). An explicit treatment of a simpler, 1D case is presented in Appendix A. There, we can find analytical solutions for both the original discrete problem and the differential equation, and estimate the error of the approximation as a function of $x$ and $(l - d)$.

These considerations suggest that the applicability of the PDE is nonuniform in terms of the variables $(x, y)$ and parameter $d$. The continuous approximation gives a good agreement if:

- the death rate is high,
  \[ |l - d| \ll 1, \quad (21) \]
  or
- the death rate is low and the initial configuration is outside a boundary layer,
  \[ |l - d| \sim 1 \text{ and } x + y > a, \quad (22) \]

where $a$ is the thickness of the boundary layer, which is characterized by strong gradients of the solution.

Note that the PDE becomes applicable to a reasonable degree of accuracy when the difference $(d - l)$ is small compared to 1, but smallness of $(d - l)$ compared to $1/N$ is not required; in fact it can still be large compared to $1/N$.

\(^2\) The PDE was solved by using the finite difference method in combination with the boundary layer technique, where the solution was found analytically away from the corner $(0, 0)$ by solving the first order PDE (19) with the $1/N$ terms neglected, and then using this as the boundary condition inside a small (of the order $10/N$) domain near $(0, 0)$. A similar boundary layer method is later used for the discrete problem (15) and (16), see Section 4.1.
3.5. The diffusion-free approximation

If we ignore higher derivatives in the PDE obtained in the previous section, we obtain the following first order equation,

\[ z_{xt}x[l(1-u)-d] + z_{yt}[l(xu+y)-dy] = 0, \quad z_{xt}^\text{wt}|_{x+y=1} = Nx. \]  

(23)

This equation is valid for the expected number of wild-type cells; for other cases, the boundary condition is modified in the obvious way. In many cases (see below) the solution of Eq. (23) approximates the true solution away from a boundary layer near \( x = y = 0 \). Eq. (23) can be solved by the method of characteristics. The characteristics are defined parametrically as solutions of the following ODEs:

![Fig. 4. The expected number of wild type cells, \( z_{\text{wt}}^\text{i} \), and expected number of mutants, \( z_{\text{mut}}^\text{i} \), plotted against the initial number of wild type cells, \( i \) (in all cases, the initial number of mutants is zero). The parameters are \( N = 1000, u = 10^{-3} \) and \( d = 0 \) in figures (a) and (b); \( d = 0.4 \) in figures (c) and (d). Each plot contains two graphs: the squares connected by dashed lines represent the solution of the original, discrete problem (15, 16) or (17, 18). Solid lines represent solutions of PDE (19). We can see that the approximation by the PDE works a lot better for higher values of \( d \) ((c) and (d)).](image-url)
\[
\begin{align*}
\dot{x} &= x[l(1-u) - d], \\
\dot{y} &= l(xu + y) - dy.
\end{align*}
\] (24)

The general solution is
\[
\begin{align*}
x(t) &= x_0 e^{(l(1-u) - d)t}, \\
y(t) &= (x_0 + y_0) e^{(l-d)t} - x(t).
\end{align*}
\] (26)

The solution of the PDE at a point \((x, y)\) is given by
\[
\begin{align*}
z_{wt,\text{diff-free}}(x, y) &= \frac{x + y}{x + y} \left( \frac{x + y}{N} \right)^{\frac{l}{l + d}}.
\end{align*}
\] (28)

Translated into the discrete value, \(z_{ij}^{\text{wt}}\), this gives
\[
\begin{align*}
z_{ij}^{\text{wt}} &\approx z_{ij}^{\text{wt, diff-free}} = \frac{i}{i+j} N \left( \frac{N}{i+j} \right)^{\frac{l}{l + d}}.
\end{align*}
\] (29)

The corresponding value for the number of mutants is given by
\[
\begin{align*}
z_{ij}^{\text{mut}} &\approx z_{ij}^{\text{mut, diff-free}} = N - z_{ij}^{\text{wt, diff-free}} = N \left( 1 - \frac{i}{i+j} \left( \frac{N}{i+j} \right)^{\frac{l}{l + d}} \right).
\end{align*}
\]

Notice that setting \(i = M_0\) and \(j = 0\) in the above expression gives exactly the solution of the fixed-time problem, (5). Therefore, the fixed-time problem is a diffusion-free approximation of the fixed-size problem. We will investigate its applicability in Section 4.1. In the next subsection we will identify a special parameter regime where the exact solution for the mean is possible (see also Section D.3 for an exact solution of the \(d = 1/2\) case).

### 3.6. The critical regime: an exact solution

Let us define the pair \((l_*, d_*)\) such that
\[
d_* = l_*(1-u)
\] (30)

(and \(l_* + d_* = 1\) as usual). We have
\[
d_* = \frac{1-u}{2-u}.
\] (31)

We will call the case where \(d = d_*\) the critical regime because it has very special properties, which are outlined in this section. Similarly, \(0 \leq d < d_*\) will be called the subcritical regime, and \(d_* < d \leq 1/2\) the supercritical regime.

In order to see why the point \(d = d_*\) is in some sense critical, let us take a look at the equations for the characteristics of the diffusion-free system, (26) and (27). The slope of the characteristics in the three regimes is presented in Fig. 5. In the subcritical regime, \(d < d_*\), Fig. 5(a), the characteristics have a positive slope and they map any set \(x + y = \alpha\) with \(0 < \alpha < 1\) onto the set \(x + y = 1\). In the supercritical regime, \(d > d_*\), Fig. 5(c), the characteristics have a negative slope and they map any set \(x + y = \alpha\) with \(0 < \alpha < 1\) onto a subset of the set \(x < \alpha, x + y = 1\). Finally, in the critical case where \(d = d_*\), the characteristics are vertical lines, 5(b).
It turns out that in the critical regime, we can find the exact analytical solution for the master equation. To see this, it is more convenient to work with the quantity $z_{wt}^{ij}$, the expected number of wild-type cells as $t \to \infty$. Using Eq. (30), we obtain from Eqs. (17) and (18):

\begin{eqnarray}
2(i + j)z_{ij}^{wt} - i(z_{i+1,j}^{wt} + z_{i-1,j}^{wt}) - j(z_{i,j+1}^{wt} + z_{i,j-1}^{wt}) &=& (i + j) \frac{u}{1-u} (z_{i,j+1}^{wt} - z_{i,j}^{wt}), \\
z_{0j}^{wt} &=& 0, \quad z_{N-j,j}^{wt} = N-j, \quad 0 \leq j \leq N.
\end{eqnarray}

It is easy to check that the solution

$$z_{ij}^{wt} = i$$

satisfies the equation and the boundary conditions. In particular, we have

$$z_{1,0}^{wt} = 1.$$  

In other words, the expected number of wild type cells in the critical regime is one. Setting $d = l(1 - u)$ in Eq. (14), we obtain

$$E^*(wt) = \frac{1 - (1 - u)^N}{u},$$

and

$$E^*(mut) = N - E^*(wt) = \frac{(1 - u)^N - 1 + Nu}{u}. \quad (36)$$

This exact formula can be simplified in the three important regimes:

- Small mutation rate limit: if $Nu \ll 1$, we have
  $$E^*(mut) \approx uN(N - 1)/2.$$

- Large population size limit: if $Nu \gg 1$, we have
  $$E^*(mut) \approx \frac{Nu - 1}{u} = N(1 - 1/(Nu)).$$
Note that the mean number of wild-type cells at size $N$ is given by

$$E^*(wt) = \frac{1}{u}. \quad (37)$$

- Intermediate case: if $Nu = 1$ and $N \to \infty$, we obtain

$$E^*(mut) \approx N/e,$$

where $e \approx 2.7183$ is the base of the natural logarithm.

