Rigorous hitting times for binary mutations

Evolutionary Computation, **7**(2):167–203, 1999.

J. Garnier, L. Kallel and M. Schoenauer

CMAP – UMR CNRS 7641 , Ecole Polytechnique,
Palaiseau 91128, France.
E-mail: *garnier@cmapx.polytechnique.fr*
E-mail: *kallel@cmapx.polytechnique.fr*
E-mail: *marc@cmapx.polytechnique.fr*

Abstract

In the binary evolutionary optimization framework, two mutation operators are theoretically investigated. For both the standard mutation, in which all bits are flipped independently with the same probability, and the *1-bit-flip* mutation, which flips exactly one bit per bitstring, the statistical distribution of the first hitting times of the target are thoroughly computed (expectation and variance) up to terms of order l (the size of the bitstrings) in two distinct situations: without any selection, or with the deterministic (1+1)-ES selection on the OneMax problem. In both cases, the *1-bit-flip* mutation convergence time is smaller by a constant (in terms of l) multiplicative factor. These results extend to the case of multiple independent optimizers.

1 Introduction

One known drawback of Evolutionary Algorithms as function optimizers is the amount of computational efforts they require to reach their goal. Hence, a lot of work has been devoted to the study of theoretical properties of of these algorithms, aiming at bounding their complexity, or time-to-solution.
In real-valued spaces, detailed results have been obtained for some particular functions (e.g. the sphere and the corridor models) since the early days of Evolution Strategies [36, 43]. The most recent and powerful results provide bounds for the Euclidean distance to the solution in different situations: Beyer [8, 9, 10, 11] handles sophisticated variants of standard ES (i.e. based on Gaussian mutations, with or without self-adaptive mutation parameters), while Rudolph [41, 40] studies the local convergence speed for a larger class of mutation operators.

The situation is not so clear when if comes to optimization in the discrete bitstring space: Though some results prove the global convergence of evolutionary algorithms, very few works succeeded in addressing the issue of convergence time.
Initial investigations about Genetic Algorithms (GAs) concentrated on a constructive view of the algorithms, based on the Schema Theorem and the Building Blocks hypothesis [27, 22]. Though the Schema Theory has been widely discussed and commented upon [4, 47, 16], it has also been usefully generalized [35], leading to the heuristic approaches to the design of efficient algorithms [44]. But it has never led to any strong theoretical result about convergence, or complexity, of the algorithm itself.

The main theoretical results in the field of evolutionary binary function optimization have been obtained using Markov chain analyses. Though considered a marginal operator in the GA paradigm [27, 22], the mutation operator is at the core of all convergence results obtained so far in that domain, responsible for the ergodicity of the Markov chain representing the evolutionary process: Early works [19, 17, 33, 45] consider simplified models; Rudolph convergence results [38] directly rely on the positivity [1] of the mutation operator, while their recent extension by Agapie [3] relaxes the strong hypothesis of positive mutations to the existence of a finite chain of mutations linking any two points of the space with non-zero probability (irreducible and diagonal positive transition mutation matrix). Still in the Markov framework but using completely different techniques (the Friedlin-Wentzell theory of stochastic perturbation of dynamical systems), Cerf [12, 13] obtains a lower bound for the population size allowing to prove finite time convergence results for an inhomogeneous Genetic Algorithm. His results are also primarily based on mutation (but he shows that an acceleration of convergence is clearly brought by crossover).
None of these results makes any particular assumption on the fitness function. But on the other hand, they do not give any indication about the complexity of the algorithm: the hitting time of the associated Markov chain, i.e. the expectation of the number of generations required to reach the optimum, remains unknown.

The first results about complexity [5, 32, 6, 41] have been obtained on the simple well-studied OneMax problem (the fitness is the number of ones in the bitstring, as introduced by Ackley [2]), though some other particular functions were studied as well (e.g. in [39, 18]). And again, most of such works concentrate on the mutation operator[2]. These results will be recalled in more details in section 2.

In that streamline of research, this paper investigates two types of binary mutation operators, namely the standard bitwise mutation (all bits are flipped independently with the same probability) and the mutation that flips exactly k bits of each bitstring, randomly chosen with uniform probability.
The efficiency of these operators are compared through the derivation of exact hitting times of the corresponding Markov chains. Two situations are studied: First, the operators are applied on an arbitrary fitness landscape, without selection, on a single bitstring, e.g. in a (1,1)-ES algorithm. Second, the coupling with deterministic selection, the (1+1)-ES algorithm, is investigated for the OneMax problem.
In both cases, the exact expectations of the average hitting times, as well as their standard deviations, are calculated up to terms of order l (the size of the bitstrings).

The paper is organized as follows: In section 2, both mutation operators are summarized, and the existing theoretical results about convergence time are recalled. The main results of the paper are also introduced. Section 3 presents the proofs of the results in the case of the (1,1)-ES, i.e. without selection, while section 4 derives the hitting times of the two mutation operators coupled with selection. The final section presents some hints for future directions of research. Some basic results in probability theory, intensively used throughout the demonstrations, are recalled in the Appendix.

[1] there exists a mutation that links any two points of the space with non-zero probability

[2] though the influence of crossover has been investigated by Culberson [15], who shows that the number of local optima in CX landscape increase exponentially in problem size. Nevertheless, to circumvent these local optima, it is essential that crossover takes place between fairly close parents[28].

2 Comparing Binary Mutation Operators

2.1 Framework

From now on, the search space will be the binary space $\{0,1\}^l$ for some $l \in \mathbb{N}$. The usual mutation operator in the binary space for evolutionary methods is the standard $\frac{c}{l}$ mutation, which flips all bits independently with probability $\frac{c}{l}$, for some $c \in \mathbb{R}^+$. However, a common practice is also to consider the *k bit flip* mutation, witch flips exactly $k \in \mathbb{N}$ randomly chosen bits in each individual.
Note that both above operators are homogeneous operators, i.e. they do not change along evolution. Whereas nearly all theoretical results only consider the homogeneous case, Cerf's results [12, 13] emphasize a decreasing mutation rate (tightly coupled to an increasing selection pressure) to reach finite time convergence – but again, without hitting time estimation. On the other hand, decreasing mutation rate along the generations is a common practice, even though very little theoretical results support it (in [7], an hyperbolic decreasing scheme is advocated). Once again, practice is one step beyond theoretical studies.

2.2 State of the Art

As briefly recalled in the introduction, there has been some work about the expected complexity of some simple evolutionary algorithms:
Bäck [5] derived the transition probabilities of the (1+1)-ES for the population Markov chain on the OneMax problem for the $\frac{c}{l}$ mutation. In the same framework, Mühlenbein [32] considers a simpler Markov chain that has (provable) worse performance. He gives an upper bound for convergence time, and proves that this bound is minimal for the $\frac{1}{l}$ mutation, equal to $l\log(l)$. Bäck [6] numerically approximates the expected absorption time of the transition matrix, and validates the $l\log(l)$ bound as a close approximation of the true behavior. Rudolph [41] widely comments on these studies, as well as on the validity of the approximations. He also proves a $l\log(l)$ upper bound for the (1+1)-ES using the *1-bit-flip* mutation.
Recently, Droste & al. [18] were able to improve on those results by actually showing that these upper bounds are reached: they prove that the complexity of the (1+1)-ES with $\frac{c}{l}$ mutation is indeed in $O(l\log(l))$. Moreover, they generalize this result to any linear binary function.

However, the above mentioned works cannot be used to actually compare the $\frac{c}{l}$ and the *1-bit-flip* mutations operators, as they provide order of magnitude results without giving the multiplicative constants. The purpose of this paper is to achieve such comparison by actually giving more accurate expansions of the convergence times as l goes to infinity.

Some hints have been given, though, in the work of Rudolph. In [39], he shows that for large values of l the (1+1)-ES using $\frac{1}{l}$ mutation is a lot quicker than the single-bit-flip hill-climber in optimizing long k-path problem with fixed k values (polynomial versus exponential time in l). The difference between the two algorithms is that the (1+1)-ES can do more than one bit-flip at the same time: The $\frac{1}{l}$ mutation flips one bit per bitstring *in average* only.
However, an immediate consequence is that the $\frac{1}{l}$ mutation (or more generally the $\frac{c}{l}$ mutation) is also "able" to do no bit-flip at all on some individuals: Rudolph also mentions that the distribution of the random variable K representing the number of flipped bits per bitstring with the $\frac{c}{l}$ mutation is well-known to be

$$P[K = k] = (C_k^l)(\frac{c}{l})^k(1 - \frac{c}{l})^{l-k} \simeq \frac{e^{-c}}{k!}c^k$$

A straightforward consequence is that, for large values of l, a fraction of e^{-c} of the mutations does not modify the bitstring at all (about one third for $c = 1$).

This raises a serious question about the efficiency of such mutation: such a big amount of non-mutated individuals causes a consequent waste of time: even if the fitness does not have to be evaluated for unmodified bitstrings, such an inefficient mutation still leads to additional computational time.

2.3 Summary of results

In any evolutionary algorithm, the strong interactions between selection and evolution operators (crossover, mutation, ...) makes it difficult to efficiently compare the effects of different operators without taking into account the characteristics of the selection, i.e. the fitness function.
A standard method to avoid such a bias is to start by considering the effects of each operator alone: selection without crossover or mutation (see the studies of *takeover time* [24, 14]), crossover or mutation in the absence of selection [46, 23].

With respect to the already known results cited here above which provide asymptotic bounds for the expected hitting times with a "big-Oh" precision, we shall give exact estimates for the expected hitting times for the different processes, which allows a rigorous comparison between the $\frac{c}{l}$ and the *1-bit-flip* mutations operators. Further we shall give the complete statistical distributions of the hitting times. Such results are in particular necessary if one wants to get the expected first hitting time T_N of N independent processes. If one only has an estimate of the expected time without an estimate of the corresponding variance, one has absolutely no idea of what the order of T_N will be. If the variance if almost zero, all hitting times take the same value and one has $\mathbb{E}[T_N] \simeq \mathbb{E}[T_1]$. But if the variance is large, one can expect that $\mathbb{E}[T_N] \ll \mathbb{E}[T_1]$, since one of the independent process will certainly take an exceptional low value that can be taken by T_1.

The first part of this study is restricted to the study of both the $\frac{c}{l}$ and the *1-bit-flip* mutations in the absence of selection, i.e. inside a simple (1,1)-ES algorithm. Such algorithm can also be viewed as a random walk directed by the corresponding mutation. The purpose of this part of the paper is then to evaluate as precisely as possible the statistical distribution of the first hitting time of a given target point.
The main result (see Section 3) in that case is that the expectation of the first hitting time for the $\frac{c}{l}$ mutation is larger by a factor $(1 - e^{-c})^{-1}$ than that of the *1-bit-flip* mutation. The variances and the complete statistical distributions at the first order with respect to l of both hitting times are also derived. It is found that the hitting time of the $\frac{c}{l}$ mutation (resp. *1-bit-flip* mutation) obeys an exponential distribution with mean $\frac{2^l}{(1-e^{-c})}$ (resp. 2^l). Furthermore, if we compare the sequences of visited points by both mutations, we shall see that the $\frac{c}{l}$ mutation jumps over some points visited by the *1-bit-flip* mutation. Hence, when the first process reaches the target point for the first time, the second process misses it with probability $1 - \frac{1-e^{-c}}{c}$. Finally, the case of several independent processes is addressed. The first hitting time of N independent $\frac{c}{l}$ mutations (resp. *1-bit-flip* mutations) obeys an exponential distribution with mean $\frac{2^l}{N(1-e^{-c})}$ (resp. $\frac{2^l}{N}$).

The second part of this study brings together the mutation operators and the simple deterministic (1+1)-ES selection (a single parent generates a single offspring, and the fittest of both becomes the next parent).
It has already been argued [41] that is easier to prove useful results on the OneMax problem than on any other binary function optimization problem, as increasing the fitness is equivalent to decreasing the Hamming distance to the optimum.
This paper indeed buys this argument, as all computations of that part (Section 4) make intensive use of that fitness-distance equivalence to precisely study the statistical distribution of the first hitting times of the global maximum (all ones) when using either mutation operator. It is proved that the hitting time $\bar{T}$ of the $\frac{c}{l}$ mutation (resp. the hitting time T of the *1-bit-flip* mutation) can be expanded up to terms of order $o(l)$ as: $\bar{T} = \frac{e^c}{c} l \ln l - \frac{e^c}{c} l \ln Z + lC(c)$ (resp. $T = l \ln l - l(\ln Z + \ln 2)$), where Z obeys an exponential distribution with mean 1 and density $p(z) = e^{-z} \mathbb{I}_{z \geq 0}$, and $C(c)$ is some c-dependent constant. Therefore the dominant term is deterministic and is equal to $\frac{e^c}{c} l \ln l$ (resp. $l \ln l$). Furthermore a random corrective term proportional to l involves a normalized standard deviation (that is to say the standard deviation divided by the mean) of the order of $(\ln l)^{-1}$. We then address the case of N independent processes, and prove that the first hitting time $\bar{T}_N$ of N independent $\frac{c}{l}$ mutations (resp. the first hitting times T_N of N independent *1-bit-flip* mutations) can be expanded as $\bar{T}_N = \frac{e^c}{c} l \ln l - \frac{e^c}{c} l \ln Z_N + lC(c)$ (resp. $T_N = l \ln l - l(\ln Z_N + \ln 2)$), where Z_N obeys a distribution with density $p_N(z) = N(1 - e^{-z})^{N-1} e^{-z} \mathbb{I}_{z \geq 0}$. Hence there is very little gain when using several independent processes. This is due to the weak variances of the distributions of T and $\bar{T}$.

