



# Precise Runtime Analysis for Plateau Functions

Denis Antipov, Benjamin Doerr

## ► To cite this version:

Denis Antipov, Benjamin Doerr. Precise Runtime Analysis for Plateau Functions. ACM Transactions on Evolutionary Learning and Optimization, 2021, 1 (4), pp.1-28. 10.1145/3469800 . hal-03418304

**HAL Id: hal-03418304**

**<https://hal.science/hal-03418304>**

Submitted on 7 Nov 2021

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

# Precise Runtime Analysis for Plateau Functions\*

Denis Antipov  
ITMO University  
St. Petersburg  
Russia

and

Laboratoire d'Informatique (LIX)  
CNRS  
École Polytechnique  
Institute Polytechnique de Paris  
Palaiseau  
France

Benjamin Doerr  
Laboratoire d'Informatique (LIX)  
CNRS  
École Polytechnique  
Institute Polytechnique de Paris  
Palaiseau  
France

November 2, 2021

---

\*This work is a significantly extended version of the PPSN 2018 paper [AD18]. It completes the original work by including the mathematical proofs, which were omitted

## Abstract

To gain a better theoretical understanding of how evolutionary algorithms (EAs) cope with plateaus of constant fitness, we propose the  $n$ -dimensional  $\text{PLATEAU}_k$  function as natural benchmark and analyze how different variants of the  $(1 + 1)$  EA optimize it. The  $\text{PLATEAU}_k$  function has a plateau of second-best fitness in a ball of radius  $k$  around the optimum. As evolutionary algorithm, we regard the  $(1 + 1)$  EA using an arbitrary unbiased mutation operator. Denoting by  $\alpha$  the random number of bits flipped in an application of this operator and assuming that  $\Pr[\alpha = 1]$  has at least some small sub-constant value, we show the surprising result that for all constant  $k \geq 2$ , the runtime  $T$  follows a distribution close to the geometric one with success probability equal to the probability to flip between 1 and  $k$  bits divided by the size of the plateau. Consequently, the expected runtime is the inverse of this number, and thus only depends on the probability to flip between 1 and  $k$  bits, but not on other characteristics of the mutation operator. Our result also implies that the optimal mutation rate for standard bit mutation here is approximately  $k/(en)$ . Our main analysis tool is a combined analysis of the Markov chains on the search point space and on the Hamming level space, an approach that promises to be useful also for other plateau problems.

## 1 Introduction

This work aims at making progress on several related subjects—we aim at understanding how evolutionary algorithms optimize non-unimodal<sup>1</sup> fitness functions, what mutation operators to use in such settings, how to analyze the behavior of evolutionary algorithms on large plateaus of constant fitness, and in particular, how to obtain runtime bounds that are precise including the leading constant.

The recent work [DLMN17] observed that a large proportion of the theoretical work in the past concentrates on analyzing how evolutionary algorithms optimize unimodal fitness functions and that this can lead to misleading recommendations on how to design evolutionary algorithms. Based on a precise analysis of how the  $(1 + 1)$  EA optimizes jump functions, it was

---

in the conference version for reasons of space, and it extends the conference version by proving the same result for mutation operators with a sub-constant probability to flip exactly one bit, by a tail bound for the runtime, and by a wider selection of applications in Section 5.

<sup>1</sup>As common in optimization, we reserve the notion *unimodal* for objective functions such that each non-optimal search point has a strictly better neighbor.

observed that the classic recommendation to use standard bit mutation with mutation rate  $\frac{1}{n}$  is far from optimal for this function class. For jump size  $k$ , a speed-up of order  $k^{\Theta(k)}$  can be obtained by using a mutation rate of  $\frac{k}{n}$ .