It is interesting that in the critical regime, we can also find the exact solution of PDE (19). With the assumption $d = l(1 - u)$, Eq. (19) can be written as

$$z_{wt}(x+y) + \frac{1}{Nu} \left( z_{wtx}(1-u) + z_{wtw} \left(y + \frac{u(x-y)}{2}\right)\right) = 0,$$

with the boundary conditions $z_{wt}(x,1-x) = Nx$ and $z_{wt}(0,0) = 0$. The exact solution is

$$z_{wt}(x,y) = Nx \equiv z_{wt,x}(x,y). \quad (38)$$

This is of course a continuous version of the discrete solution obtained in Section 3.6, Eq. (34).

In Section 4 we obtain solutions for super- and subcritical regimes.

4. Complete results for the mean number of mutants

4.1. The boundary layer method

An exact analytical solution of systems (15)–(18) is not possible when $d \neq d_*$ or $d \neq 0$. Here we develop a very efficient approximate method for solving system (17) and (18).

The main idea is to realize that often (and the exact conditions are described below), the solution of system (17) has a boundary layer structure. Namely, away from the point $(0,0)$, the solution is described very well by the diffusion-free approximation (28). In the corner near the point $(0,0)$, there is a boundary layer where the full system must be solved. This is illustrated in Fig. 6. The solution of the discrete system, Eqs. (17) and (18) is plotted together with the diffusion-free solution, (29), for several values of $d$. Note that for clarity, the graphs are shifted with respect to each other in the vertical direction. We can see that the boundary layer, that is, the region where the diffusion approximation does not hold, grows with $d$.

The reason for the boundary layer structure is that away from the origin, the coefficients in front of the higher derivatives in PDE (19) become a lot smaller than those in front of the lower derivatives, and the PDE acquires a singular structure.

Let us define the boundary region by $0 < a \leq x + y$, $x > 0$, $y > 0$. In the region $0 < x + y < a$, $x > 0$, $y > 0$, we will rescale the variables $x$ and $y$ by introducing new coordinates, $X$ and $Y$, such that in the corner region, $0 < X + Y < 1$ we have:

$$X = \frac{x}{a}, \quad Y = \frac{y}{a}.$$
The PDE can be rewritten in terms of the new coordinates:

\[ z_X(P_- - P_+) + z_Y(P_+ - P_-) + \frac{1}{2Na} [z_{XX}(P_- + P_+) + z_{YY}(P_+ + P_-)] = 0. \]  

(39)

The goal of this approach is to make sure that the resulting equation does not have a boundary layer structure, that is, the coefficients in front of the second and first derivatives are of the same order of magnitude. In order to achieve this we need

\[ a \sim \frac{1}{|l-d|N}. \]  

(40)

If we can do this with \( a \ll 1 \), then the second derivatives only matter in a small (of size comparable to \( a \)) corner of the domain. Outside this corner we have an analytical diffusion-free solution. The full PDE only needs to be solved inside the boundary layer. The boundary condition is given by

\[ z_{\text{diff}}|_{x+y=1} = \gamma N \left( \frac{1}{a} \right)^{-\frac{\gamma}{2d}}. \]

It is easy to see that the boundary layer method works as long as

\[ |l - d| \gg 1/N. \]

Depending on the applicability of the PDE (Appendix B), there are two cases:

**Lower death rates**, \(|l - d| \sim 1\) (in practice, \(|l - d| > 0.01\)). In this case, the PDE is a good approximation outside the boundary layer, but it is not applicable within the boundary layer. However, the good news is that the boundary layer in this case is extremely thin, and the number of states it contains is relatively small (it is approximately \( \sim \left( \frac{1}{l-d} \right)^2 \)). Therefore, instead of solving the PDE, we can actually solve the original, discrete system of linear algebraic equations, Eq. (17), inside the boundary layer. To truncate the system of algebraic equations, we use the diffusion-free

Fig. 6. The boundary layer structure of the solution. Numerical solutions of system (17, 18), \( z_{\text{diff}} \) (connected circles), are plotted together with the diffusion-free approximation, solid lines, for different values of \( d \), for \( N = 100 \) and \( u = 1/N \). The graphs are shifted vertically with respect to each other.
solution to define the boundary condition at the edge of the boundary layer. This system can be easily solved numerically because of the small number of equations.

Higher death rates, $|l - d| \ll 1$. In this case the same approach holds in theory, but the number of equations inside the boundary layer becomes very large and the system is difficult to solve. On the other hand, in this case the discrete solution is well approximated by the solution of the PDE, even inside the boundary layer. Thus, one could use the numerical solution of the PDE inside the boundary layer, with the boundary conditions given by the diffusion-free approximation. Note that since the singular structure of the initial PDE does not appear in the scaled version, Eq. (39), the step-size for the numerical solution of the PDE can be taken to be much larger than $1/N$, thus reducing the number of (algebraic) equations that need to be solved.

4.2. Comparison with the results of stochastic simulations

Here we discuss several examples of application of the boundary layer method, and compare the corresponding approximate solution with the solution obtained by stochastic numerical simulations. In Fig. 7 we present three cases: (a) $uN \ll 1$, (b) $uN = 1$ and (c) $uN \gg 1$. Each figure shows plots for three cases, $N = 10^4$, $N = 10^5$ and $N = 10^6$. The conditional expectation of the number of mutants, $E(mut)$, scaled by $N$, is plotted against $d$. The discrete points represent stochastic simulations, and the solid lines correspond to our calculations. The boundary-layer method is applied for values of $d$ between 0 and 0.49. We can see excellent agreement between the stochastic simulations and the boundary-layer method of solution.

Note that the time-complexity of the stochastic code grows with the population size $N$ and also it is inversely proportional to the distance between $l$ and $d$, because $N/(l - d)$ is an average number of steps it takes to grow from 1 cell to $N$ cells. The average time complexity of the code is given by $O(NM/(l - d))$, where $M$ is the number of independent simulations. The most extreme values of our simulations are $N = 10^6$, $M = 5 \times 10^4$, $d = 0.495$, the time complexity in this case is $5 \times 10^{12}$. This estimates the average number of random numbers generated during the simulation. The random generator used in our experiments is gsl_rnd_mt19937 in the GNU Scientific Library.