3 Mutations without selection: the (1,1)-ES

3.1 Notations and conventions

State space. In this paper l stands for an integer, and we shall develop our theory in the asymptotic framework $l \gg 1$. The state space of the processes that will be studied is $E = \{0,1\}^l$, whose cardinality is 2^l. We shall consider the Hamming distance d defined between two strings $X, X' \in E$ by:

$$d(X, X') = \sum_{i=1}^{l} \mathbb{I}_{X_i \neq X'_i}.$$

Process X. The process $X(n)$ is a process starting from some point X^0, whose evolution is governed by the following way. At each time n, you build the string $X(n)$ by inverting each bit of $X(n-1)$ independently with probability c/l. The process $X(n)$ is a Markov chain whose stationary transition matrix of size $2^l \times 2^l$ defined by:

$$P_{X,X'} := \mathbb{P}(X(1) = X' \,|\, X(0) = X),$$

is given by:

$$P_{X,X'} = (cl^{-1})^{d(X,X')} \left(1 - cl^{-1}\right)^{l - d(X,X')}.$$

The process X corresponds to a (1,1)-ES with a c/l mutation.

Process $\bar{X}$. The process $\bar{X}(n)$ is a process starting from some point X^0, whose evolution is governed by the following way. At each time n, you choose a random integer uniformly over $\{1, ..., l\}$. Then you invert the corresponding bit of the string $\bar{X}(n-1)$. The resulting string is $\bar{X}(n)$. The process $\bar{X}$ is a Markov chain whose stationary transition matrix is:

$$\bar{P}_{X,X'} = \begin{cases} 0 & \text{if } d(X,X') \neq 1 \\ l^{-1} & \text{if } d(X,X') = 1 \end{cases}$$

The process $\bar{X}$ corresponds to a (1,1)-ES with a 1-bit-flip mutation.

Goal of the study. We fix some point X^f and we look for the statistical distributions of the respective first hitting times T and $\bar{T}$ of the point X^f by the chains $X(n)$ and $\bar{X}(n)$.

3.2 Main results

In this subsection we begin with stating the main results. The proofs of which follow in the next subsections.

PROPOSITION **3.1**
If $d(X^0, X^f) \neq 0$, then:

$$\mathbb{E}[T] \simeq \frac{2^l}{1-e^{-c}}, \qquad \mathbb{E}[\bar{T}] \simeq 2^l,$$

in the sense that $2^{-l} \times \mathbb{E}[T]$ (resp. $2^{-l} \times \mathbb{E}[\bar{T}]$) goes to $1/(1-e^{-c})$ (resp. 1) when $l \to \infty$.

Note that whatever c, the (1,1)-ES with a c/l mutation is slower than the (1,1)-ES with a 1-bit-flip mutation.

REMARK **3.2**
We can compare these results with the mean values the first hitting times of X^f by some other simple processes. Let us consider the following processes:
1. Random process
At each time you choose some string uniformly over the state space E. The first hitting time obeys the distribution:

$$\mathbb{P}(T_1 = n) = 2^{-l}(1-2^{-l})^{n-1}$$

whose mean is:

$$\mathbb{E}[T_1] = 2^l.$$

2. Exhaustive process
You choose some arbitrary order relation over E and the strings of E are tested one after the other according to the order relation. The first hitting time obeys a uniform distribution over $\{1, ..., 2^l\}$ whose mean is:

$$\mathbb{E}[T_2] = 2^{l-1} + \frac{1}{2}.$$

The second method presents the advantage that it never tests twice the same string, and consequently it looses less time than the first one. It has to be noticed that the strategy corresponding to the Markov chain $X(n)$ is the less efficient.

We can be more precise and give the complete statistical distributions of T and $\bar{T}$ in the framework $l \gg 1$.

PROPOSITION **3.3**
If $d(X^0, X^f) \neq 0$, then $2^{-l}(1-e^{-c})T$ (resp. $2^{-l}\bar{T}$) converges in distribution to a random variable with exponential distribution and mean 1 as $l \to \infty$.

This Proposition has two straightforward corollaries.

COROLLARY **3.4**
If $d(X^0, X^f) \neq 0$, then the standard deviations of the variables T and $\bar{T}$ are equal to their respective means as $l \to \infty$.

$$\mathbb{E}\left[(T - \mathbb{E}[T])^2\right] \simeq \frac{2^{2l}}{(1-e^{-c})^2}, \qquad \mathbb{E}\left[(\bar{T} - \mathbb{E}[\bar{T}])^2\right] \simeq 2^{2l}.$$

COROLLARY **3.5**
Assume that we deal with a set of N processes X_i (resp. $\bar{X}_i$) which evolve independently. Then the global first hitting time T (resp $\bar{T}$), which is the infimum of the N independent hitting times, is such that the statistical distribution of $T2^{-l}$ (resp. $\bar{T}2^{-l}$) converges to that of an exponential random variable with mean $N^{-1}(1-e^{-1})^{-1}$ (resp. N^{-1}). In particular its expectation is:

$$\mathbb{E}[T] \simeq \frac{2^l}{N(1-e^{-c})}, \qquad \mathbb{E}[\bar{T}] \simeq \frac{2^l}{N},$$

The expected first hitting time is N times smaller than that of a single process. Hence, the number of evaluations remains the same, and there is no advantage to increase N.

3.3 Study of the chain $\bar{X}(n)$

Let us introduce the auxiliary Markov process $Z(n)$ with state space $\{0, ..., l\}$ starting from $Z(0) = k_f := d(X^0, X^f)$ and with transition matrix $\bar{Q}$:

$$\bar{Q}_{k,k'} = \begin{cases} 0 & \text{if } k' \neq k-1, k+1 \\ 1 - \frac{k}{l} & \text{if } k' = k+1 \\ \frac{k}{l} & \text{if } k' = k-1 \end{cases}$$

It is easy to check that:

$$d(\bar{X}(n), X^f) \stackrel{law}{=} Z(n). \tag{1}$$

Obviously the first hitting time of the point 0 by $Z(n)$ corresponds to the first hitting time $\bar{T}$ of the point X^f by the chain $\bar{X}(n)$.
Let us denote by ν_k the first hitting time of the point k by the chain $Z(n)$ and by $\bar{\nu}_k$ the first hitting time ≥ 1 of the point k. These times coincide if the chain starts from a point distinct from k. But if the chain starts from k, then $\nu_k = 0$ while $\bar{\nu}_k \geq 2$.
Remark: The Markov chain Z is known in the mathematical and physical literature under the term "Ehrenfest model". One can find a nice introduction to this model in [29, Section 4.6], and one can also find in [31, Section 7.3] the computation of the expectation of the first hitting time ν_0 when starting from $l/2$, so that Lemma 3.6 is not really new. Nevertheless, we shall use Lemma 3.6 and intermediate results derived in its proof to exhibit the complete distribution of ν_0, and not only its expectation. The complete distribution is indeed necessary to estimate the variance of ν_0, and also to find the distribution and the expectation of the infimum of the first hitting times of N independent process $(\bar{X}_i)_{i=1,...,N}$.

LEMMA **3.6**
1. If $k_f = 0,1,2,3$, we have for any l[3]*:*

$$\mathbb{E}_0[\nu_0] = 0,$$
$$\mathbb{E}_1[\nu_0] = 2^l - 1,$$
$$\mathbb{E}_2[\nu_0] = \frac{l}{l-1}2^l - \frac{2l}{l-1},$$
$$\mathbb{E}_3[\nu_0] = \frac{l^2-l+2}{(l-2)(l-1)}2^l - \frac{3l^2-3l+2}{(l-2)(l-1)}.$$

2. If $k_f = 1,...,l$, then we have for $l \gg 1$:

$$\mathbb{E}_{k_f}[\nu_0] \simeq 2^l,$$

which means that $2^{-l} \times \mathbb{E}_{k_f}[\nu_0]$ goes to 1 as $l \to \infty$.

Proof. Step 1. Statistical expectations of the hitting times.
Obviously $\mathbb{E}_0[\nu_0] = 0$. Let us assume in the following that $k_f \geq 1$. There exists a unique invariant probability measure for the chain $Z(n)$ (i.e. $\mu\bar{Q} = \mu$) which is given by:

$$\mu(k) = 2^{-l}C_l^k.$$

By the general theory of Markov chains, since $Z(n)$ is irreducible and recurrent [37, ch.5, exercise 3.14], the first hitting time $\bar{\nu}_k$ satisfies:

$$\mathbb{E}_k[\bar{\nu}_k] = \frac{1}{\mu(k)}.$$

In particular: $\mathbb{E}_0[\bar{\nu}_0] = 2^l$. Starting from 0, the first jump is non random: it occurs at time 1 and the process goes to the point 1. Thus:

$$\mathbb{E}_0[\bar{\nu}_0] = \mathbb{E}_0\left[\bar{\nu}_0 | \, Z(1) = 1\right] = 1 + \mathbb{E}_0\left[\bar{\nu}_0 - 1 | \, Z(1) = 1\right].$$

Applying the weak Markov property (cf Appendix): $\mathbb{E}_0[\bar{\nu}_0] = 1 + \mathbb{E}_1[\nu_0]$, which implies:

$$\mathbb{E}_1[\nu_0] = 2^l - 1. \tag{2}$$

Let $k \in \{0, ..., l-1\}$. Starting from $k+1$, the chain Z jumps either in k or in $k+2$. We can therefore split $\mathbb{E}_{k+1}[\nu_k]$ into two terms:

$$\mathbb{E}_{k+1}[\nu_k] = \mathbb{E}_{k+1}[\nu_k | Z(1) = k]\mathbb{P}_{k+1}(Z(1) = k) + \mathbb{E}_{k+1}[\nu_k | Z(1) = k+2]\mathbb{P}_{k+1}(Z(1) = k+2)$$

which implies by the weak Markov property:

$$\mathbb{E}_{k+1}[\nu_k] = 1 \times \bar{Q}_{k+1,k} + (1 + \mathbb{E}_{k+2}[\nu_k])\, \bar{Q}_{k+1,k+2}. \tag{3}$$

The Markov chain X only goes to neighbored levels of $\{0,1\}^l$, and therefore the Markov chain Z has to reach $k'-1$,...,k while going from k' to k (with $k < k'$). Using the strong Markov property we then get for any $k' > k$:

$$\mathbb{E}_{k'}[\nu_k] = \sum_{j=k}^{k'-1} \mathbb{E}_{j+1}[\nu_j] \tag{4}$$

[3] $\mathbb{E}_k$ (resp. $\mathbb{P}_k$) stands for the expectation (resp. the probability) with respect to the statistical distribution of the chain $Z(n)$ starting from $Z(0) = k$.