Jump functions are difficult to optimize because the optimum is surrounded by a large set of search points of very low fitness (all search points in Hamming distance 1 to  $k - 1$  from the optimum). However, local optima are not the only feature which makes functions difficult to optimize. Another challenge for most evolutionary algorithms are large plateaus of constant fitness. On such plateaus, the evolutionary algorithm learns little from evaluating search points and consequently performs an unguided random walk. To understand this phenomenon in more detail, we propose a class of fitness functions very similar to jump functions. A *plateau function* with plateau parameter  $k$  is identical to a jump function with jump size  $k$  except that the  $k - 1$  Hamming levels around the optimum do not have a small fitness, but have the same second-best fitness as the  $k$ -th Hamming level. Consequently, these functions do not have true local optima (in which an evolutionary algorithm could get stuck for longer time), but only a plateau of constant fitness. Our hope is that this generic fitness function with a plateau of scalable size may aid the understanding of plateaus in evolutionary computation in a similar manner as the jump functions have led to many useful results about the optimization of functions with true local optima, e.g., [DJW02, JW02, DDK15, BDK16, DFK<sup>+</sup>16, FKK<sup>+</sup>16, COY17, COY18, DLMN17, DFK<sup>+</sup>18, WVHM18, HS18, Doe19b, Doe19a].

When trying to analyze how evolutionary algorithms optimize plateau functions, we observe that the active area of theoretical analyses of evolutionary algorithms has produced many strong tools suitable to analyze how evolutionary algorithms make true progress (e.g., various form of the fitness level method [Weg01, Sud13, DL16, CDEL18] or drift analysis [HY01, DJW12, LW14, DK19]), but much less is known on how to analyze plateaus. This is not to mean that plateaus have not been analyzed previously, see, e.g., [GKS99, JW01, DHN07, BFH<sup>+</sup>09, FHN09, NSW09, FHN10], but these results appear to be more ad hoc and less suitable to derive generic methods for the analysis of plateaus. In particular, with the exception of [GKS99], we are not aware of any results that determine the runtime of an evolutionary algorithm on a fitness function with non-trivial plateaus precisely including the leading constant (whereas a decent number of very precise results have recently appeared for unimodal fitness functions, e.g., [BDN10, DFW11, Wit13, LOW17, HPR<sup>+</sup>18, DDL19, HW19]).

Such precise results are necessary for our further goal of understanding the influence of the mutation operator on the efficiency of the optimization process. Mutation is one of the most basic building blocks in evolutionary

computation and has, consequently, received significant attention also in the runtime analysis literature. We refer to the discussion in [DLMN17] for a more extensive treatment of this topic and only note here that even small changes of the mutation operator or its parameters can lead to a drastic change of the efficiency of the algorithm [DJK08, DJS<sup>+</sup>13].

**Our results:** Our main result is a very general analysis of how the simplest mutation-based evolutionary algorithm, the  $(1 + 1)$  EA, optimizes the  $n$ -dimensional plateau function with plateau parameter  $k \in \mathbb{N}$ , which is considered as a constant and does not depend on  $n$  when  $n$  tends to the positive infinity. We allow the algorithm to use any unbiased mutation operator (including, e.g., one-bit flips, standard bit mutation with an arbitrary mutation rate, or the fast mutation operator of [DLMN17]) as long as the operator flips exactly one bit with probability  $\omega(n^{-\frac{1}{2k-2}})$ . This assumption is natural, but also ensures that the algorithm can reach all points on the plateau. Denoting the number of bits flipped in an application of this operator by the random variable  $\alpha$ , we prove that the expected optimization time (number of fitness evaluations until the optimum is visited) is

$$\frac{n^k}{k! \Pr[1 \leq \alpha \leq k]} (1 + o(1)).$$

This result, tight apart from lower order terms only, is remarkable in several respects. It shows that the performance depends very little on the particular mutation operator, only the probability to flip between 1 and  $k$  bits has an influence. The absolute runtime is also surprising — it is the size of the plateau times the waiting time until we flip between 1 and  $k$  bits.