Compared to the method of stochastic simulations, the boundary layer method described above is considerably less computationally intensive. From formula (40) it is clear that if $|l - d| \gg 1/N$, the resulting boundary layer is quite thin. For instance, if $d = 0.49$, we only need to solve a system of algebraic equations, system (17) and (18), with $0 \leq i, j \leq 99$. This is equivalent to inverting a matrix of size $5000 \times 5000$. This can be done in a matter of seconds. An equivalent full stochastic simulation with $N = 10^6$ takes 100 h to run on desktop-class machine (3 GHz Xeon processor).

The boundary layer method is also clearly superior to a direct solution of the discrete system (17) and (18). The boundary layer method, as pointed out before, requires an inversion of a $5000 \times 5000$ matrix. The discrete system requires an inversion of an $N^2/2 \times N^2/2$ matrix. Computationally, the two methods are the same for $N \approx 100$, but as $N$ grows, solving the discrete system becomes more expensive. Biologically relevant values of $N$ are $10^{10}$ and larger, which makes the direct solution of system (17) and (18) impossible.

Another advantage of the boundary-layer method is that it is independent of $N$. Even though the width of the boundary layer, Eq. (40), depends on $N$, the number of nodes inside the boundary layer is only defined by $l - d$. Therefore, we can get results for any population sizes.
A disadvantage of the boundary-layer method is that it breaks down for the values of $d$ such that $|l - d| \sim 1/N$; there, other methods must be used. This is described in Section 4.3.

4.3. The solution structure in the sub- and supercritical regimes

The boundary layer method works well if $|l - d| \gg 1/N$. On the other hand, we have the exact solution for $d = d_*$. Finally, certain analytical methods apply if $d$ is very close to $l$, see Appendix C.2. The detailed analysis of all parameter regimes is presented in Appendices B and C. Here, for simplicity of exposition, we summarize the results.

Fig. 7. The boundary-layer method of calculating the average number of mutants, compared to the results of stochastic simulations. The diamonds, triangles and squares, represent stochastic simulations for $N = 10^4$, $N = 10^5$ and $N = 10^6$, respectively. The boundary-layer calculations are shown by solid lines.
The first thing we emphasize is that the solution behaves qualitatively differently depending both on the death rate and on the product $uN$. The summary of all the regimes is presented in Fig. 8. We will distinguish the small mutation rate limit, $uN \ll 1$, Fig. 8(a), and the large population size limit, $uN \gg 1$, Fig. 8(b).

**Small mutation rate limit.** If $uN \ll 1$, Fig. 8(a), we have three distinct subcritical regimes. It is convenient to define

$$d = d_\ast - \epsilon,$$

where the critical $d_\ast$ satisfies $d_\ast = l_\ast (1 - u) = \frac{1-u}{2-u}$. We distinguish the following subcases:

- **Strongly subcritical regime**, $\epsilon > 0$, $\frac{\epsilon}{\epsilon \gg 1/N}$. The solution has a boundary layer structure, such that away from the origin, $(0,0)$, it is well approximated by the diffusion-free solution, Eq. (28). Inside the boundary layer, the solution can be easily found. If $|l - d| \sim 1$, we need to solve the original discrete system (in this case the boundary layer is very thin, and this computation is very easy). If $1/N \ll |l - d| \ll 1$, we can approximate the solution inside the boundary layer with the solution of PDE (19). The applicability of the PDE approach is shown in Fig. 8 (c). This regime is denoted by (A) in Fig. 8.

- **Nearly critical regime**, $\epsilon \ll 1/N$. The solution can be well approximated by the critical solution, Eq. (38). This regime is denoted by (C) in Fig. 8.

- **Weakly subcritical regime**, $\epsilon > 0$, $\frac{\epsilon}{\epsilon \sim 1/N}$. In this intermediate regime, diffusion-free and critical solutions are not good approximations. The system can be approximated by PDE (19) and solved numerically. This regime is denoted by (B) in Fig. 8.

**Large population sizes:** $uN \gg 1$, $d = 1/2$. The solution is close to the large-population, high-death-rate solution, Eq. (52), away from the boundary $x + y = 1$.

![Fig. 8. A schematic map of all the regimes. The horizontal axes denotes the death rate, $d$. (a) Small mutation rates: the proximity of $d$ to $d_\ast$ is measured by $\epsilon$. (b) Large population sizes, the proximity of $d$ to $1/2$ is measured by $\delta$. (c) The applicability of the PDE approximation. Solution types are denoted by letters in circles. (A) the solution has a boundary-layer structure and it is well described by the diffusion-free solution, Eq. (28), away from $(0,0)$; (B) the solution can be obtained by a numerical solution of PDE (19); no other simplifications have been found; (C) the solution is close to critical, Eq. (38); (D) the solution is close to the large-population, high-death-rate solution, Eq. (52), away from the boundary $x + y = 1.$](image-url)
In the supercritical regime, \( \epsilon > 0 \), the behavior is the same as in the nearly critical regime, that is, for both positive and negative values of \( \epsilon \), as long as \( \epsilon \ll 1/N \), the solution is close to critical.

4.3.1. Large population size limit

If \( uN \gg 1 \), Fig. 8(b), there is only one subcritical regime. For all positive \( \epsilon \), i.e. for all \( d < d_* \), the solution has a boundary-layer structure and it is well described by the diffusion-free solution outside of a boundary layer. As \( d \) approaches \( d_* \) from below, the solution gradually gets closer and closer to the critical one; regime (A) in Fig. 8. For negative values of \( \epsilon \), we have three distinct supercritical regimes. To describe them, it is convenient to define

\[
d = 1/2 - \delta.
\]

We distinguish the following cases:

- Nearly critical regime, \( \delta \gg 1/N \). The solution is very well described by the diffusion-free solution, Eq. (28). The term ‘nearly-critical’ is slightly misleading because unless \( |d - d_*| \ll 1/N \), the solution is not close to the critical one.
- Strongly-supercritical regime, \( \delta \ll 1/N \). This case is denoted by (D) in Fig. 8. Now, the solution has a boundary layer structure, but it is very different from structure (A). Namely, away from the boundary \( x + y = 1 \), the solution is very well described by the large-population, high-death-rate solution, Eq. (52) in Appendix C.3. Inside the boundary layer, the PDE has to be solved numerically.
- Weakly-supercritical regime, \( \delta \sim 1/N \). In this intermediate regime, diffusion-free and critical solutions are not good approximations. The system can be approximated by PDE (19) and solved numerically ((B) in Fig. 8).

Note that regardless of the value of \( uN \), a boundary-layer structure is observed as long as \( |l - d| \gg 1/N \), and the differential equation is a good approximation as long as \( |l - d| \ll 1 \).

5. Discussion

5.1. Summary

We have studied the Luria–Delbruck distribution of the number of mutants in a stochastically growing colony of mutating cells, for a fixed colony size, \( N \). The cells are characterized by a growth rate, \( l \), and a death rate, \( 0 \leq d \leq l \). The mutation rate is denoted by \( u \). The probability distribution is described by a system of linear algebraic equations, (13), of size \( N(N - 1)/2 \). We have derived equations for the moments, (15)–(18), and concentrated on solving the system for the first moments.