Furthermore, substituting Eq. (4) with $k'=2$ into (3), we get the inductive relation:

$$\mathbb{E}_{k+2}[\nu_{k+1}] = (\mathbb{E}_{k+1}[\nu_k]-1)\,\frac{k+1}{l-(k+1)} - 1. \tag{5}$$

Substituting (2) and (5) into the sum (4) with $k=0$, we get the complete expression of $\mathbb{E}_{k_f}[\nu_0]$, valid for any l. We can also bound $\mathbb{E}_{k_f}[\nu_0]$ for any $k_f \geq 1$ by:

$$\mathbb{E}_1[\nu_0] \leq \mathbb{E}_{k_f}[\nu_0] \leq \mathbb{E}_1[\nu_0] + \mathbb{E}_{k_f}[\nu_1]. \tag{6}$$

In the following two steps, we are going to prove that the last term of the right-hand member of this estimate is negligible compared to $\mathbb{E}_1[\nu_0]$ in the asymptotic framework $l \to \infty$. This is the main issue of the proof, and although it sounds surprising, it can be explained heuristically. Indeed, when the chain Z is at level 1, it has a chance $1/l$ to reach 0. It is much likely to go to level 2, and then to go the middle levels which contain almost all states. Thus, the chain needs to visit level 1 about $O(l)$ times to finally succeed in reaching level 0. So we could expect that $\mathbb{E}_1[\nu_0] \sim l\mathbb{E}_1[\bar{\nu_1}] \sim l\mathbb{E}_{k_f}[\nu_1]$. Hence $\mathbb{E}_{k_f}[\nu_1] \sim 2^l/l$. Let us turn to the proof of this statement.
Step 2. If $k \leq [l/2]$, then $\mathbb{E}_k[\nu_1] \leq 2^{l+1}/(l-1)$.
From (5) we compute:

$$\begin{aligned}
\mathbb{E}_2[\nu_1] &= \frac{2^l}{l-1} - \frac{l+1}{l-1}, && (7)\\
\mathbb{E}_3[\nu_2] &= \frac{2^{l+1}}{(l-2)(l-1)} - \frac{l(l+1)+2}{(l-2)(l-1)}. && (8)
\end{aligned}$$

Besides, for any $k \leq l/2$, Eq. (5) also implies that:

$$\mathbb{E}_{k+1}[\nu_k] \leq \mathbb{E}_k[\nu_{k-1}].$$

Summing with respect to $k=3,...$ starting from (8) yields that, for any $k \leq 2$:

$$\mathbb{E}_k[\nu_2] \leq \frac{2^{l+1}}{(l-1)(l-2)} \times \left(\frac{l}{2}-2\right) \leq \frac{2^l}{l-1}.$$

Summing with (7) establishes the desired result.
Step 3. If $k \geq [l/2]$, then $\mathbb{E}_k[\nu_{[l/2]}] \leq l^2/4$.
If $Z(0)=l$, then the chain has no other choice but to jump to $l-1$:

$$\mathbb{E}_l[\nu_{l-1}] = 1. \tag{9}$$

Eq. (5) also reads as: $\mathbb{E}_k[\nu_{k-1}] = 1 + \frac{l-k}{k}(\mathbb{E}_{k+1}[\nu_k]+1)$, which implies that, for any $k \geq l/2$: $\mathbb{E}_k[\nu_{k-1}] \geq 2 + \mathbb{E}_{k+1}[\nu_k]$. By a recursive argument starting from (9), we get that: $\mathbb{E}_{k+1}[\nu_k] \leq 2(l-k)$. Summing with respect to k yields the desired result.
Step 4. Conclusion
Injecting the statements of Steps 2 and 3 into:

$$\mathbb{E}_{k_f}[\nu_1] \leq \sup_{k\leq[l/2]} \mathbb{E}_k[\nu_1] + \sup_{k\geq[l/2]} \mathbb{E}_k[\nu_{[l/2]}]$$

we get that, for any $k_f \geq 1$:

$$\mathbb{E}_{k_f}[\nu_1] \leq \frac{3}{l}2^l. \tag{10}$$

which yields the result of the Lemma in view of Eq. (6). □

REMARK **3.7**
This proof also shows that 2^l steps are necessary to go from 1 to 0 (in distance to target point), while coming from any $k_f > 1$ to 1 in distance takes $O(2^l/l)$ steps.

Let us now study the statistical distribution of $2^{-l}\nu_0$.

LEMMA **3.8**
If $k_f \geq 1$, then $2^{-l}\nu_0$ converges in distribution to a random variable with exponential distribution and mean 1 as $l \to \infty$.

Proof.
Step 1. Tightness of $2^{-l}\nu_0$.
By Lemma 3.6:

$$\sup_{l\in\mathbb{N}} \mathbb{E}_{k_f}[2^{-l}\nu_0] < \infty$$

which is sufficient to establish the tightness (and hence the relative compactness) of the sequence of variables $2^{-l}\nu_0$ [20, Theorem II-2-2]. Tightness insures that there exists at least one weak limit of the sequence $(2^{-l}\nu_0)_{l\in\mathbb{N}}$.
Step 2. Convergence in distribution.
Let a, b be two positive real numbers. We get from the strong Markov property that:

$$\mathbb{P}_k\left(\nu_0 \geq [2^l a] + [2^l b] \middle| \nu_0 \geq [2^l a]\right) = \sum_{j=1}^{l} \mathbb{P}_k\left(Z([2^l a]) = j \middle| \nu_0 \geq [2^l a]\right) \mathbb{P}_j(\nu_0 \geq [2^l b]) \quad (11)$$

On the one hand the following obvious estimate holds true for any $j \geq 1$:

$$\mathbb{P}_j(\nu_0 \geq [2^l b]) \geq \mathbb{P}_1(\nu_0 \geq [2^l b]). \quad (12)$$

On the other hand, conditioning with respect to the first time when the chain Z hits 1:

$$\mathbb{P}_j(\nu_0 \geq [2^l b]) = \sum_{t_1=0}^{\infty} \mathbb{P}_1(\nu_0 \geq [2^l b] - t_1)\mathbb{P}_j(\nu_1 = t_1)$$

If $\delta > 0$, then Markov inequality yields:

$$\mathbb{P}_j(\nu_1 \geq [2^l\delta]) \leq \frac{\mathbb{E}_j[\nu_1]}{[2^l\delta]}.$$

Combining with Eq. (10) implies that:

$$\mathbb{P}_j(\nu_1 \geq [2^l\delta]) \leq \frac{3}{\delta l}.$$

We then get:

$$\mathbb{P}_j(\nu_0 \geq [2^l b]) \leq \mathbb{P}_1(\nu_0 \geq [2^l b] - [2^l\delta])\left(1 - \frac{3}{\delta l}\right) + \frac{3}{\delta l}. \quad (13)$$

Substituting (12) and (13) into (11) establishes:

$$\begin{aligned} &\liminf_{l\to\infty} \mathbb{P}_1(\nu_0 \geq [2^l b]) \leq \liminf_{l\to\infty} \mathbb{P}_k\left(\nu_0 \geq [2^l a] + [2^l b] \middle| \nu_0 \geq [2^l a]\right) \\ \leq &\limsup_{l\to\infty} \mathbb{P}_k\left(\nu_0 \geq [2^l a] + [2^l b] \middle| \nu_0 \geq [2^l a]\right) \leq \limsup_{l\to\infty} \mathbb{P}_1(\nu_0 \geq [2^l b] - [2^l\delta]). \end{aligned} \quad (14)$$

Let ν be a weak limit of $2^{-l}\nu_0$. Such a limit exists by Step 1, but needs not to be unique, and may depend on the starting point k. Multiplying by 2^{-l} both members inside the $\mathbb{P}(...)$ terms of (14), we obtain:

$$\mathbb{P}_1(\nu \geq b) \leq \mathbb{P}_k(\nu \geq a+b|\nu \geq a) \leq \mathbb{P}_1(\nu \geq b-\delta)$$

In this expression δ is an arbitrary positive real number. Letting $\delta \to 0$, we find that the distribution of ν does not depend on k (take $a=0$, $b>0$) and satisfies for any $a,b>0$:

$$\mathbb{P}(\nu \geq b) = \mathbb{P}(\nu \geq a+b|\nu \geq a).$$

Lemma 5.2 in the Appendix then implies that ν is a random variable with exponential distribution. Since its mean is 1, we then get the uniqueness of the weak limit, and consequently the convergence in distribution of $2^{-l}\nu_0$ to this variable. □

3.4 Study of the chain $X(n)$

The approach is very similar to that of the previous section. Let us introduce the process Z:

$$Z(n) = d(X(n), X^f).$$

$(Z(n))_{n\in\mathbb{N}}$ is a Markov chain with transition matrix Q whose coefficients are for $j,j' \in \{0,...,l\}$:

$$Q_{j,j'} = \sum_{k=0}^{j}\sum_{k'=0}^{l-j} \mathbb{1}_{k'-k=j'-j} C_j^k C_{l-j}^{k'} (cl^{-1})^{k+k'}(1-cl^{-1})^{l-k-k'}.$$

The chain Z is recurrent and its unique invariant probability measure is μ:

$$\mu(j) = 2^{-l}C_l^j. \tag{15}$$

Let us adopt the following notations: ν_k stands for the first hitting time of the point k by the chain $Z(n)$ and $\bar{\nu}_k$ stands for the first hitting time ≥ 1 of the point k. The expectation with respect to the statistical distribution of the chain Z starting from k is denoted by $\mathbb{E}_k$.

LEMMA **3.9**
If $k_f \geq 1$, then:

$$\lim_{l\to\infty} 2^{-l}\mathbb{E}_{k_f}[\nu_0] = \frac{1}{1-e^{-c}}.$$

Proof. In the following, the notation $\mathbb{E}_p$ is a shorthand for the expectation with respect to the statistical distribution of the chain Z starting from $Z(0)$ which obeys the distribution $\mathbb{P}_p(Z(0)=k) = p(k)$. Moreover for $k=0,...,l$, we denote by δ_k the distribution over $\{0,...,l\}$ given by $\delta_k(j)=1$ if $j=k$ and 0 otherwise.
Step 1. For any $k=0,...,l$, we have $\mu(k)\left(\mathbb{E}_{\delta_k Q}[\nu_k]+1\right)=1$.
Conditioning with respect to the first jump, we have for any initial probability measure p_0 over $\{0,...,l\}$:

$$\begin{aligned}\mathbb{E}_{p_0}[\nu_k] &= \sum_{i\neq k}\sum_{j=0}^{l} p_0(i)Q_{i,j}(\mathbb{E}_j[\nu_k]+1)\\ &= \sum_{j=0}^{l}(p_0Q)(j)(\mathbb{E}_j[\nu_0]+1) - p_0(k)(\sum_{j=0}^{l} Q_{k,j}\mathbb{E}_j[\nu_k]+1),\end{aligned}$$

which also reads as:

$$\mathbb{E}_{p_0}[\nu_k] = (\mathbb{E}_{p_0 Q}[\nu_k] + 1) - p_0(k)(\mathbb{E}_{\delta_k Q}[\nu_0] + 1). \tag{16}$$

Let p_0 be a probability measure over $\{0, ..., l\}$. Defining inductively $p_{n+1} = p_n Q$, applying (16) inductively with respect to n and summing, we find that for any n:

$$\mathbb{E}_{p_n}[\nu_k] = \mathbb{E}_{p_0}[\nu_k] + n - \sum_{j=0}^{n-1} p_j(k) \left(\mathbb{E}_{\delta_k Q}[\nu_k] + 1\right). \tag{17}$$

By the ergodic theorem for Markov chains [37, ch. 6, theorem 2.8] p_n converges to the unique invariant probability measure μ as n goes to infinity. Dividing by n Eq. (17), and letting $n \to \infty$, we get the desired result.
Step 2. If $1 \le k \le l-2$, then $\mathbb{E}_{k+1}[\nu_k] \le 2^l e^c c^{-1} (C_{l-1}^k)^{-1}$.
Applying Step 1, since $(\delta_k Q)(k+1) = Q_{k,k+1}$ we get:

$$\mu(k) Q_{k,k+1} \mathbb{E}_{k+1}[\nu_k] \le 1.$$

The probability $Q_{k,k+1}$ to go from k to $k+1$ is bounded below by the probability that one of the $l-k$ bits which coincide with those of X^f is inverted while the other are not:

$$Q_{k,k+1} \ge (1 - cl^{-1})^{l-1}(l-k)cl^{-1} \ge ce^{-c}(1 - k/l).$$

From the expression (15) of the invariant probability measure we have $\mu(k) \times (1 - k/l) = 2^{-l} C_{l-1}^k$ which yields the result.
Step 3. $\lim_{l\to\infty} 2^{-l} \sup_{1 \le k_0 < k_1 \le l-1} \mathbb{E}_{k_1}[\nu_{k_0}] = 0.$
Let us consider some $1 \le k_0 < k_1 \le l-1$.

$$\mathbb{E}_{k_1}[\nu_{k_0}] \le \sum_{k=k_0}^{k_1-1} \mathbb{E}_{k+1}[\nu_k] \le \sum_{k=1}^{l-2} \mathbb{E}_{k+1}[\nu_k].$$

The estimate of Step 2 is sufficient to bound this term uniformly with respect to $1 \le k_0 < k_1 \le l-1$ by $e^c c^{-1} 2^{l+1}/l$.
Step 4. $\lim_{l\to\infty} 2^{-l} \sup_{1 \le k_0 < k_1 \le l-1} \mathbb{E}_{k_0}[\nu_{k_1}] = 0.$
This step can be deduced from the previous one from the invariance property of the distribution of the chain Z with respect to the transformation $\{0, ..., l\} \mapsto \{l, ..., 0\}$.
Step 5. $\lim_{l\to\infty} 2^{-l} \sum_{j=1}^{\infty} \frac{c^j e^{-c}}{j!} \mathbb{E}_j[\nu_0] = 1.$
A simple computation shows that:

$$(\delta_0 Q)(j) = Q_{0,j} = C_l^j (cl^{-1})^j (1 - cl^{-1})^{l-j} \text{ for } j \in \{0, ..., l\}.$$