A similar-looking result was obtained in [GKS99], namely that the expected runtime of the  $(1 + 1)$  EA with 1-bit mutation and with standard bit mutation with rate  $\frac{1}{n}$  on the needle function is (apart from lower order terms) the size of the plateau times the probability to flip a positive number of bits (which is 1 for 1-bit mutation and  $(1 - o(1))(1 - \frac{1}{e})$  for standard bit mutation with rate  $\frac{1}{n}$ ). Our result is different from that one in that we consider constrained plateaus of arbitrary (constant) radius  $k \geq 2$ , and more general in that we consider a wide class of unbiased mutation operators. Despite the difference in the plateaus, the expected runtime is surprisingly similar, which is the size of the plateau times the expected number of iterations until we flip between 1 and  $k$  bits (where for the needle function we can take  $k = n$ ).

We note that there is a substantial difference between the case  $k = n$  and  $k$  constant. Since the needle function consists of a plateau containing the whole search space apart from the optimum, the optimization time in this case is just the hitting time of a particular search point when doing an

undirected random walk (via repeated mutation) on the hypercube  $\{0, 1\}^n$ . For  $\text{PLATEAU}_k$  with constant  $k$ , the plateau has a large boundary. More precisely, almost all<sup>2</sup> search points of the plateau lie on its outer boundary and furthermore, all these search points have almost all their neighbors outside the plateau. Hence a large number of iterations (namely almost all) are lost in the sense that the mutation operator generates a search point outside the plateau (and different from the optimum), which is not accepted. Interestingly, as our result shows, the optimization of such restricted plateaus is not necessarily significantly more difficult (relative to the plateau size) than the optimization of the unrestricted needle plateau.

Our precise runtime analysis allows to deduce a number of particular results. For example, when using standard bit mutation, the optimal<sup>3</sup> mutation rate is  $\frac{\sqrt[k]{k!}}{n}$ , that is, approximately  $\frac{k}{en}$ . This is by a constant factor less than the optimal rate of  $\frac{k}{n}$  for the jump function with jump size  $k$ , but again a factor of  $\Theta(k)$  larger than the classic recommendation of  $\frac{1}{n}$ , which is optimal for many unimodal fitness functions. Hence our result confirms that optimal mutation rates can be significantly higher for non-unimodal fitness functions. While the optimal mutation rates for jump and plateau functions are similar, the effect of using the optimal rate is very different. For jump functions, an  $k^{\Theta(k)}$  factor speed-up (compared to the standard recommendation of  $\frac{1}{n}$ ) was observed, here the influence of the mutation operator is much smaller, namely the factor  $\Pr[1 \leq \alpha \leq k]$ , which is trivially at most 1, but which was assumed to be at least some positive constant. Interestingly, our results imply that the fast mutation operator described in [DLMN17] is not more effective than other unbiased mutation operators, even though it was proven to be significantly more effective for jump functions [DLMN17] and it has shown good results in some practical problems [MB17].

So one structural finding, which we believe to be true for larger classes of problems and which fits to the result [GKS99] for needle functions, is that the mutation rate, and more generally, the particular mutation operator which is used, is less important while the evolutionary algorithm is traversing a plateau of constant fitness.

The main technical novelty in this work is that we model the optimization process via two different Markov chains describing the random walk on the plateau, namely the chain defined on the  $\Theta(n^k)$  elements of the plateau (plus

---

<sup>2</sup>in the usual asymptotic sense, that is, meaning all but a lower order fraction

<sup>3</sup>We call a mutation rate optimal when it delivers an expected runtime that differs from the truly optimal one at most by lower order terms, that is, e.g. a factor of  $(1 \pm o(1))$ . This suggests that there might be a range of optimal rates, however without proof we note that changing the mentioned optimal mutation rate by a factor of  $(1 \pm \Omega(1))$  would also increase the runtime by a  $(1 + \Omega(1))$  factor.