So far, only the case \( d = 0 \) has been solved analytically by [14]. We have obtained new analytical results for the expected number of mutants in the following special cases: (i) the critical regime: \( d = d_* = l_*(1 - u) \) and (ii) when the death rate is equal to the reproduction rate: \( d = 1/2 \). We have also analyzed the behavior of the system for general death rates, \( d \), by using a boundary-layer approximation and a continuous approximation. We have constructed a phase diagram for the
behavior of the system in the whole interval $0 \leq d \leq 1/2$. In particular, for $|l - d| \gg 1/N$ we have designed a very efficient boundary layer method for calculating the means. All the results are in excellent agreement with stochastic numerical simulations.

5.2. The fixed-size problem versus the fixed-time problem

We have seen that the fixed-time problem (equation (5) or (28)) can be a good approximation to the fixed-size problem in certain parameter regimes, but it fails for other regimes. The diffusion-free formula works well if $d$ is not too close to $l$, and the initial number of cells is large. The intuition behind this is clear. If the death rate is small or zero, then the probability for the colony to die out is small, and with a high probability it will grow in a roughly-deterministic fashion to large numbers of cells. The same holds for systems which start growing from a large initial number of cells, which corresponds to the solution of the equations away from the origin. The larger is the initial number of cells, the smaller is the role of stochastic effects. The fixed-time approximation breaks down when stochastic effects are strong, namely, for $|l - d| \sim 1/N$, and for $d \ll l$, $M_0 \sim 1$, where $M_0$ is the initial number of cells.

In this paper we find that the fixed-time problem is a ‘diffusion-free’ approximation of the fixed-size problem. The results for the mean number of mutants in the fixed-time formulation can be obtained from the fixed-size problem by (1) forming a continuous approximation to the equations for the mean number of mutants and (2) truncating it at the first order of the $1/N$ expansion, where $N$ is the size of the colony. Here we discuss some further insights connected with the diffusion-free approximation.

Let us formulate a continuous time stochastic process with the number of wild type cells, $i$, and the number of mutants, $j$, subject to the following transition probabilities:

$$
P_{(i,j)\rightarrow(i+1,j)} = L(1-u)i \Delta t, \quad P_{(i,j)\rightarrow(i-1,j)} = Di \Delta t, \quad P_{(i,j)\rightarrow(i,j+1)} = L(j+ui) \Delta t, \quad P_{(i,j)\rightarrow(i,j-1)} = Dj \Delta t,
$$

for an infinitesimal time-interval, $\Delta t$. From the Kolmogorov forward equation we can derive the following equations for the average number of wild type cells, $\langle x \rangle$, and the average number of mutants, $\langle y \rangle$:  

$$
\dot{x} = L(1-u)\langle x \rangle - D\langle x \rangle, \\
\dot{y} = (L-D)\langle y \rangle + Lu\langle x \rangle.
$$

These equations are identical to the equations for the characteristics obtained in (24) and (25). The solution, Eq. (28), is identical to that obtained from the fixed-time system, formula (5). The reason for the differences in the behavior of the exact solution and the diffusion-free solution is transparent from the present interpretation of the characteristic equations. The quantity $\langle x \rangle/(\langle x \rangle + \langle y \rangle)$ is different from $z_{1,0}^{mut}/N$: the former quantity is the fraction of wild type cells in a typical path, and the latter quantity is the typical fraction of wild type cells given that the total number of cells is $N$. The two are close only in special cases, where stochastic effects are not too strong.

In the derivation of the fixed-time formula for the average number of mutants, (5), the wild-type cells were assumed to grow deterministically. In our fixed-size approach, all cells grow
stochastically. If we introduce the stochastic growth in the continuous-time model, then the conditional average number of mutants subject to the colony not going extinct is given approximately by $(y)/(1 - d/l)$, see Eqs. (43) and (44), which for $d > 0$ is even larger than the value in formula (5). In other words, making the growth of the wild-type stochastic in the fixed-time approach makes the agreement only worse. The reason for this is the incorrect relation between the time-variable and the size of the colony, $(x(t) + y(t)) = N$.

In Fig. 9 we compare the boundary-layer solution obtained by the method of Section 4.1 with the diffusion-free solution, Eq. (28), which is the same as Eq. (5). We plot the relative difference 

$$\frac{E(mut) - E(Y)}{E(mut)},$$

between the expected number of mutants, $E(mut)$, calculated by the boundary-layer method, and $E(Y)$, the diffusion-free formula (28). The value of $d$ runs from 0 to 0.49. The relative difference between the fixed-size solution and the fixed-time solution is greater for the small mutation rate case, Fig. 9(a); it is also significant for the large population size case, Fig. 9(b). The relative difference continues to grow for $d > 0.49$, but in that region the boundary-layer method breaks down. For values of $d$ very close to $1/2$ we can use the analytical formulas for $E(mut)$ (not shown). For example, in the limit where $d \rightarrow l$, the average number of mutants of the fixed-size problem behaves in the following way: for large population sizes ($uN \gg 1$) it tends to $N$ and for small mutation rates ($uN \ll 1$) it tends to $uN^2/2 \ll 1$. The corresponding solution of the fixed-time problem, Eq. (5), tends to $N$ no matter what the value of $uN$ is.

5.3. Biological implications

Once more we emphasize the difference between the fixed-size and fixed-time approaches. Both approaches can be applied to problems in oncology and other areas. The fixed time problem gives rise to a distribution of mutants created after a certain time after the initiation of the tumor. This can be relevant for in vitro experiments where a colony of cancer cells grows for a fixed duration of time, and then treatment is applied. Similarly, this approach can describe in vivo experiments where tumors are implanted in mice and then therapeutic agents are introduced after a fixed

![Fig. 9. The relative difference between the fixed-time calculation and a fixed-size calculation with the boundary-layer method, for different values of N, and (a) $uN = 0.1$, (b) $uN = 10$.](image-url)
time-lapse. In a real clinical situation, however, it is impossible to determine the tumor age at the start of treatment. Instead, we can measure the size of the tumor. Therefore, a more relevant approach in this case is the fixed-size model.

In this paper we have found the following interesting trends in the expected number of mutants in a population of a given size.

- The cellular population with a higher death rate will contain a larger number of mutants than a population of an equal size but with a smaller death rate. This is a consequence of the fact that the average number of cell divisions that it takes to go from one cell to \( N \) cells is larger for populations with a higher death rate, which increases the total expected number of mutants created.

- A larger population will contain a larger percentage of mutants. For example, for small mutation rates, \( uN \ll 1 \), the fraction of mutants at size \( N \) is given by \( uN \), that is, it is proportional to the colony size! In the limit of very large populations, such that \( uN \gg 1 \), the fraction of mutants approaches 100% as \( N \to \infty \). These results are superficially counterintuitive, because there is no selective advantage to the mutants: they divide at exactly the same rate as the wild-type cells. Still, we see an effect resembling selection: the proportion of mutants grows as the colony expands. The reason for this of course is not selection, but the fact that we are dealing with unidirectional mutations. Once a mutant is created, its offspring can only be mutants. At the same time, each division of wild-type cells has a chance of creating a new mutant (thus decreasing the percentage of wild-type cells). Therefore, in the limit \( N \to \infty \), the colony will contain an increasing fraction of mutants.