Applying Step 1 with $k = 0$ yields $\mu(0)\left(\mathbb{E}_p[\nu_0] + 1\right) = 1$. Since $\mu(0) = 2^{-l}$ and $p(j) \to \frac{c^j e^{-c}}{j!}$ as $l \to \infty$, we get the desired result.
Step 6. For any $j \ge 1$, $\lim_{l\to\infty} 2^{-l} \mathbb{E}_j[\nu_0] = \dfrac{1}{1 - e^{-c}}$.
From the obvious estimates

$$\mathbb{E}_1[\nu_0] \le \mathbb{E}_1[\nu_j] + \mathbb{E}_j[\nu_0], \qquad \mathbb{E}_j[\nu_0] \le \mathbb{E}_1[\nu_0] + \mathbb{E}_j[\nu_1]$$

which hold true for any $j \geq 1$, we get that:

$$|\mathbb{E}_j[\nu_0] - \mathbb{E}_1[\nu_0]| \leq \max(\mathbb{E}_j[\nu_1], \mathbb{E}_1[\nu_j]).$$

Applying the results of Step 3 (in the case $k_0 = 1$) and Step 4 (in the case $k_0 = 1$), we obtain:

$$\limsup_{l\to\infty} 2^{-l} \sup_{j\leq l-1} |\mathbb{E}_j[\nu_0] - \mathbb{E}_1[\nu_0]| = 0. \tag{18}$$

It remains to estimate $\mathbb{E}_l[\nu_0]$. From the relation

$$\mathbb{E}_l[\nu_0] = 1 + \sum_{j=0}^{l} \mathbb{P}_l(Z(1) = j)\mathbb{E}_j[\nu_0]$$

we get that:

$$\mathbb{E}_l[\nu_0] = \frac{1 + \sum_{j=0}^{l-1} \mathbb{P}_l(Z(1) = j)\mathbb{E}_j[\nu_0]}{1 - \mathbb{P}_l(Z(1) = l)}.$$

Since $\mathbb{P}_l(Z(1) = l) = Q_{l,l} = (1-c/l)^l \simeq e^{-c}$ the Lebesgue dominated convergence theorem and (18) implies that:

$$\limsup_{l\to\infty} 2^{-l}|\mathbb{E}_l[\nu_0] - \mathbb{E}_1[\nu_0]| = 0. \tag{19}$$

Substituting (18) and (19) into Step 5 and applying the dominated convergence theorem finally establish that:

$$\lim_{l\to\infty} 2^{-l}\mathbb{E}_1[\nu_0] = \frac{1}{1-e^{-c}}.$$

Substituting into (18) and (19) completes the proof of the Lemma. □

LEMMA **3.10**
If $k_f \geq 1$, then $2^{-l}(1 - e^{-c})\nu_0$ converges in distribution to a random variable with exponential distribution and mean 1 as $l \to \infty$.

Proof. The proof of the convergence in distribution of $2^{-l}\nu_0$ is based on the very same arguments as those of the proof of Lemma 3.8. □

3.5 Heuristic comparison between the chains X and $\bar{X}$

The statistical distributions of the chain $X(n)$ and of the corresponding stopping time T can actually be derived from the statistical distributions of the chain $\bar{X}(n)$ and of the corresponding stopping time $\bar{T}$. However the reader must be warned that this derivation is not as rigorous as the proofs of the previous Lemmas, in the sense that there is a gap between the statement of Proposition 3.11 which definitively holds true and its application that will be pointed out here below. Nevertheless, the arguments presented in this section are rather intuitive, and we feel that they are worth noticing.
Let us introduce a sequence of independent and identically distributed random variables $(N_j)_{j\in\mathbb{N}}$ with Poisson distribution:

$$\mathbb{P}(N_1 = j) = \frac{c^j e^{-c}}{j!}, \tag{20}$$

and denote by S_n the partial sums:

$$S_n = \sum_{j=1}^{n} N_j.$$

The following proposition is based on the fact that the distribution of N_n is in the asymptotic framework $l \to \infty$ equivalent to the statistical distribution of the number of inverted bits during the n-th transition of the process X.

PROPOSITION **3.11**
In the asymptotic framework $l \to \infty$ we have:

$$(X(n))_{n\in\mathbb{N}} \overset{law}{\simeq} (\bar{X}(S_n))_{n\in\mathbb{N}}.$$

Proof. Let X^0, X^1 be two strings of the state space E and denote $d := d(X^0, X^1)$. On the one hand, denoting $\bar{P}^j$ is the j-th power of the matrix $\bar{P}$:

$$\begin{aligned}
\mathbb{P}\left(\bar{X}(S_{n+1}) = X^1 \middle| \bar{X}(S_n) = X^0\right) &= \sum_{j=0}^{\infty} \mathbb{P}(N_{n+1} = j)\mathbb{P}\left(\bar{X}(j) = X^1 \middle| \bar{X}(0) = X^0\right) \\
&= \sum_{j=0}^{\infty} \frac{c^j e^{-c}}{j!} (\bar{P}^j)_{X^0,X^1} \\
&\overset{l\to\infty}{\simeq} c^d e^{-c} l^{-d},
\end{aligned}$$

because $(\bar{P}^d)_{X^0,X^1} = d! l^{-d}$ and $(\bar{P}^j)_{X^0,X^1} = O(l^{-d-1})$ for any $j \neq d$. On the other hand, :

$$\begin{aligned}
\mathbb{P}\left(X(n+1) = X^1 | X(n) = X^0\right) &= (cl^{-1})^d (1 - cl^{-1})^{l-d} \\
&\overset{l\to\infty}{\simeq} c^d e^{-c} l^{-d}.
\end{aligned}$$

The comparison of these estimates implies the statement of the Lemma. □

In the following discussion, we shall apply the relation $X(.) = \bar{X}(S_.)$ without taking care of the small error terms. Although we do know that this not completely rigorous because the small error terms which are generated at every transition could cumulate and become non-negligible, we shall obtain by this way a nice interpretation of the relations between the expressions of the expected hitting times T and $\bar{T}$ of the chains X and $\bar{X}$ that have been proved in the above sections.

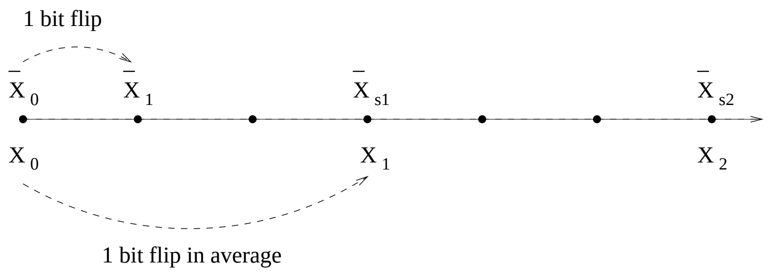

FIG 1 : *The chain $X(n)$ jumps over some strings visited by the chain $\bar{X}(n)$.*

By the renewal theory (cf Appendix and (43)), since $\mathbb{E}[N_1] = c$, the speed rate of S_n is $c \times n$. Therefore we could expect from the relation $X(.) = \bar{X}(S_.)$ that $X(.)$ would hit X^f about c times faster than $\bar{X}(.)$. But the process $S_.$ does not take all integer values.

Consequently, when $\bar{X}(.)$ hits X^f for the first time at time ν_0, the process $S_.$ may jump over the value ν_0, which implies that $X = \bar{X}(S_.)$ has some probability p_{miss} to miss X^f, if ν_0 is strictly between two successive times S_n, S_{n+1}. This probability can be computed by the renewal theory (cf Appendix and (44)) in the asymptotic framework $\nu_0 \gg 1$, which holds true when $l \gg 1$:

$$p_{miss} = 1 - \frac{\mathbb{P}(N_1 = 0)}{\mathbb{E}[N_1]} = 1 - \frac{1 - e^{-c}}{c}.$$

If ν_0^j stands for the j-th hitting time of X^f by the chain $\bar{X}(.)$, then the first hitting time T of $X = \bar{X}(S_.)$ is given by the first ν_0^j (divided by the speed rate c) which corresponds exactly to S_n for some n. The expectation of T can be split over the number J which is the first j when ν_0^j coincides with some S_n:

$$\mathbb{E}[T] = \frac{1}{c}\sum_{j=0}^{\infty} \mathbb{E}_{k_f}[\nu_0^j | J = j]\mathbb{P}(J = j).$$

By the strong Markov property, the excursions of the chain $\bar{X}(.)$ between two successive hitting times ν_0^j and ν_0^{j+1} are independent and identically distributed. Thus:

$$\mathbb{E}[T] = \frac{1}{c}\sum_{j=1}^{\infty} \left(\mathbb{E}_{k_f}[\nu_0] + \mathbb{E}_0[\bar{\nu}_0](j-1)\right)\mathbb{P}(J = j).$$

Since $\mathbb{P}(J = j) = p_{miss}^{j-1}(1 - p_{miss})$ and $\mathbb{E}_0[\bar{\nu}_0] = 2^l$, we finally get that:

$$\mathbb{E}[T] = \frac{1}{c}\left(\mathbb{E}_{k_f}[\nu_0] + \frac{c - 1 + e^{-c}}{1 - e^{-c}}\mathbb{E}_0[\bar{\nu}_0]\right) \overset{l\to\infty}{\simeq} 2^l \frac{1}{1 - e^{-c}}. \tag{21}$$

4 Mutation and selection: the (1+1)-ES

4.1 Main results

In the same way as above, two processes are compared: $\bar{X}$ corresponds to a (1+1)-ES with a 1-bit-flip mutation per individual. And X corresponds to a (1+1)-ES with a c/l mutation.

Let us consider first the process $\bar{X}$ starting from X^0 whose evolution is governed in the following way. At each time n, you choose a random integer uniformly over $\{1, ..., l\}$. Then you invert the corresponding bit of the string $\bar{X}(n-1)$. The resulting string is denoted by $\tilde{X}(n-1)$. If $d(\tilde{X}(n-1), X^f) < d(\bar{X}(n-1), X^f)$, then $\bar{X}(n) = \tilde{X}(n-1)$, otherwise $\bar{X}(n) = \bar{X}(n-1)$. The transition matrix of the chain $\bar{X}(n)$ is consequently:

$$\bar{P}_{X,X'} = \begin{cases} 1 - d(X, X^f)l^{-1} & \text{if } d(X, X') = 0, \\ l^{-1} & \text{if } d(X, X') = 1 \text{ and } d(X', X^f) = d(X, X^f) - 1, \\ 0 & \text{otherwise.} \end{cases}$$

PROPOSITION **4.1**

1. Assuming that the starting point X^0 is such that $d(X^0, X^f) = k$, then the expectation of $\bar{T}$ is:

$$\mathbb{E}_k[\bar{T}] = l\sum_{j=1}^{k} j^{-1} \overset{l\geq k\gg 1}{\simeq} l(\ln k + C)$$

where C is the Euler's constant whose value is $C \simeq 0.58$.

2. Assuming that the starting point X^0 is chosen uniformly over the state space $E = \{0,1\}^l$, the normalized first hitting time $\bar{T}/(l \ln l)$ converges in probability to 1 as $l \to \infty$. For $l \gg 1$, the expectation of $\bar{T}$ can be expanded as:

$$\mathbb{E}[\bar{T}] \stackrel{l \gg 1}{\simeq} l(\ln l + C - \ln 2)$$

up to terms of order $o(l)$, while the normalized standard deviation obeys:

$$\bar{\sigma} := \frac{\sqrt{\mathbb{E}[\bar{T}^2] - \mathbb{E}[\bar{T}]^2}}{\mathbb{E}[\bar{T}]} \stackrel{l \gg 1}{\simeq} \frac{\pi}{\sqrt{6}(\ln l + C - \ln 2)}.$$

The convergence rate of the normalized standard deviation towards 0 is about $(\ln l)^{-1}$, which proves rigorously the convergence in probability of $\bar{T}/(l \ln l)$ towards 1, but this convergence rate is rather slow for practical purposes. The normalized standard deviation is actually of order 1 for a large band of values of l. In order to get $\bar{\sigma} < 20\%$ (resp. $\bar{\sigma} < 10\%$), it is necessary that $l > 7.10^2$ (resp. $l > 4.10^5$). We can actually be much more precise and give a very accurate description for the statistical distribution of $\bar{T}$ in the asymptotic framework $l \to \infty$.

PROPOSITION **4.2**
Assuming that the starting point X^0 is chosen uniformly over the state space $E = \{0,1\}^l$, the normalized first hitting time $(\bar{T} - l \ln l)/l$ converges in distribution as $l \to \infty$ to $-\ln 2 - \ln Z$, where Z is a random variable with an exponential density $p(z) = \mathbb{1}_{z \geq 0} e^{-z}$.

Note that we get back the results of Proposition 4.1 since $\mathbb{E}[-\ln Z] = C$ and $\mathbb{E}[(\ln Z)^2] = \pi^2/6 + C^2$, where C is the Euler's constant.