the optimum) and the chain obtained from aggregating these into the total mass on the Hamming levels. Due to the symmetry of the process, one could believe that it suffices to regard only the level chain. The chain defined on the elements, however, has some nice features which the level chain is missing, among others, a symmetric transition matrix (because for any two search points  $x$  and  $y$  on the plateau, the probability of going from  $x$  to  $y$  is the same as the probability of going from  $y$  to  $x$ ). This symmetry allows us to analyse the speed of convergence to some distribution over the points of the plateau by using some ideas similar to the ones used in [Vit00] for the analysis of the rapidly mixing Markov chains. For this reason, we find it fruitful to switch between the two chains. Exploiting the interplay between the two chains and using classic methods from linear algebra, we find the exact expression for the expected runtime.

The most valuable insight given by this approach is that the mixing of the probability mass over the plateau is very fast. More precisely, we show that independently of the first position on the plateau, in slightly more than  $\Theta(\sqrt{n} \log(n))$  iterations we are almost equally likely to be at any point of the plateau. A similar mixing argument was used to prove the upper bound on the runtime of the  $(1 + 1)$  EA on the LEADINGONES with strong prior noise in [Sud20]. There, however, only an exponential mixing time was shown, although the author conjectures that it should be polynomial. Our analysis based on the interplay of two Markov chains is problem-specific (e.g., we base our arguments on the symmetry of the plateau), but we are optimistic that the observed behavior of a small mixing time can be also seen on other plateaus which are not too easy to leave.

The rest of the paper has the following structure. In Section 2 we describe the  $(1 + 1)$  EA, the operators it uses and the problem on which we analyse the algorithm. In Section 3 we list the mathematical means that are used in our analysis. We also introduce the central tool of our analysis — the two Markov chains, show their properties and the connection between the two chains. In Section 4 we prove the main result of this work, which is, the precise runtime of the  $(1 + 1)$  EA on the  $\text{PLATEAU}_k$  function for constant  $k$ . The corollaries from the main result, which are, the precise runtime of different variants of the  $(1 + 1)$  EA, are shown in Section 5. Finally, we summarize the results in Section 6.

## 2 Problem Statement

We consider the maximization of a function defined on the space of bit-strings of length  $n$  which resembles the ONEMAX function, but has a plateau of

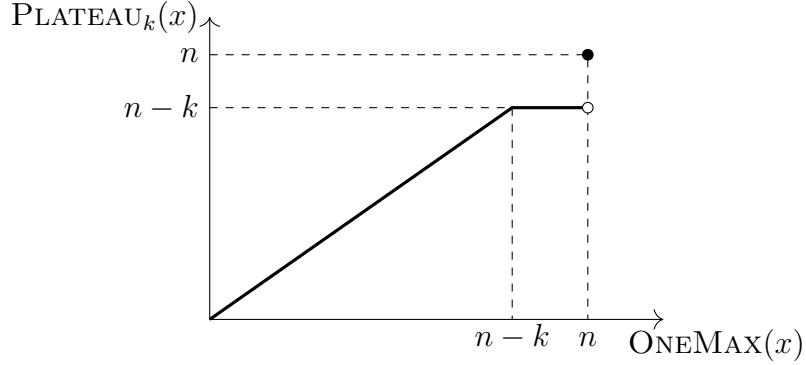


Figure 1: Plot of the PLATEAU function. As a function of unitation, the function value of a search point  $x$  depends only on the number of one-bits in  $x$ .

second-highest fitness of radius  $k$  around the optimum. We call this function  $\text{PLATEAU}_k$  and define it as follows.

$$\text{PLATEAU}_k(x) := \begin{cases} n - k, & \text{if } n - k < \text{ONEMAX}(x) < n, \\ \text{ONEMAX}(x), & \text{otherwise,} \end{cases}$$

where  $\text{ONEMAX}(x) := \|x\|_1$  is the number of one-bits in  $x \in \{0, 1\}^n$ .