Comprehensive solutions of the fixed-time and fixed-size problems now open the door for more complex problems. For instance, one direction is the fixed-size problem in the case of advantageous and disadvantageous mutants. Another extension of the present work is a more detailed analysis of the variance of the number of mutants, beyond the partial treatment given in the Appendix.

Acknowledgement

N.K. thanks the support of the Sloan Fellowship and the NIH grant 1R01AI058153-01A2. The work of PB and LW has been supported by grants from NIH (LM-07443-01) and NSF (EIA-0321390). We thank Alex Sadovsky for useful discussions.

Appendix A. Continuous approximation in the 1D case

It is interesting to see how the continuous approximation works in the 1D case. For this purposes we will use a usual asymmetric Markov chain with absorbing boundary conditions at 0 and \( N \). We have only one variable, \( 0 \leq i \leq N \), which increases by 1 with rate \( l \) and decreases by 1 with rate \( d \), \( l + d = 1 \). The probability to get absorbed in state \( i = N \) starting from state \( i \) is denoted by \( h_i \). We have
The diffusion approximation with \( x = i/N \) reads,

\[
\frac{1}{2N} h'' + h'(l - d) = 0, \quad h(0) = 0, \quad h(1) = 1.
\]

The solution is

\[
h_{\text{cont}}(x) = \begin{cases} 
\frac{e^{-(l-d)2Nx}}{e^{-(l-d)2N} - 1}, & l \neq d, \\
x, & l = d.
\end{cases}
\]

If \( l = d \), the continuous solution is the limit of the discrete solution. Let us determine all the regimes where the continuous solution is a good approximation for the discrete one. Using \( l = 1 - d \), we have (with \( i = xN \) in the discrete case):

\[
h_{\text{discr}}(x) = \frac{(d/N)^{xN} - 1}{(d/N)_{xN} - 1}, \quad h_{\text{cont}} = \frac{e^{2(1-d)Nx}}{e^{2(1-d)N} - 1}.
\]

Let us set \( d = 1/2 - \delta \). Two cases can be distinguished.

- \( xN \gg 1 \). If \( \delta Nx \gg 1 \), then

\[
\left( \frac{d}{1-d} \right)^{xN} \ll 1, \quad e^{2(1-d)Nx} \ll 1,
\]

and both solutions are exponentially close to 1, that is, they are very close to each other. If \( xN\delta \sim 1 \), or \( xN\delta \ll 1 \), expressions (46) are not small. In this case, let us denote \( N \equiv \beta/\delta \). We have

\[
h_{\text{discr}}(x) - h_{\text{cont}}(x) = \frac{16\beta e^{4\beta(1-x)}(1 - x + e^{4\beta x} - e^{x}) \delta^2}{3(e^{4\beta} - 1)^2} + O(\delta^4).
\]

We conclude that for \( Nx \gg 1 \), the continuous solution is a good approximation for the discrete solution.

- \( xN \sim 1 \). If \( \delta \gg 1/N \), we can replace the denominators of expressions for \( h_{\text{discr}} \) and \( h_{\text{cont}} \) by \((-1)\).

For \( \delta \ll 1 \), we have in this case,

\[
h_{\text{discr}}(x) - h_{\text{cont}}(x) = \frac{16\delta^3}{3} (1 + O(\delta)).
\]

In the case where \( N\delta \ll 1 \), we have

\[
h_{\text{discr}}(x) - h_{\text{cont}}(x) = \frac{8N\delta^3}{3} x(1 - x)(1 + O(N\delta)).
\]

Our analysis shows that for large \( N \), the continuous solution is a good approximation for the discrete solution if
• $xN \sim 1$, or
• $x \sim 1/N$ and $|l - d| \ll 1$.

The latter case is of especial importance since we are often interested in the behavior of a colony starting from one cell, that is, $x = 1/N$. Note that the conditions above are equivalent to those for the 2D case, Eqs. (21) and (22).

Appendix B. The subcritical regime

Depending on the value of $\epsilon$ (definition (41)), there are several regions where solution methods are different.

B.1. Strongly subcritical death rates

We define strongly subcritical death rates as satisfying $\epsilon > 0$ and $\epsilon \gg \frac{1}{N}$. In this case, the boundary layer width satisfies $a \ll 1$, because $|l - d| \gg 1/N$, and therefore the boundary layer method is applicable. This means that we only need to solve the system numerically in a small (of the size $\sim a$) corner of the domain.

Let us examine the applicability conditions for the continuous approximation, Eqs. (21) and (22). In the case of condition (22), the width of the boundary layer, $a$, is comparable with the step-size, $1/N$, and inside this layer the gradients of the solution are very high. This shows that the PDE is not a good description inside the boundary layer. On the contrary, in the case of condition (21), the boundary layer contains many grid points, $a \gg 1/N$. Even though the PDE has a boundary layer structure, the gradients are reasonably small and the continuous approximation works even inside the boundary layer.

B.2. Weakly-subcritical and nearly-critical death rates

We will call the death rates satisfying $\epsilon > 0$ and $\epsilon \ll \frac{1}{N}$ the nearly-critical regime. In this case, we have $|l - d| \ll 1$ (we assume that $u \ll 1$), and therefore the discrete system is well approximated by the PDE in the entire domain $0 < x + y < 1$. Let us write down PDE (19) in terms of $\epsilon$:

$$\begin{align*}
\frac{\partial w}{\partial t} &\equiv x(2 - u) + \frac{1}{2N} \left[ z_{xx} \left( \frac{2(1 - u)}{2 - u} - \epsilon u \right) + z_{yy} \left( \frac{2(1 - u)}{2 - u} \right) \right]
\end{align*}$$

(47)

Neglecting the terms multiplying $\epsilon$, we obtain the equation

$$z_{xy}(x + y) + \frac{1}{N}\left( z_{xx} + z_{yy} \right) = 0$$

with the boundary conditions $z_{xx}(x, 1 - x) = Nx$ and $z_{yy}(0, 0) = 0$. This equation is exact in the critical case $\epsilon = 0$, and it is solved by $z_{xy}(x, y) = z_{xx}(x, y) = Nx$, Eq. (38). Note that this result
is independent of the magnitude of the quantity $Nu$. This critical solution is a good approximation for the actual solution in the nearly-critical regime.

It is interesting to also examine the behavior of the diffusion-free solution (Eq. (28)) in this regime. There are two limiting cases:

- if $u \ll \epsilon \ll 1$, the exponent $lu/(l - d) \ll 1$, and we have
  \[ z_{w, \text{diff-free}}(x, y) \approx \frac{Nx}{x + y}. \]
  We can see that in this case the diffusion-free approximation does not work. On the other hand,
- if $\epsilon \ll u \ll 1$, the exponent $lu/(l - d) \approx 1$, and we have
  \[ z_{w, \text{diff-free}}(x, y) \approx Nx. \]

In this case the diffusion-free approximation works well.