COROLLARY **4.3**
Assume that we deal with a set of N processes $\bar{X}_i$ which evolve independently. The global first hitting time $\bar{T}$ is the infimum of the N independent hitting times.
1. If N is fixed (independent of l), then there exists a random variable Z_N with probability density $p_N(z) = N(1-e^{-z})^{N-1} e^{-z} \mathbb{1}_{z \geq 0}$ and mean $\sum_{j=1}^N j^{-1}$ such that the global first hitting time $\bar{T}$ satisfies:

$$\frac{\bar{T} - l \ln l}{l} \stackrel{l \to \infty}{\longrightarrow} -\ln 2 - \ln Z_N. \tag{22}$$

2. If $N = N(l)$ grows sub-exponentially to infinity as $l \to \infty$ (more exactly if $N(l) \to \infty$ and $\frac{\ln \ln N(l)}{\ln l / \ln \ln l} \to 0$ as $l \to \infty$), then

$$\frac{\bar{T} - l \ln l + l \ln \ln N(l)}{l} \stackrel{l \to \infty}{\longrightarrow} -\ln 2. \tag{23}$$

with probability 1.

This corollary means in particular that the gain when using N independent processes is very weak, of the order of $\ln \ln N$. To get a significant gain, it is necessary to deal with an exponentially large (as a function of l) number of processes.

Let us now consider the process X starting from X^0 whose evolution is governed in the following way. At each time n, you build a string $\tilde{X}(n-1)$ by inverting each bit of $X(n-1)$ independently with probability c/l. If $d(\tilde{X}(n-1), X^f) < d(X(n-1), X^f)$, then $X(n) = \tilde{X}(n-1)$, otherwise $X(n) = X(n-1)$. The transition matrix of the chain $X(n)$ is consequently:

$$P_{X,X'} = \begin{cases} P_{X,X} & \text{if } d(X,X') = 0, \\ (cl^{-1})^{d(X,X')}\left(1 - cl^{-1}\right)^{l-d(X,X')} & \text{if } d(X',X^f) < d(X,X^f), \\ 0 & \text{otherwise.} \end{cases}$$

with $P_{X,X} = 1 - \sum_{k=0}^{d(X,X^f)} \sum_{k'=0}^{l-d(X,X^f)} \mathbb{I}_{k<k'} C^k_{d(X,X^f)} C^{k'}_{l-d(X,X^f)} (cl^{-1})^{k+k'} \left(1 - cl^{-1}\right)^{l-k-k'}$.

PROPOSITION **4.4**
Assuming that the starting point X^0 is chosen uniformly over the state space $\{0,1\}^l$, the normalized first hitting time $T/(l \ln l)$ converges in probability to e^c/c as $l \to \infty$.
For $l \gg 1$, the expectation of T can be expanded as:

$$\mathbb{E}[T] \overset{l \gg 1}{\simeq} l \left(\frac{e^c}{c} \ln l + R(c)\right)$$

up to terms of order $o(l)$. The function $c \mapsto R(c)$ is plotted in Figure 2. In particular $R(1) \simeq -1.90$.

Note that the function $c \mapsto c^{-1}e^c$ is minimal at $c = 1$. The mutation rate l^{-1} is therefore optimal.

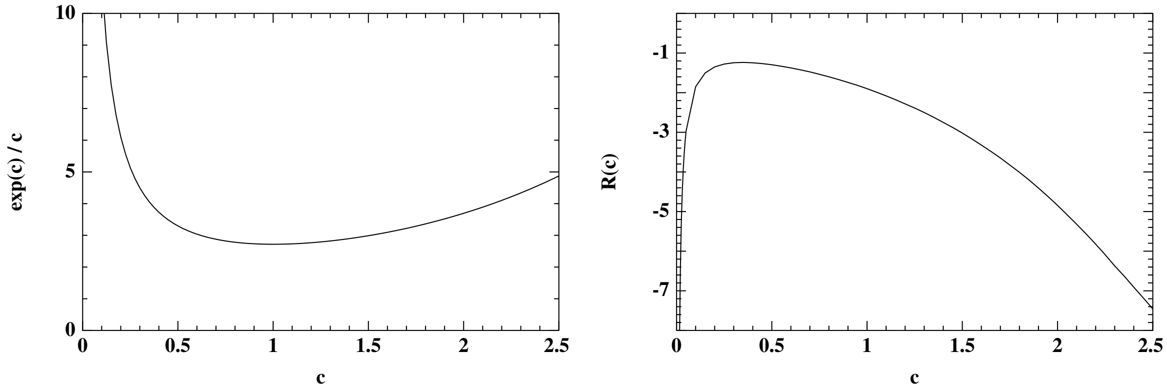

FIG 2 : *Functions $c \mapsto c^{-1}e^c$ and $c \mapsto R(c)$*

We can actually be much more precise and give a very accurate description for the statistical distribution of T in the asymptotic framework $l \to \infty$.

PROPOSITION **4.5**
Assuming that the starting point X^0 is chosen uniformly over the state space $E = \{0,1\}^l$, the normalized first hitting time $(T - c^{-1}e^c l \ln l)/l$ converges in distribution as $l \to \infty$:

$$\frac{T - c^{-1}e^c l \ln l}{l} \overset{l \to \infty}{\longrightarrow} -\frac{e^c}{c}(\ln Z + C) + R(c),$$

where Z is a random variable with an exponential density $p(z) = \mathbb{I}_{z \geq 0} e^{-z}$, $C \simeq 0.58$ is the Euler's constant and $R(c)$ is plotted in Figure 2.

Corollary 4.6

Assume that we deal with a set of N processes X_i which evolve independently. The global first hitting time T is the infimum of the N independent hitting times.
1. If N is fixed (independent of l), then there exists a random variable Z_N with probability density $p_N(z) = N(1-e^{-z})^{N-1}e^{-z}\mathbb{1}_{z\geq 0}$ and mean $\sum_{j=1}^{N} j^{-1}$ such that the global first hitting time $\bar{T}$ satisfies:

$$\frac{T - c^{-1}e^c l \ln l}{l} \overset{l\to\infty}{\longrightarrow} R(c) - \frac{e^c}{c}(\ln Z_N + C). \tag{24}$$

2. If $N = N(l)$ grows sub-exponentially to infinity as $l \to \infty$ (more exactly if $N(l) \to \infty$ and $\frac{\ln\ln N(l)}{\ln l/\ln\ln l} \to 0$ as $l \to \infty$), then

$$\frac{T - c^{-1}e^c l \ln l + c^{-1}e^c l \ln\ln N(l)}{l} \overset{l\to\infty}{\longrightarrow} R(c) - \frac{e^c}{c}C. \tag{25}$$

with probability 1.

4.2 Study of the chain $\bar{X}$

4.2.1 Proof of Proposition 4.1

Let us introduce the auxiliary process $Z(n)$ with state space $\{0,...,l\}$ defined by:

$$Z(n) = d(\bar{X}(n), X^f). \tag{26}$$

Z is a Markov chain starting from $Z(0) = d(X^0, X^f)$ whose jumps are governed by the stationary transition matrix $\bar{Q}$:

$$\bar{Q} = \begin{pmatrix} 1 & 0 & \dots & \dots & \dots & 0 \\ \frac{1}{l} & 1-\frac{1}{l} & 0 & \dots & \dots & 0 \\ 0 & \frac{2}{l} & 1-\frac{2}{l} & 0 & \dots & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & 1-\frac{1}{l} & \frac{1}{l} & 0 \\ 0 & \dots & \dots & 0 & 1 & 0 \end{pmatrix}$$

Obviously the first hitting time of the point 0 by $Z(n)$ corresponds to the first hitting time $\bar{T}$ of the point X^f by the chain $\bar{X}(n)$. Starting from k, the first hitting time of the point 0 by the chain Z is:

$$\bar{T} = \sum_{j=1}^{k} \tau_j, \tag{27}$$

where the τ_j's are the times necessary for the chain to jump from j to $j-1$. By the strong Markov property, the τ_j's are independent random variables with respective distributions ($(Q_{j,j})^n$ is the n-th power of the coefficient $Q_{j,j}$):

$$\mathbb{P}(\tau_j = t) = (1-Q_{j,j})(Q_{j,j})^{t-1} = \frac{j}{l}\left(1-\frac{j}{l}\right)^{t-1} \quad \text{for } t \in \mathbb{N}^*.$$

In particular their first and second order moments are:

$$\mathbb{E}[\tau_j] = \frac{l}{j}, \qquad \mathbb{E}[\tau_j^2] = \frac{2l^2 - jl}{j^2}.$$

We can then deduce the expectation of $\bar{T}$:

$$\mathbb{E}_k[\bar{T}] = \sum_{j=1}^{k} \mathbb{E}[\tau_j] = l \sum_{j=1}^{k} j^{-1} \stackrel{k \gg 1}{\simeq} l(\ln k + C),$$

up to terms of order $o(l)$. If the chain $\bar{X}$ starts from a point X^0 which is chosen uniformly over the state space $\{0,1\}^l$, then the chain Z starts from the point k_0 which obeys the distribution μ:

$$\mu(k_0) = 2^{-l} C_l^{k_0}. \tag{28}$$

and then:

$$\mathbb{E}_\mu[\bar{T}] = \sum_{k=0}^{l} \mu(k) \mathbb{E}_k[\bar{T}].$$

Using Stirling's formula [1](see Appendix):

$$\mathbb{E}_\mu[\bar{T}] \stackrel{l \gg 1}{\simeq} l \left(\ln l + C + \sqrt{l} \int_0^1 e^{-lf(a)} \frac{\ln(a)}{\sqrt{2\pi a(1-a)}} da \right),$$

where $f(a) = \ln 2 + a \ln(a) + (1-a)\ln(1-a)$. Since f is minimal at $a = 1/2$, $f(1/2) = 0$, $f'(1/2) = 0$ and $f''(1/2) = 4$, we get by using Laplace technique that:

$$\mathbb{E}_\mu[\bar{T}] \stackrel{l \gg 1}{\simeq} l\,(\ln l + C - \ln 2)$$

up to terms of order $o(l)$. The calculation of the second order moment of $\bar{T}$ can be carried out by the same way:

$$\begin{aligned}
\mathbb{E}_k[\bar{T}^2] &= \sum_{j=1}^{k} \mathbb{E}[\tau_j^2] + \sum_{i \neq j=1}^{k} \mathbb{E}[\tau_i]\mathbb{E}[\tau_j] = l^2 \left(\sum_{j=1}^{k} j^{-2} \right) - l \left(\sum_{j=1}^{k} j^{-1} \right) + l^2 \left(\sum_{j=1}^{k} j^{-1} \right)^2 \\
&\stackrel{l \gg 1}{\simeq} l^2 \left(\frac{\pi^2}{6} + C^2 + (\ln k)^2 + 2C \ln k \right) - l(\ln k + C).
\end{aligned}$$

So the normalized standard deviation defined by:

$$\bar{\sigma}_k := \frac{\sqrt{\mathbb{E}_k[\bar{T}^2] - \mathbb{E}_k[\bar{T}]^2}}{\mathbb{E}_k[\bar{T}]}$$

satisfies:

$$\bar{\sigma}_k \stackrel{l \gg 1}{\simeq} \frac{\pi}{\sqrt{6}(\ln k + C)}.$$

Starting from a random point point X^0 chosen uniformly over the state space E, we have using (28):

$$\mathbb{E}_\mu[\bar{T}^2] \stackrel{l \gg 1}{\simeq} l^2 \left\{ (\ln l)^2 + (2C - 2\ln 2) \ln l + \left(\frac{\pi^2}{6} + C^2 + (\ln 2)^2 - 2C \ln 2 \right) \right\}$$

up to terms of order $o(l^2)$. As a consequence the normalized standard deviation defined by:

$$\bar{\sigma}_\mu := \frac{\sqrt{\mathbb{E}_\mu[\bar{T}^2] - \mathbb{E}_\mu[\bar{T}]^2}}{\mathbb{E}_\mu[\bar{T}]}$$

satisfies:

$$\bar{\sigma}_\mu \overset{l \gg 1}{\simeq} \frac{\pi}{\sqrt{6}(\ln l + C - \ln 2)}$$

which goes to 0 as $l \to \infty$ at rate $(\ln l)^{-1}$, which implies that $\bar{T}/(l \ln l)$ converges in probability to 1.

Remark **4.7**
From Eq.(27) we can deduce a closed form expression for the histogram of $\bar{T}$, which holds true for any l. Unfortunately these expressions are all the more complicated as l is large.