Notice that the plateau of the function  $\text{PLATEAU}_k(x)$  consists of all bit-strings that have at least  $n - k$  one-bits, except the optimal bit-string  $x^* = (1, \dots, 1)$ . See Fig. 1 for an illustration of  $\text{PLATEAU}_k$ .

To compare the results of our analysis to the best runtime which could be obtained by an algorithm using only unbiased operators, we note that the *unary unbiased black-box complexity* (see [LW12] for the definition) of  $\text{PLATEAU}_k$  is  $\Theta(n \log n)$  for all constants  $k$ . While this implies that there is a unary unbiased black-box algorithm finding the optimum of  $\text{PLATEAU}_k$  in  $O(n \log n)$  time, such results generally do not indicate that a problem is easy for reasonable evolutionary algorithms. For example, in [DDK14] it was shown that the NP-complete partition problem also has a unary unbiased black-box complexity of  $O(n \log n)$ .

**Lemma 1.** *For all constants  $k$ , the unary unbiased black-box complexity of the  $\text{PLATEAU}_k$  function is  $\Theta(n \log n)$ .*

*Proof.* The lower bound follows from the  $\Omega(n \log n)$  lower bound for the unary unbiased black-box complexity of  $\text{ONEMAX}$  shown in [LW12]. Since we can



write  $\text{PLATEAU}_k = f \circ \text{ONEMAX}$  for a suitable function  $f$  (such that  $f(x) = x$ , if  $x \notin [n - k..n]$  and  $f(x) = n - k$  otherwise), any algorithm solving  $\text{PLATEAU}_k$  can be transferred into an algorithm which treats all points with fitness in  $[n - k..n - 1]$  as points with fitness  $(n - k)$  and therefore solving  $\text{ONEMAX}$  in the same time.

The upper bound follows along the same lines as the  $O(n \log n)$  upper bound for the unary unbiased black-box complexity of  $\text{JUMP}_k$ , see [DDK15] and note that the algorithm given there contains a sub-routine which, in expected constant time, for a given constant radius  $r$  determines the Hamming distance  $H(x, x^*)$  of a point  $x$  from the optimum  $x^*$  without evaluating search points  $y$  with  $H(y, x^*) \leq r$ . Note that the Hamming distance from the optimum determines the  $\text{ONEMAX}$  value of  $x$ . Hence with this routine one can optimize both jump and plateau functions by simulating an  $O(n \log n)$  black-box algorithm for  $\text{ONEMAX}$ .  $\square$

To understand how evolutionary algorithms optimize plateau functions, we consider the most simple evolutionary algorithm, the  $(1 + 1)$  EA shown in Algorithm 1. However, we allow the use of an arbitrary unbiased mutation operator. A mutation operator  $\text{MUTATE}$  for bit-string representations is called *unbiased* if it is symmetric in the bit-positions  $[1..n]$  and in the bit-values 0 and 1. This is equivalent to saying that for all  $x \in \{0, 1\}^n$  and all automorphisms  $\sigma$  of the hypercube  $\{0, 1\}^n$  (respecting Hamming neighbors) we have  $\sigma^{-1}(\text{MUTATE}(\sigma(x))) = \text{MUTATE}(x)$ , which is an equality of distributions. The notation of unbiasedness was introduced (also for higher-arity operators) in the seminal paper [LW12].

---

**Algorithm 1:** The fast  $(1 + 1)$  EA with a generic mutation operator maximizing  $f : \{0, 1\}^n \rightarrow \mathbb{R}$

---

```

1  $x \leftarrow$  random bit string of length  $n$ ;
2 while not terminated do
3    $y \leftarrow \text{MUTATE}(x)$ ;
4   if  $f(y) \geq f(x)$  then
5      $x \leftarrow y$ ;
6   end
7 end
```

---

For our purposes, it suffices to know that the set of unbiased mutation operators consists of all operators which can be described as follows. First, we choose a number  $\alpha \in [0..n]$  according to some probability distribution and then we flip exactly  $\alpha$  bits chosen uniformly at random. Examples for

unbiased operators are the operator of Random Local Search, which flips a single random bit, or standard bit mutation, which flips each bit independently with probability  $\frac{1}{n}$ . Note that in the first case  $\alpha$  is always equal to one, whereas in the latter  $\alpha$  follows a binomial distribution with parameters  $n$  and  $\frac{1}{n}$ . This characterization can be derived from [DKLW13, Proposition 19]. It was explicitly stated in [DDY20].