The same conclusion can be obtained by examining the first- and second-derivative terms in Eq. (47).

Finally, we consider the borderline case, $\epsilon > 0$ and $\epsilon \sim \frac{1}{N}$, which we call weakly subcritical. There are two limiting cases:

- The limit of large population sizes, $uN \gg 1$. In this case, for $\epsilon \sim 1/N$, the solution is well described by the boundary-layer structure with the diffusion-free solution away from the origin.
- The limit of small mutation rates, $uN \ll 1$. In this regime, the solution does not have a boundary layer structure, and neither is it described by the critical solution. The only solution method we know for this case is a numerical solution of the PDE.

**B.3. Summary of the results for subcritical death rates**

The behavior of the solution differs depending on the value of the product $uN$.

In the case of small mutation rates, $uN \ll 1$, strongly subcritical, weakly subcritical and nearly critical regimes are qualitatively different, which is illustrated in Fig. 10, left. There, the solid lines denote solutions $z_{w}(x, 0)$ of PDE (19), and the dashed lines represent diffusion-free solutions (28), for different values of the parameter $\epsilon$ (Eq. (41)). We can see that in the strongly-subcritical regime (thin lines), the solution of the PDE has a boundary-layer structure such that it is well approximated by the diffusion-free solution away from $(0, 0)$. In the weakly-subcritical regime (the lines of intermediate thickness), the solution $z_{w}(x, y)$ is not close to the diffusion-free solution. Neither is it close to the critical solution, Eq. (38). Finally, in the nearly-critical regime, the function $z_{w}$ approaches the critical solution $Nx$, see the thickest line in Fig. 10, left. Note that in the latter case, the diffusion-free solution is not a good approximation because $\epsilon = u$; in order for $z_{w, \text{diff-free}}$ to approach the critical solution, we must have $\epsilon \ll u$.

The picture is a lot simpler in the limit of large population sizes, $uN \gg 1$, see Fig. 10, right. For any positive $\epsilon$, the solution $z_{w}(x, y)$ has a boundary-layer structure and is well described by the diffusion-free solution. Fig. 10, right, shows the behavior of the solution in strongly subcritical,
weakly subcritical, and nearly critical regimes, for $uN = 10$. We can see that the diffusion-free solution is very close to the full solution of the PDE in all the regimes, except for a boundary layer. As $\epsilon/C_1$ decreases from being much greater than $1/N$ to being much smaller than $1/N$, both $z^{\text{wt}}(x,y)$ and the diffusion-free solution gradually turn into the critical solution, $N_x$, Eq. (38).

The difference in the behavior between these two cases, $uN \ll 1$ and $uN \gg 1$, is further illustrated in Fig. 11. There, we plot the numerical solutions of system (17) and (18), $z^{\text{mut}}_{i0}$ (connected circles), and the diffusion-free approximation, solid lines, for different values of $d$, in the subcritical regime in the two limiting cases. The observed difference in the behavior can be explained intuitively by estimating the width of the boundary layer, Eq. (40), at the critical value of $d$. We have $d_* \approx 1/2 - u/4$, and $|l_* - d_*| \sim u$. This means that at $d = d_*$, the boundary layer width is $a \sim (uN)^{-1}$. If $uN \gg 1$, we have $a \ll 1$, and the problem has a boundary layer structure. If $uN \ll 1$, we have $a \gg 1$, which indicates that there is no boundary layer. It is easy to see that in the $uN \ll 1$ case, the boundary layer can only exist for values of $d$ much below $1/2 - 1/N$. In our notation, this corresponds to $\epsilon \gg 1/N$.

Appendix C. The supercritical regime

We will refer to death rates satisfying $\epsilon < 0$, or $d_* < d \leq 1/2$, as supercritical. The value of $\epsilon$ in definition (41) is restricted from below, $\epsilon \geq -u/[4(1 - u/2)]$. There are several distinct supercritical regimes.

C.1. The limit of small mutation rates

Let us suppose that $uN \ll 1$. In this case, we have $|l - d| \ll 1/N$. The behavior of the system is well approximated by the PDE where we set $l = d$:

$$l u(x) z^{\text{wt}}_x - z^{\text{wt}}_y + \frac{1}{2N} (z^{\text{wt}}_x x + z^{\text{wt}}_y y) = 0.$$  \hspace{1cm} (48)

Since $uN \ll 1$, the diffusion terms dominate, and we can see that the solution is well approximated by the critical solution, Eq. (38).
C.2. The analytical solution

It is possible to derive analytical solutions of system (15) and (16) in the supercritical regime with $uN \ll 1$. This is done by the method of iterations. Let us rewrite Eq. (15) in the form:

$$(i + j)z_{ij}^{mut} - i(lz_{i+1,j}^{mut} + dz_{i-1,j}^{mut}) - j(lz_{i,j+1}^{mut} + dz_{i,j-1}^{mut}) = liu(z_{i,j+1}^{mut} - z_{i+1,j}^{mut}).$$

This system can be solved in the case where $u = 0$:

$$z_{ij}^{(0)} = \frac{Nj}{i + j} \frac{1 - (d/l)^{i+j}}{1 - (d/l)^N},$$

(49)

that is, the expected fraction of mutants is given by the initial fraction of mutants, $j/(i + j)$, times the probability of the colony to not go extinct, $\frac{1 - (d/l)^{i+j}}{1 - (d/l)^N}$. We can now solve Eqs. (15) and (16) under the assumption that $uN$ is small. We represent the solution as a series, $z_{ij} = z_{ij}^{(0)} + \sum_{l=1}^{\infty} u^l z_{ij}^{(l)}$. At the $m$th order, we have the following inhomogeneous equation:

$$(i + j)z_{ij}^{(m)} = i\left(lz_{i+1,j}^{(m)} + dz_{i-1,j}^{(m)}\right) + j\left(lz_{i,j+1}^{(m)} + dz_{i,j-1}^{(m)}\right) + li\left(-z_{i+1,j}^{(m-1)} + z_{i,j+1}^{(m-1)}\right), \quad m \geq 1

z_{0j}^{(m)} = z_{N-j,j}^{(m)} = 0.$$

Explicit analytical solutions of this iterative scheme are possible in some special cases. In what follows we calculate the mean number of mutants for $d = 1/2$. Similar calculations for the variance are presented in Section D.

Note that if $l = d = 1/2$, the probability of non-extinction needs to be interpreted as the limit of expression (14) as $d \to l$; we have $P_{\text{non-ext}}_j = j/N$. Thus the first of boundary conditions (16) becomes $z_{0,j} = j$. The expression for the zeroth order solution, (49), also simplifies: $z_{ij}^{(0)} = j$. Iterating up to the third order, we obtain the following solution for $E(mut) = Nz_{10}$:

$$E(mut) = \frac{uN(N-1)}{2} - \frac{u^2N(N-1)(N-2)}{6} + \frac{7}{144}u^3N(N-1)(N-2)(N-17/7) + O([uN]^4).$$

(50)
Note that the first two terms of this expansion coincide with the Taylor expansion of the critical solution, Eq. (36).