Proposition **4.8**
Let us denote $p_j(t) = \frac{j}{l}\left(1 - \frac{j}{l}\right)^{t-1}$ *for* $t \in \mathbb{N}^*$ *and* $j = 1, ..., l$.
1. Assuming that the starting point X^0 *is such that* $d(X^0, X^f) = k$, *then:*

$$\mathbb{P}_k(\bar{T} = t) = p_1 \star ... \star p_k(t),$$

where $p \star q(t) = \sum_{t'=1}^{t-1} p(t')q(t-t')$.
2. Assuming that the starting point X^0 *is chosen uniformly over the state space* $E = \{0,1\}^l$, *then:*

$$\mathbb{P}(\bar{T} = t) = \sum_{k=1}^{l} 2^{-l} C_l^k \mathbb{P}_k(\bar{T} = t).$$

In Figure 3 are plotted the theoretical histograms of $\bar{T}$ under the assumption that the starting point X^0 is chosen uniformly over the state space E for different values of l.

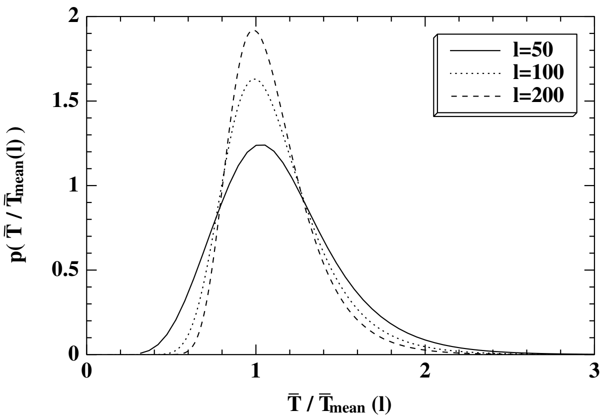

Fig 3 : *Theoretical histograms of* $\bar{T}$. *The scales are normalized by the quantity* $\bar{T}_{mean}(l) := l(\ln l + C - \ln 2)$.

4.2.2 Proof of Proposition 4.2

Let us denote by T_r^l the random variable which depends on l:

$$T_r^l := \frac{\bar{T} - l \ln l + l \ln 2}{l}.$$

Let α be a real number larger than -1. From the representation (27), the Laplace transform of T_r^l is:

$$\mathbb{E}_k[e^{-\alpha T_r^l}] = e^{\alpha(\ln 2 - \ln l)} \mathbb{E}\left[e^{-\alpha(\tau_1+\ldots+\tau_k)/l}\right].$$

Since the τ_j's are independent:

$$\mathbb{E}_k[e^{-\alpha T_r^l}] = e^{-\alpha(\ln 2 - \ln l)} \prod_{j=1}^{k} \mathbb{E}\left[e^{-\alpha\tau_j/l}\right].$$

Computing the Laplace transforms of the variables τ_j's whose distributions are geometric is easy:

$$\mathbb{E}\left[e^{-\alpha'\tau_j}\right] = e^{-\alpha'}\left(\frac{l}{j}(1-e^{-\alpha'}) + e^{-\alpha'}\right)^{-1}$$

so that (with $\alpha' = \alpha/l$, we have $1 - e^{-\alpha'} \simeq \alpha/l$):

$$\mathbb{E}_k[e^{-\alpha T_r^l}] = e^{-\alpha(\ln 2 - \ln l)} e^{-k\alpha/l} \left[\prod_{j=1}^{k}\left(1 + \frac{\alpha}{j} - \frac{\alpha}{l}\right)\right]^{-1} \tag{29}$$

up to terms of order $O(l^{-1})$. We first deal with the product (multiplied by $k!$):

$$\begin{aligned}
\prod_{j=1}^{k}\left(j + \alpha(1 - \frac{j}{l})\right) &= \prod_{j=1}^{k}(j+\alpha)\prod_{j=1}^{k}\left(1 - \frac{1}{l}\frac{\alpha j}{j+\alpha}\right) \\
&= \frac{\Gamma(k+\alpha+1)}{\Gamma(\alpha+1)} \exp\left(\sum_{j=1}^{k}\ln\left(1 - \frac{\alpha}{l}\frac{j}{j+\alpha}\right)\right)
\end{aligned}$$

where Γ is the Euler's Gamma function $\Gamma(x) := \int_0^\infty t^{x-1}e^{-t}dt$. In the asymptotic $l \to \infty$ the sum inside the exponential is equivalent to:

$$\sum_{j=1}^{k}\ln\left(1 - \frac{\alpha}{l}\frac{j}{j+\alpha}\right) \overset{l\gg 1}{\simeq} -\frac{\alpha}{l}\sum_{j=1}^{k}\frac{j}{j+\alpha} \overset{l\gg 1}{\simeq} -\frac{k}{l}\alpha,$$

up to terms of order $O(kl^{-2})$. Substitution into (29) then yields that:

$$\mathbb{E}_k[e^{-\alpha T_r^l}] \overset{l\to\infty}{\longrightarrow} 2^{-\alpha} l^{-\alpha} \frac{\Gamma(k+1)\Gamma(\alpha+1)}{\Gamma(k+\alpha+1)}.$$

Since $\mathbb{E}_\mu[.] = \sum_{k=0}^{l} 2^{-l} C_l^k \mathbb{E}_k[.]$ and $\Gamma(k+1)/\Gamma(k+\alpha+1) \simeq k^{-\alpha}$ for $k \gg 1$, we obtain:

$$\mathbb{E}_\mu[e^{-\alpha T_r^l}] \overset{l\to\infty}{\longrightarrow} \Gamma(\alpha+1) \tag{30}$$

for any $\alpha > 0$. This proves the convergence in distribution of T_r^l as $l \to \infty$ to the unique variable T_r such that $\mathbb{E}[e^{-\alpha T_r}] = \Gamma(\alpha+1)$. It is easy to identify this variable in terms of a random variable Z with exponential distribution and mean 1: $T_r = -\ln Z$.

4.2.3 Proof of Corollary 4.3

Since $\bar{T} = \inf_{i=1,\ldots,N}\bar{T}_i$ and the $\bar{T}_i$ are independent and obey the same distribution:

$$\mathbb{P}\left(\bar{T} \geq \bar{t}\right) = \mathbb{P}\left(\bar{T}_1 \geq \bar{t}\right)^N. \tag{31}$$

Choosing $\bar{t} = l\ln l - l\ln 2 + lt$ and using Proposition 4.2:

$$\mathbb{P}\left(\bar{T}_1 \geq \bar{t}\right) \stackrel{l\to\infty}{\longrightarrow} \mathbb{P}(-\ln Z_1 \geq t) = \mathbb{P}(Z_1 \leq e^{-t})$$

where Z_1 is an exponential variable with mean 1, so that:

$$\mathbb{P}\left(\bar{T}_1 \geq \bar{t}\right) \stackrel{l\to\infty}{\longrightarrow} 1 - \exp(-e^{-t}).$$

If N is independent of l, substituting into (31) establishes:

$$\mathbb{P}\left(\frac{\bar{T} - l\ln l + l\ln 2}{l} \geq t\right) \stackrel{l\to\infty}{\longrightarrow} \left(1 - \exp(-e^{-t})\right)^N.$$

Therefore the convergence (22) holds true if Z_N is the random variable whose repartition function is:

$$\mathbb{P}(Z_N \leq z) = \left(1 - \exp(-z)\right)^N.$$

It means that Z_N obeys the statistical distribution of the supremum of N independent random variables with exponential density and mean 1. This proves the first statement of the corollary.

Let us turn to the case $N = N(l)$. Denoting

$$T_{r,1}^l := \frac{\bar{T}_1 - l\ln l + l\ln 2}{l}$$

and following carefully the proof of Proposition 4.2 and the error terms, it can be shown that (30) still holds true for large α, but much smaller than l:

$$\mathbb{E}[e^{-\alpha T_{r,1}^l}] = \Gamma(\alpha+1) \text{ if } 1 \ll \alpha \ll l.$$

Using a Tauberian theorem (Theorem XII-5-4-[21] which establishes connections between the Laplace transform and the repartition function of a random variable), we then get that, if $t \ll \Gamma^{-1}(l)$:

$$\mathbb{P}\left(T_{r,1}^l \leq t\right) \stackrel{l\gg 1}{\simeq} \exp\left(-e^{-t}\right).$$

Since $\Gamma^{-1}(l) \simeq \ln l/\ln\ln l$, substituting into (31) establishes that, for any $t \ll \ln l/\ln\ln l$:

$$\mathbb{P}\left(\frac{\bar{T} - l\ln l + l\ln 2}{l} \geq t\right) \stackrel{l\gg 1}{\simeq} \left(1 - \exp\left(-e^{-t}\right)\right)^{N(l)}.$$

Let $\delta > 0$. Choosing $t = -\ln[(1+\delta)\ln N(l)]$ (which is smaller than $\ln l/\ln\ln l$ by assumption), we have $e^{-t} = N(l)^{1+\delta}$ so that:

$$\mathbb{P}\left(\frac{\bar{T} - l\ln l + l\ln 2}{l} \geq t\right) \stackrel{l\gg 1}{\simeq} \left(1 - \frac{1}{N(l)^{1+\delta}}\right)^{N(l)} \stackrel{l\gg 1}{\simeq} 1$$

since $N(l) \to \infty$. Choosing $t = -\ln[(1-\delta)\ln N(l)]$, we have $e^{-t} = N(l)^{1-\delta}$ so that:

$$\mathbb{P}\left(\frac{\bar{T} - l\ln l + l\ln 2}{l} \geq t\right) \stackrel{l\gg 1}{\simeq} \left(1 - \frac{1}{N(l)^{1-\delta}}\right)^{N(l)} \stackrel{l\gg 1}{\simeq} 0$$

which completes the proof of the Corollary.

4.3 Study of the chain X

4.3.1 Proof of Proposition 4.4

The strategy of the proof is similar to that of Proposition 4.1. However the details are a little more complicated because the transition matrix of the chain $X(n)$ is more complicated than the one of the chain $\bar{X}(n)$. Let us introduce the auxiliary process $Z(n)$ with state space $\{0,...,l\}$ defined by:

$$Z(n) = d(X(n), X^f). \tag{32}$$

Z is a Markov chain starting from $Z(0) = d(X^0, X^f)$. The coefficients of the corresponding transition matrix Q are given by:

$$Q_{j,j-j'} = \begin{cases} \displaystyle\sum_{k=0}^{(l-j)\wedge(j-j')} C_j^{k+j'} C_{l-j}^k (cl^{-1})^{2k+j'}(1-cl^{-1})^{l-2k-j'} & \text{if } j' \in \{1,...,j\}, \\ Q_{j,j} & \text{if } j' = 0, \\ 0 & \text{if } j' \notin \{0,...,j\}. \end{cases} \tag{33}$$

with $Q_{j,j} = 1 - \sum_{j'=1}^{j} Q_{j,j-j'}$.
Let us assume that the chain Z arrives at some point j at some time t. We then denote by τ_j the waiting time for the next jump, and by N_j the length of this jump. By the strong Markov property, the random variables τ_j and N_j are independent and their respective statistical distributions are given by:

$$\begin{aligned} \mathbb{P}(\tau_j = t) &= (1 - Q_{j,j})Q_{j,j}^{t-1} \text{ for } t \in \mathbb{N}^*, & (34) \\ \mathbb{P}(N_j = n) &= \frac{Q_{j,j-n}}{1 - Q_{j,j}} \text{ for } n = 1,...,j. & (35) \end{aligned}$$

The statistical distribution of the first hitting time T can be expressed in terms of the variables N_j and τ_j. In particular, conditioning with respect to the first jump yields:

$$\mathbb{P}_j\left(T = t\right) = \sum_{t'=1}^{t}\sum_{n=1}^{j} \mathbb{P}_j\left(T = t | \tau_j = t', N_j = n\right) \mathbb{P}(\tau_j = t')\mathbb{P}(N_j = n).$$

Since the weak Markov property insures that:

$$\mathbb{P}_j\left(T = t | \tau_j = t', N_j = n\right) = \mathbb{P}_{j-n}\left(T = t - t'\right),$$

we get the inductive relation:

$$\mathbb{P}_j\left(T = t\right) = \sum_{t'=1}^{t}\sum_{n=1}^{j} \mathbb{P}_{j-n}\left(T = t - t'\right) \mathbb{P}(\tau_j = t')\mathbb{P}(N_j = n). \tag{36}$$

Indeed, iterating this formula establishes the following closed-form expression which holds true for any k and l:

$$\mathbb{P}_k\left(T = t\right) = \sum_{J=1}^{k} \sum_{n_1+...+n_J=k} \prod_{i=1}^{J} \mathbb{P}(N_{k-n_1-...-n_{i-1}} = n_i) p_k \star p_{k-n_1} \star ... \star p_{k-n_k-...n_J}(t). \tag{37}$$

where $p \star q(t) = \sum_{t'=1}^{t-1} p(t')q(t-t')$. If we assume that the starting point X^0 is chosen uniformly over the state space $\{0,1\}^l$, then we can express the histogram of T as follows:

$$\mathbb{P}\left(T=t\right) = \sum_{k=0}^{l} 2^{-l} C_l^k \mathbb{P}_k\left(T=t\right). \tag{38}$$

We now turn to the study of the expectation of T in order to exhibit its behavior for large values of l. Multiplying (36) by t and summing over t yields the inductive relation:

$$\mathbb{E}_j[T] = \mathbb{E}[\tau_j] + \sum_{n=1}^{j} \mathbb{E}_{j-n}[T]\mathbb{P}(N_j = n). \tag{39}$$

Iterating with respect to t establishes the following closed-form expression which holds true for any k and l:

$$\mathbb{E}_k[T] = \sum_{J=1}^{k} \sum_{n_1+\ldots+n_J=k} \left(\sum_{i=0}^{J-1} \mathbb{E}[\tau_{k-n_1-\ldots-n_i}] \right) \prod_{i=1}^{J} \mathbb{P}(N_{k-n_1-\ldots-n_{i-1}} = n_i). \tag{40}$$

In the asymptotic framework $l \gg 1$, this complicated expression gets much simpler.
Step 1. There exist functions $f_{\tau,c}$ and $f_{N,c}$ which do not depend on l such that, for $l \gg 1$: $\mathbb{E}[\tau_j] = f_{\tau,c}(j/l)$ and $\mathbb{E}[N_j] = f_{N,c}(j/l)$.

$$\frac{1}{f_{\tau,c}(a)} = 1 - e^{-c} \sum_{k=0}^{\infty} \frac{(c(1-a))^k}{k!} \left(\sum_{j=0}^{k} \frac{(ca)^j}{j!} \right), \tag{41}$$

$$f_{N,c}(a) = c(2a-1) + \frac{c^2(1-a)a\left(\sum_{k=0}^{\infty} \frac{(c^2(1-a)a)^k}{k!(k+1)!}\right) + ca\left(\sum_{k=0}^{\infty} \frac{(c^2(1-a)a)^k}{k!^2}\right)}{e^c f_{c,\tau}(a)}. \tag{42}$$

This step results from the long but straightforward calculations of the expectations of N_j and τ_j whose statistical distributions are given by (33)-(35).

Step 2. If $a \in (0,1]$, then we have for $l \gg 1$: $\mathbb{E}_{[al]}[T] = \sum_{j=1}^{[al]} \frac{f_{\tau,c}(j/l)}{f_{N,c}(j/l)}$.
Let $0 < \delta \ll 1$, such that $1 \ll \delta l \ll l$. By Step 1, we get that over the interval $j \in \{[al], \ldots, [al] - [\delta l]\}$, the statistical distributions of N_j and τ_j are almost identical to the ones of $N_{[al]}$ and $\tau_{[al]}$. Using the inductive relation (39):

$$\mathbb{E}_{[al]}[T] \simeq \sum_{k=1}^{\infty} \sum_{n_1+\ldots+n_{k-1} \leq [\delta l] < n_1+\ldots+n_k} \prod_{j=1}^{k} \mathbb{P}(N_{[al]} = n_j) \left(k\mathbb{E}[\tau_{[al]}] + \mathbb{E}_{[al]-n_1+\ldots+n_k}[T] \right).$$

By the renewal theory (see Appendix and (43)):

$$\frac{1}{[\delta l]} \sum_{k=1}^{\infty} k\mathbb{P}\left(N_1 + \ldots + N_{k-1} \leq [\delta l] < N_1 + \ldots + N_k\right) \overset{l\to\infty}{\simeq} \frac{1}{\mathbb{E}[N_{[al]}]} = \frac{1}{f_{N,c}(a)},$$

which implies that:

$$\mathbb{E}_{[al]}[T] \simeq \frac{f_{\tau,c}(a)}{f_{N,c}(a)}[\delta l] + \mathbb{E}_{[al]-[\delta l]}[T].$$

Iterating this process completes the proof of Step 2.
Step 3. Computation of $\mathbb{E}_\mu[T]$ *for* $l \gg 1$.
The function $a \mapsto \frac{f_{c,\tau}(a)}{f_{c,N}(a)}$ has a singularity near the point $a = 0$ of the type a^{-1}. Nevertheless $g_c := a \mapsto \frac{f_{c,\tau}(a)}{f_{c,N}(a)} - \frac{e^c}{ca}$ is a smooth function. As a consequence:

$$\mathbb{E}_{[al]}[T] = \sum_{j=1}^{[al]} \frac{e^c l}{cj} + \sum_{j=1}^{[al]} g_c(j/l)$$

which yields:

$$\mathbb{E}_{[al]}[T] = l\left(c^{-1}e^c \ln l + c^{-1}e^c \ln a + c^{-1}e^c C + G_c(a)\right)$$

where G_c is the integral of g_c:

$$G_c(a) = \int_0^a g_c(s)ds.$$

As stated in the proof of Proposition 4.1, if the chain X starts from a point X^0 which is chosen uniformly over the state space $\{0,1\}^l$, then the chain Z starts from the point k_0 which obeys the distribution (28) so that:

$$\mathbb{E}_\mu[T] = \sum_{k=0}^{l} \mu(k)\mathbb{E}_k[T].$$

Using Stirling's formula [1] to expand $\mu(k)$:

$$\mathbb{E}_\mu[T] = l\left(c^{-1}e^c \ln l + c^{-1}e^c C + eC + \sqrt{l}\int_0^1 e^{-lf(a)} \frac{c^{-1}e^c \ln a + G_c(a)}{\sqrt{2\pi a(1-a)}} da\right)$$

where $f(a) = \ln 2 + a\ln(a) + (1-a)\ln(1-a)$. Using Laplace technique to compute the integral then establish the result by choosing $R(c) = -c^{-1}e^c \ln 2 + c^{-1}e^c C + G_c(1/2)$.
Step 4. Conclusion.
Adopting the same strategy as in the proof of Proposition 4.1, long but straightforward calculations establish that:

$$\frac{\sqrt{\mathbb{E}[T^2] - \mathbb{E}[T]^2}}{\mathbb{E}[T]} \xrightarrow{l\to\infty} 0$$

at rate $(\ln l)^{-1}$, which implies that $T/(l \ln l)$ converges in probability to e.

Remark **4.9**
In the above proof we have established a closed form expression (37-38) for the histogram of T which holds true for any l. This exact expression is unfortunately not very useful, because very complicated. Nevertheless a computer can deal with it whenever l does not exceed a few hundreds. In Figure 4 are plotted the theoretical histograms of $\bar{T}$ for different values of l.

4.3.2 Proofs of Proposition 4.5 and Corollary 4.6

Although the calculations are a little more complicated, the proofs are exactly the same as the ones of Proposition 4.2 and Corollary 4.3 respectively.

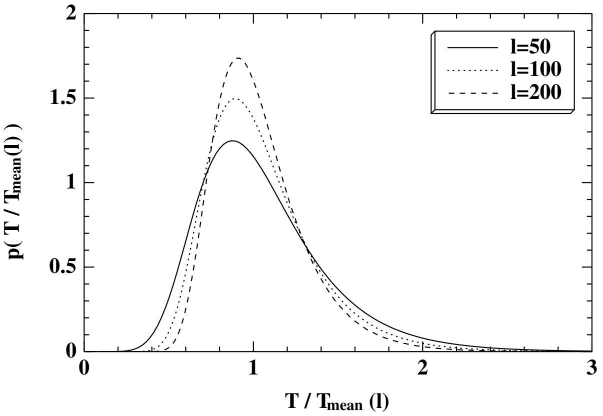

FIG 4 : *Theoretical histograms of T in the case $c = 1$, which corresponds to the (1,1)-ES with mutation rate $1/l$. The scales are normalized by the quantity $T_{mean}(l) := l(e \ln l + R(1))$, $R(1) \simeq -1.90$.*

5 Discussion and Future Directions

A first outcome of the results of this paper is a clear advantage of the *1-bit-flip* mutation over the $\frac{c}{l}$ mutation, both in the absence of selection, and when using the (1+1)-ES selection on the OneMax problem.
However, this paper does not claim that the 1-bit-mutation should be preferred to the c/l mutation, as it is clear that the 1-bit-mutation with non linear functions cannot avoid any local optimum. Rather, we point out the fact that a balance should be found between the ability of a mutation to escape from local optima, and its efficiency in terms of convergence speed. The results of this paper suggest that the positivity [4] property of mutations can result in a slower convergence: The c/l-mutations are able to escape from any local optima as they are positive operators; but, on both flat and linear landscapes, they are slower than the simple one-bit-flip mutation.

Another point to consider, when trying to analyze these theoretical results in the light of practical work, is that the hitting times derived for both mutations were computed in terms of number of mutations performed before reaching the target point of the search space. In the framework of Evolutionary Optimization, a common practice is to consider that the main source of CPU consumption of an Evolutionary Algorithm comes from the computations of the fitnesses. In that perspective (also considered in [18]), a mutation that does not flip any bits does not cost any supplementary fitness evaluation. As was recalled in section 2, this happens e^{-c} of the times with the $\frac{c}{l}$ mutation. So considering the number of fitness computation instead of the number of steps, the *1-bit-flip* and the $\frac{c}{l}$ mutations require the same effort to reach their goal in the absence of selection. On the other hand, there is still a factor $\frac{e^c-1}{c}$ greater than 1 in favor of the $\frac{c}{l}$ mutation in the case of the (1+1)-ES on the OneMax problem. Note that this factor goes to 1 when c goes to 0: the $\frac{c}{l}$ mutation then becomes as slow as the *1-bit-flip* mutation. Indeed, the process spends e^c of the time doing nothing, ce^c mutating one bit, and the remaining time $(1 - e^c - ce^c)$ mutating more than one bit. Neglecting the steps where nothing happens, the process spends $\frac{ce^{-c}}{1-e^{-c}} \simeq 1 - \frac{c}{2}$ $[c \to 0]$ mutating exactly one bit, and therefore looks like the *1-bit-flip* mutation. But this is true only in terms of fitness evaluations, as most of its effort is then devoted to generate random numbers and not doing anything else.

[4] a mutation that links any two points of the space with non-zero probability

As already mentioned in section 2, some practitioners of Evolutionary Algorithms actually use inhomogeneous decreasing schemes for the mutation rate (see also the theoretical results of [13]). Indeed, there might be some benefit in starting with large mutation and decreasing the mutation strength to the *1-bit-flip* mutation in the final generations. On the other hand, if the landscape can be locally seen as a OneMax-like problem to the nearest local optima, then, the *1-bit-flip* mutation will be quicker than the $\frac{c}{l}$ mutation to reach that local optima.
Anther way to pursue this work in the same line would be to study a mixed mutation, identical to the $\frac{c}{l}$ mutation, but ensuring that at least one bit of the bitstring is flipped. And of course, the ultimate step along that line would be to study some self-adaptive mutation strength, as the one proposed in [34, 7].

A natural direction for future work is to try to extend the results of the (1+1)-ES algorithm to other types of functions or algorithms. First, it can be noted that the results can be easily applied to any linear or site-wise optimizable binary function (as flipping one bit from 0 to 1 always improves the fitness). Second, note that the influence of a population also requires more investigation.
On the other hand, the $\frac{c}{l}$ mutation certainly is the most widely used mutation operator on the binary search space. But it should be noted that, in the binary framework, one generally uses a GA-like algorithm, involving another type of selection, and considering the crossover operator to be the main operator: the mutation operator is considered a background operator in the GA literature [27, 22]. Thus, any generalization of the tendencies proved in the (1+1)-ES framework is hazardous.

Finally, another question, which is still far ahead, is whether it is possible to tell which is the right mutation for a given landscape: the exploration properties alone (their behavior inside the (1,1)-ES) are certainly not enough, since their effect is biased in the complete algorithm, e.g. by the selection operator. Turned the other way round, that question becomes: Is it possible to design a selection scheme to ensure a maximal exploitation of the potential of a given mutation?

References

[1] M. Abramowitz and I. Stegun. *Handbook of mathematical functions.* Dover publications, New York, 1965.

[2] D. H. Ackley. *A connectionnist Machine for Genetic Hill-climbing.* Kluwer, Boston, 1987.

[3] Alexandru Agapie. Genetic algorithms: minimal conditions for convergence. In J.-K. Hao, E. Lutton, E. Ronald, M. Schoenauer, and D. Snyers, editors, *Artificial Evolution'97*, LNCS. Springer Verlag, October 1997.

[4] J. Antonisse. A new interpretation of schema notation that overturns the binary encoding constraint. In J. D. Schaffer, editor, *Proceedings of the* 3^{rd} *International Conference on Genetic Algorithms*, pages 86–91. Morgan Kaufmann, June 1989.

[5] T. Bäck. The interaction of mutation rate, selection, and self-adaptation in genetic algorithm. In R. Manner and B. Manderick, editors, *Proceedings of the* 2^{nd} *Conference on Parallel Problems Solving from Nature*, pages 85–94. Morgan Kaufmann, 1992.