C.3. The limit of large population sizes

In the opposite limit, \( uN \gg 1 \), the supercritical solutions behave quite differently. We have three regimes, depending on the value \( d \) defined in Eq. (42).

**Nearly critical values:** \( d \gg 1/N \). In this case, the solution of the PDE has a boundary layer structure, that is, it can be approximated by the diffusion-free solution. It deviates significantly from the critical solution unless \( j \sim d/C_0 \), i.e.,

\[
\frac{z_{\text{wt}}}{u^v} = 0.
\]

This equation can be solved together with the boundary conditions \( z_{\text{wt}}(0,0) = 0 \) and \( z_{\text{wt}}(0,1) = 0 \). We obtain \( z_{\text{wt}}(0, y) = 0 \). For \( x \gg 0 \), let us now try to approximate Eq. (48) by its first-derivative component,

\[
z_x - z_y = 0.
\]

Solutions of this equation satisfy \( z_{\text{wt}}(x, y) = f(x + y) \), for some function \( f \). Using \( z_{\text{wt}}(0, y) = 0 \) as a boundary condition, we obtain \( z_{\text{wt}}(x, y) = 0 \). The diffusion-free solution (28) has the same limiting behavior as \( d \to l \): the exponent becomes very large and \( z_{\text{wt}}(x, y) \to 0 \) pointwise.

It is more informative to consider the expected number of mutants, \( z_{\text{mut}} \), rather than the expected number of wild-type cells, \( z_{\text{wt}} \). For this variable, Eq. (48) can be approximated by (51) and solved with the boundary condition \( z_{\text{mut}}(0, y) = Ny \):

\[
z_{\text{mut}}(x, y) = N(x + y) \equiv z_{\text{mut},1/2}(x, y).
\]

The superscript 1/2 refers to the limiting value of the death rate, \( d = 1/2 \). For lack of a better term, we call this solution the ‘large-population, high-death-rate solution’.

Note that the solution procedure that we use for both \( z_{\text{wt}} \) and \( z_{\text{mut}} \) is not well justified. Indeed, the boundary condition is given along the line \( x = 0 \), where the first-order approximation (51) does not hold. However, intuitively, we can see how solution (52) can be obtained. Let us consider the characteristics (26) and (27), Fig. 5. As \( d \to 1/2 \), the characteristics map the solution along the line \( x + y = \alpha \) with \( 0 < \alpha < 1 \) onto a very narrow region of the line \( x + y = 1 \), near \( y = 1 \). In the limit \( d \to 1/2 \), they converge to one point, \( (0,1) \), giving \( z_{\text{wt}}(x, y) = 0 \). For \( d \) near 1/2, the characteristics can be approximated by straight lines \( x + y = C \) with \( 0 < C < 1 \), except in a very narrow boundary layer near the vertical \( x = 0 \). Inside that region, they shoot up to reach a point near \( (0,1) \). Indeed, by expanding functions (26) and (27) in terms of small \( \delta \), Eq. (42) yields:

\[
x(t) = -lut + x_0 + O(\delta), \quad y(t) = lut + y_0 + O(\delta).
\]

This shows that for the zero-order approximation, we have \( x(t) + y(t) = x_0 + y_0 \equiv C \). Solution (52) approximates the characteristics by the straight lines \( x + y = C \), ignoring the boundary layer.
This gives a result (formula (52)) which very well approximates the behavior of the PDE away from the boundary $x + y = 1$.

Fig. 12, left, shows the limiting behavior of the solution $z_{\text{mut}}(x,0)$ as $uN \rightarrow \infty$. We can see that this function approaches its limit $z_{\text{mut}}(x,0) = Nx$, Eq. (52), pointwise; the convergence is faster away from the boundary $x + y = 1$.

C.4. Summary of the results for supercritical death rates

As with the subcritical regime, the behavior of the solution differs depending on the value of the product $uN$.

In the case of small mutation rates, $uN \ll 1$, the behavior is very simple: both the solutions of PDE (19), $z^{w}(x, y)$, and the diffusion-free solutions (28) are very close to the critical solution, Eq. (38), for all $\epsilon < 0$.

In the limit of large population sizes, $uN \gg 1$, the picture is more complicated, see Fig. 12, right. In the example of the figure, we use $uN = 10^3$. In the critical regime and nearly-critical regime, the solution $z^{w}(x, y)$ is well approximated by diffusion-free solution (28), the two thinnest lines in Fig. 12, right. In the weakly-subcritical regime, there is a noticeable difference between the full and the diffusion-free solution. This difference continues to grow as $d$ approaches 1/2, and the exact solution tends to a limiting shape, where for a large fraction of the interval $x \in [0,1]$, the solution is well approximated by $z^{w} = 0$, and the average number of mutants, $z_{\text{mut}}(x, y)$, tends to the large-population, high-death-rate solution, Eq. (52). This approximation becomes exact in the limit $uN \rightarrow \infty$, see also Fig. 12, left.

Appendix D. The variance

Here we present analytical results for the values of the variance, in several special cases.

Denote by $v_{ij}^{\text{mut}}$ the expected value of the square of the number of mutants starting from point $(i, j)$. It satisfies:
The standard deviation, or the width of the distribution, is \( \sqrt{D.2} \). The critical case for small mutation rates.

In the special case of the zero death rate, the variance can be obtained by using the probability distribution \( (7) \) and formula \( (8) \), see [14]. The variance is given by \( \text{Var(mut)} = E^2(N) - E^1(N) - E^1(N^2) \). We obtain,

\[
\text{Var(mut)} = Nu(N - \log N - \gamma)[1 + O(u \log N)].
\]

The standard deviation, or the width of the distribution, is \( \sqrt{uN(1 + [\log N + \gamma]/N)} \), i.e., to the lowest order, it scales with the system size, \( N \).

**D.1. The case \( d = 0 \)**

In the special case of the zero death rate, the variance can be obtained by using the probability distribution \( (7) \) and formula \( (8) \), see [14]. The variance is given by \( \text{Var(mut)} = E^2(N) - E^1(N) - E^1(N^2) \). We obtain,

\[
\text{Var(mut)} = Nu(N - \log N - \gamma)[1 + O(u \log N)].
\]

The standard deviation, or the width of the distribution, is \( \sqrt{uN(1 + [\log N + \gamma]/N)} \), i.e., to the lowest order, it scales with the system size, \( N \).