[6] T. Bäck. Optimal Mutation Rate in Genetic Search. In S. Forrest, editor, *Proceedings of the 5^{th} International Conference on Genetic Algorithms*, pages 2–8. Morgan Kaufmann, 1993.

[7] T. Bäck and M. Schütz. Intelligent mutation rate control in canonical gas. In Z. W. Ras and M. Michalewicz, editors, *Foundation of Intelligent Systems 9th International Symposium, ISMIS '96*, pages 158–167. Springer Verlag, 1996.

[8] H.-G. Beyer. Toward a theory of evolution strategies: Some asymptotical results for the $(1,+\lambda)$-theory. *Evolutionary Computation*, 1(2):165–188, 1993.

[9] H.-G. Beyer. Toward a theory of evolution strategies: The (μ, λ)-theory. *Evolutionary Computation*, 2(4):381–407, 1994.

[10] H.-G. Beyer. Toward a theory of evolution strategies: On the benefit of sex - the $(\mu/\mu, \lambda)$-theory. *Evolutionary Computation*, 3(1):81–111, 1995.

[11] H.-G. Beyer. Toward a theory of evolution strategies: Self-adaptation. *Evolutionary Computation*, 3(3):311–347, 1995.

[12] R. Cerf. *Une théorie assymptotique des algorithmes génétiques.* PhD thesis, Université de Montpellier II, March 1994.

[13] R. Cerf. An asymptotic theory of genetic algorithms. In J.-M. Alliot, E. Lutton, E. Ronald, M. Schoenauer, and D. Snyers, editors, *Artificial Evolution*, volume 1063 of *LNCS*. Springer Verlag, 1996.

[14] U. Chakraborty, K. Deb, and M. Chakraborty. Analysis of selection algorithms: A markov chain approach. *Evolutionary Computation*, 4(2):133–168, Summer 1996.

[15] J. Culberson and Jonathan Lichtner. On searching 'alpha'-ary hypercubes and related graphs. In R. K. Belew and M. D. Vose, editors, *Foundations of Genetic Algorithms 4*, pages 263–290. Morgan Kaufmann, 1996.

[16] R. Das and D. Whitley. The only challenging problems are deceptive: Global search by solving order-1 hyperplanes. In R. K. Belew and L. B. Booker, editors, *Proceedings of the 4^{th} International Conference on Genetic Algorithms*, pages 166–173. Morgan Kaufmann, 1991.

[17] T. E. Davis and J. C. Principe. A markov chain framework for the simple genetic algorithm. *Evolutionary Computation*, 1(3):269–292, 1993.

[18] Stefan Droste, Thomas Jansen, and Ingo Wegener. A rigorous complexity analysis of the (1+1) evolutionary algorithm for linear functions with boolean inputs. In *Proceedings of the Fifth IEEE International Conference on Evolutionary Computation.* IEEE Press, 1998.

[19] A.E. Eiben, E.H.L. Aarts, and K.M. Van Hee. Global convergence of genetic algorithms: a markov chain analysis. In Hans-Paul Schwefel and Reinhard Männer, editors, *Proceedings of the 1^{st} Parallel Problem Solving from Nature*, pages 4–12. Springer Verlag, 1991.

[20] S. Ethier and T. Kurtz. *Markov processes.* Wiley, New York, 1986.

[21] W. Feller. *An introduction to probability theory and its applications.* Wiley, New York, 1971.

[22] D. E. Goldberg. *Genetic algorithms in search, optimization and machine learning.* Addison Wesley, 1989.

[23] D. E. Goldberg. The race, the hurdle and the sweet spot: Lessons from genetic algorithms for the automation of design innovation and creativity. Technical Report 98007, Illinois University, 1998.

[24] D. E. Goldberg and K. Deb. A comparative analysis of selection schemes used in genetic algorithms. In G. J. E. Rawlins, editor, *Foundations of Genetic Algorithms*, pages 69–93. Morgan Kaufmann, 1991.

[25] J. Greffenstette. Efficient implementation of algorithms. In Th. Bäck, D.B. Fogel, and Z. Michalewicz, editors, *Handbook of evolutionary computation*, pp E2.1:1-6, IOP Publishing and Oxford University Press, Bristol and New York, 1997.

[26] I. Guikhman and A. Skorokhod. Introduction à la théorie des processus aléatoires. Mir, Moscou, 1986.

[27] J. H. Holland. *Adaptation in natural and artificial systems.* University of Michigan Press, Ann Arbor, 1975.

[28] J. Horn and C.R Reeves. The genetic crossover landscape for the *onemax* problem. In J.Alander, editor, *Proceedings of the 2^{nd} Nordic Workshop on Genetic Algorithms and their Applications*, pages 27–43. University of Vaasa Press,Vaasa, Finland, 1996.

[29] M. Iosifescu. *Finite Markov processes and their applications.* Wiley, Bucarest, 1980.

[30] T. Jones and S. Forrest. Fitness distance correlation as a measure of problem difficulty for genetic algorithms. In L. J. Eshelman, editor, *Proceedings of the 6^{th} International Conference on Genetic Algorithms*, pages 184–192. Morgan Kaufmann, 1995.

[31] J.G. Kemeny and J.L. Snell. *Finite Markov chains.* Van Nostrand, Princeton, 1960.

[32] H. Mühlenbein. How genetic algorithms really work: I. mutation and hill-climbing. In R. Manner and B. Manderick, editors, *Proceedings of the 2^{nd} Conference on Parallel Problems Solving from Nature*, pages 15–25. Morgan Kaufmann, 1992.

[33] A.E. Nix and M.D. Vose. Modeling genetic algorithms with markov chains. *Annals of Mathematics and Artificial Intelligence*, 5(1):79–88, 1992.

[34] J. Obalek. *Rekombinationsoperatoren für Evolutionsstrategieren.* Diploma thesis, Universität Dortmund, Fachbereich Informatik, 1994.

[35] N. J. Radcliffe. Equivalence class analysis of genetic algorithms. *Complex Systems*, 5:183–20, 1991.

[36] I. Rechenberg. *Evolutionstrategie: Optimierung Technisher Systeme nach Prinzipien des Biologischen Evolution.* Fromman-Holzboog Verlag, Stuttgart, 1973.

[37] D. Revuz. Markov chains. North-Holland, Amsterdam, 1975.

[38] G. Rudolph. Convergence of non-elitist strategies. In Z. Michalewicz, J. D. Schaffer, H.-P. Schwefel, D. B. Fogel, and H. Kitano, editors, *Proceedings of the First IEEE International Conference on Evolutionary Computation*, pages 63–66. IEEE Press, 1994.

[39] G. Rudolph. How mutation and selection solve long path problems in polynomial expected time. *Evolutionary Computation*, 4(2):195–205, Summer 1996.

[40] G. Rudolph. Asymptotical convergence rates of simple evolutionary algorithms under factorizing mutation distributions. In J.-K. Hao, E. Lutton, E. Ronald, M. Schoenauer, and D. Snyers, editors, *Artificial Evolution'97*, pages 223–233. Springer Verlag, 1997.

[41] G. Rudolph. *Convergence Properties of Evolutionary Algorithms.* Kovac, Hamburg, 1997.

[42] G. Rudolph and J. Ziegenhirt. Computation time of evolutionary operators. In Th. Bäck, D.B. Fogel, and Z. Michalewicz, editors, *Handbook of evolutionary computation*, pp E2.2:1-4, IOP Publishing and Oxford University Press, Bristol and New York, 1997.

[43] H.-P. Schwefel. *Numerical Optimization of Computer Models.* John Wiley & Sons, New-York, 1981. 1995 – 2^{nd} edition.

[44] P.D. Surry and N.J. Radcliffe. Formal algorithms + formal representations = search strategies. In H.-M. Voigt, W. Ebeling, I. Rechenberg, and H.-P. Schwefel, editors, *Proceedings of the* 4^{th} *Conference on Parallel Problems Solving from Nature*, number 1141 in LNCS, pages 366–375. Springer Verlag, 1996.

[45] J. Suzuji. A markov chain analysis on a genetic algorithm. In S. Forrest, editor, *Proceedings of the* 5^{th} *International Conference on Genetic Algorithms.* Morgan Kaufmann, 1993.

[46] D. Thierens and D.E. Goldberg. Mixing in genetic algorithms. In S. Forrest, editor, *Proceedings of the* 5^{th} *International Conference on Genetic Algorithms*, pages 38–55. Morgan Kaufmann, 1993.

[47] S.W. Wilson. Ga-easy does not imply steepest-ascent optimizable. In R. K. Belew and L. B. Booker, editors, *Proceedings of the* 4^{th} *International Conference on Genetic Algorithms.* mk, 1991.

Appendix: Some useful results of probability theory

Markov property

A Markov chain $Z(n)$ satisfies the so-called weak and strong Markov properties [37].
Weak Markov property
For any integer n_0 and for any function ϕ which depends on the whole path:

$$\mathbb{E}\left[\phi\left((Z(n))_{n\geq n_0}\right)\middle|\,(Z(n))_{n\leq n_0}\right] = \mathbb{E}\left[\phi\left((Z(n))_{n\geq n_0}\right)\middle|\,Z(n_0)\right].$$

Besides, if the chain is homogeneous (i.e. its transition matrix is time-independent), then:

$$\mathbb{E}\left[\phi\left((Z(n))_{n\geq n_0}\right)\middle|\,Z(n_0)\right] = \mathbb{E}_{Z(n_0)}\left[\phi\left((Z(n))_{n\geq 0}\right)\right].$$

Strong Markov property
For any stopping time τ and for any function ϕ which depends on the whole path:

$$\mathbb{E}\left[\phi\left((Z(n))_{n\geq \tau}\right)\middle|\,(Z(n))_{n\leq \tau}\right] = \mathbb{E}\left[\phi\left((Z(n))_{n\geq \tau}\right)\middle|\,Z(\tau)\right].$$

Besides, if the chain is homogeneous, then:

$$\mathbb{E}\left[\phi\left((Z(n))_{n\geq\tau}\right)\middle|\, Z(\tau)\right] = \mathbb{E}_{Z(\tau)}\left[\phi\left((Z(n))_{n\geq 0}\right)\right].$$

If τ is a stopping time of the form $\tau = \inf\{n, Z(n) = k\}$, then:

$$\mathbb{E}_{Z(\tau)}\left[\phi\left((Z(n))_{n\geq 0}\right)\right] = \mathbb{E}_{k}\left[\phi\left((Z(n))_{n\geq 0}\right)\right].$$

Remark **5.1**
If ϕ depends only on one point, say $Z(n_0+k)$, and is of the form $\mathbb{I}_{z\geq a}$, then the above expectations read as probabilities:

$$\mathbb{E}\left[\phi((Z(n))_{n\geq n_0})\middle|\,(Z(n))_{n\leq n_0}\right] = \mathbb{P}\left(Z(n_0+k)\geq a\middle|\, Z(n_0)\right).$$

Renewal theory

Let $(N_j)_{j\in\mathbb{N}}$ be a sequence of independent and identically distributed random variables with distribution $\mathbb{P}(N_1 = k) = p_k > 0$, $k\in\mathbb{N}$.

If we denote by S_n the partial sums $S_n = \sum_{j=1}^{n} N_j$, then for any k, there exists a unique n_k such that $S_{n_k-1}\leq k < S_{n_k}$. The sequence n_k then satisfies [26]:

$$\frac{\mathbb{E}[n_k]}{k} \overset{k\to\infty}{\longrightarrow} \frac{1}{\mathbb{E}[N_1]}, \tag{43}$$

$$\mathbb{P}\left(\exists n, S_n = k\right) \overset{k\to\infty}{\longrightarrow} \frac{1-p_0}{\mathbb{E}[N_1]}. \tag{44}$$

A lemma about exponential random variables

The following characterization of exponential random variables is well-known in probability theory and its proof can be found for instance in [29, Section 1.4.3]. We express it in a suitable way for our purpose.

Lemma **5.2**
Let T be a $\mathbb{R}^+$-valued random variable. T obeys an exponential distribution if and only if for every $a, b\geq 0$:

$$\mathbb{P}(T\geq a+b|T\geq a) = \mathbb{P}(T\geq b).$$

Stirling's formula

If n is a large integer, then [1]:

$$n! \simeq \left(\frac{n}{e}\right)^n \sqrt{2\pi n}.$$

Laplace technique

If $f\in\mathcal{C}^3(a,b)$ reaches its minimal value at c which lies inside the interval (a,b), if $f''(c) < 0$, and if $g\in\mathcal{C}^0(a,b)\cap L^1(a,b)$, then:

$$\left|\sqrt{l}\int_a^b e^{-lf(s)}g(s)ds - \frac{g(c)\sqrt{2\pi}}{\sqrt{|f''(0)|}}\right| \overset{l\to 0}{\longrightarrow} 0.$$