**D.2. The critical case for small mutation rates.**

Let us use Eqs. \( (53) \) and \( (54) \). Substituting \( d = l(1 - u) \), we obtain the following system:

\[
2(i + j)v_{ij}^{\text{mut}} - i(v_{i+1,j}^{\text{mut}} + v_{i-1,j}^{\text{mut}}) - j(v_{i,j+1}^{\text{mut}} + v_{i,j-1}^{\text{mut}}) = (i + j)\frac{u}{1 - u}(v_{i,j+1}^{\text{mut}} - v_{i,j}^{\text{mut}}),
\]

\[
v_{0j}^{\text{mut}} = N^2 \frac{1 - (d/l)^j}{1 - (d/l)^N}, \quad v_{N-j,j}^{\text{mut}} = j^2, \quad 0 \leq j \leq N.
\]

We represent the solution in terms of a series expansion, \( v_{ij}^{\text{mut}} = \sum_{m=0}^{\infty} p^m v_{ij}^{(m)} \), where \( p = u/(1 - u) \). The zeroth order solution can be found:

\[
v_{ij}^{(0)} = j(N - i).
\]

The higher order terms satisfy the following recursion:

\[
2(i + j)v_{ij}^{(m)} - i(v_{i+1,j}^{(m)} + v_{i-1,j}^{(m)}) - j(v_{i,j+1}^{(m)} + v_{i,j-1}^{(m)}) = (i + j)(v_{i,j+1}^{(m-1)} + v_{i,j}^{(m-1)}),
\]

\[
v_{0j}^{(m)} = B_m, \quad v_{N-j,j}^{(m)} = 0, \quad m \geq 1,
\]

where \( B_m \) are terms in the Taylor expansion of the function \( N^2 \frac{1 - (d/l)^j}{1 - (d/l)^N} \) in terms of small \( p \). Here we present the results up to the fourth order:

\[
v_{1,0}^{\text{mut}} = \frac{p}{6}(2N^2 - 3N + 1) + \frac{p^2}{36}(4N^3 - 17N^2 + 14N - 1) + \frac{p^3}{2160}(-9N^4 - 295N^3 + 1085N^2 - 785N + 4) + \frac{p^4}{129600}(-108N^5 + 387N^4 + 20125N^3 - 65405N^2 + 44783N + 218) + \cdots
\]
This series converges as \( uN \to 0 \). For the variance we have

\[
\text{Var(mut)} = v_{1,0}/P_{1}^\text{non-ext} - (E(mut)^2) = \frac{uN^3}{3} [1 + O(1/N) + O(uN)],
\]

where the contribution of the second term, the square of the mean, is negligible.

### D.3. The case \( d = 1/2 \) for small mutation rates

Using Eqs. (53) and (54), we obtain the following system:

\[
2(i + j)v^{mut}_{ij} - i(v^{mut}_{i+1,j} + v^{mut}_{i-1,j}) - j(v^{mut}_{i,j+1} + v^{mut}_{i,j-1}) = iu(-v^{mut}_{i+1,j} + v^{mut}_{i,j+1}),
\]

\[
v^{mut}_{0j} = Nj, \quad v^{mut}_{N-j,j} = j^2, \quad 0 \leq j \leq N.
\]

Again, let us represent the solution as a series, \( v^{mut}_{ij} = \sum_{m=0}^\infty u^m v^{(m)}_{ij} \). The zeroth order solution is again (56). The higher order terms satisfy the following recursion:

\[
2(i + j)v^{(m)}_{ij} - i(v^{(m)}_{i+1,j} + v^{(m)}_{i-1,j}) - j(v^{(m)}_{i,j+1} + v^{(m)}_{i,j-1}) = i(-v^{(m-1)}_{i+1,j} + v^{(m-1)}_{i,j+1}),
\]

\[
v^{(m)}_{0j} = v^{(m)}_{N-j,j} = 0, \quad m \geq 1.
\]

We have, up to the fourth order,

\[
v^{mut}_{1,0} = \frac{u}{6} \left(2N^2 - 3N + 1\right) - \frac{u^2}{18} \left(N^4 - 5N^2 + 8N - 4\right) + \frac{u^3}{10800} \left(21N^4 - 530N^3 + 2755N^2 - 4870N + 2624\right)
\]

\[
+ \frac{u^4}{648000} \left(1707N^5 - 8019N^4 - 21295N^3 + 173615N^2 - 304412N + 158404\right) + \ldots
\]

Again, this series converges if \( uN \to 0 \). The variance of the number of mutants in this regime is given by Eq. (57).

### Appendix E. The fixed-time and fixed-size calculations for the variance

We have seen that fixed-time and fixed-size calculations can lead to a significant difference in the mean number of mutants. The difference between the two approaches becomes even more apparent when we consider the variance. Following the method of Section 5.2, we can derive the following equations for the second moments:

\[
\frac{d}{dt}\langle x^2 \rangle = 2\langle x^2 \rangle [L(1 - u) - D] + \langle x \rangle [L(1 - u) + D],
\]

\[
\frac{d}{dt}\langle y^2 \rangle = 2\langle y^2 \rangle (L - D) + 2Lu\langle xy \rangle + Lu\langle x \rangle + \langle y \rangle (L + D),
\]

\[
\frac{d}{dt}\langle xy \rangle = \langle xy \rangle [L(1 - u) - D + L - D] + Lu\langle x^2 \rangle.
\]

Let us set \( D = 0 \) and solve Eqs. (43), (44), (58)–(60) with the initial conditions \( \langle x \rangle (t = 0) = 1 \), \( \langle x^2 \rangle (t = 0) = 1 \), and zero for the rest of the variables. Substituting time \( t = L^{-1} \log N \), we obtain

\[
\tilde{V}(mut) = \langle y^2 \rangle - \langle y \rangle^2 = Nu(2(N/M_0 - 1) - \log(N/M_0))(1 + O(u \log N)),
\]

which is roughly the double of the correct value, formula (55). The difference comes from the following fact. In a fixed-time calculation, there is a nonzero probability to have any number
of mutants. In the fixed-size calculation, the number of mutants is restricted by $N$. The infinite ‘tail’ of the probability distribution in the fixed-time model is responsible for a larger value of the variance.

An expression similar to (61) can also be obtained from the results of [12] for a deterministically growing colony. Similarly to formula (5), we obtain

$$V(Y) = \frac{N}{L - D + Lu} \left( \frac{N}{M_0} \right)^{\frac{1}{2}} (L(1 - u) + D) - (L + D) + \frac{Lu}{L - D} \left( 2L \frac{N}{M_0} - (L + D) \right).$$  \hspace{1cm} (62)

Note that in the limit of small $uN$, the variance for the number of mutants obtained from system (58, 60) coincides with formula (62). In Fig. 13 we present a comparison of formula (62), solid lines, with the values obtained by stochastic simulations, squares. We can see that the results for the variance are not in a good agreement with the simulations in the entire range of $d$.

From Eqs. (58)–(60), we can also calculate the variance in the number of wild-type cells,

$$\tilde{V}(wt) = N^2(1 - 2u \log N + O((u \log N)^2)).$$

We can see that the variance does not disappear for $u = 0$. This is in sharp contrast with the exact result, where the variance in the number of wild-type cells is the same as that in the number of mutants (formula (55)), and it is zero for $u = 0$. The difference is easily explained by examining the system which leads to Eqs. (58)–(60). In a stochastically growing colony of cells, the size at time $t = L^{-1} \log N$ is a stochastic variable with a mean $N$ and a nonzero variance. This uncertainty contributes to the nonzero variance in the number of wild type cells even if $u = 0$.

References


N.L. Komarova et al. / Mathematical Biosciences 210 (2007) 253–290 289