**Additional assumptions:** The class of unbiased mutation operators contains a few operators which are unable to solve even very simple problems. For example, operators that always flips exactly two bits never finds the optimum of any function with unique optimum if the initial individual has an odd Hamming distance from the optimum. To avoid such artificial difficulties, we only consider unbiased operators that have at least  $\omega(n^{-\frac{1}{2k-2}})$  probability to flip exactly one bit.

As usual in runtime analysis, we are interested in the optimization behavior for large problem size  $n$ . Formally, this means that for each fixed  $k$  we view the runtime  $T_k = T_k(n)$  as a function of  $n$  and aim at understanding its asymptotic behavior for  $n$  tending to infinity. We aim at sharp results (including finding the leading constant), that is, we try to find a simple function  $\tau_k : \mathbb{N} \rightarrow \mathbb{R}$  such that  $T_k(n) = (1 + o(1))\tau_k(n)$ , which is equivalent to saying that  $\lim_{n \rightarrow \infty} T_k(n)/\tau_k(n) = 1$ . In this limit sense, however, we treat  $k$  as a constant, that is,  $k$  is a given positive integer and not also a function of  $n$ .

Finally, since the case  $k = 1$  is well-understood (PLATEAU<sub>1</sub> is the well-known ONEMAX function), we always assume  $k \geq 2$ .

## 3 Preliminaries and Notation

### 3.1 Tools from Linear Algebra

In this section we briefly review the terms, tools and facts from the linear algebra that we use in this work.

We use  $\mathbb{N}$  to denote the set of all positive integer numbers and we use  $\mathbb{N}_0$  to denote  $\mathbb{N} \cup \{0\}$ . We denote the vector of length  $n$  that consists only of ones by  $1^n$  and the vector of length  $n$  that consists only of zeros by  $0^n$ .

Given the square matrix  $A$ , the vector  $x$  is called the *left eigenvector* of the matrix  $A$  if  $xA = \lambda x$  for some  $\lambda \in \mathbb{C}$ . In this situation,  $\lambda$  is called *eigenvalue* of the matrix  $A$ . The vector  $x$  is called *right eigenvector* if  $Ax = \lambda x$  for some  $\lambda \in \mathbb{C}$ . Since in this work we regard only left eigenvectors, we call them just *eigenvectors*.

The *spectrum* of a matrix is the set of all its eigenvalues. If a matrix has size  $n \times n$ , then the number of its eigenvalues is not greater than  $n$ . For each eigenvalue there exists a corresponding *eigenspace*, that is, the linear span of all the eigenvectors that correspond to the eigenvalue.

The only point shared by any two eigenspaces that correspond to two different eigenvalues is  $0^n$ .

The *characteristic polynomial*  $\chi(\lambda)$  of matrix  $A$  is the function of  $\lambda$  that is defined as the determinant of the matrix  $A - \lambda I$ , where  $I$  is the identity matrix. The set of roots of the characteristic polynomial equals the spectrum of the matrix  $A$ .

The inner product of the vectors  $x = (x_0, \dots, x_{n-1})$  and  $y = (y_0, \dots, y_{n-1})$  is a scalar value defined by  $\langle x, y \rangle = \sum_{i=0}^{n-1} x_i y_i$ . The two vectors are *orthogonal* if their inner product is zero.

For every *diagonalizable* matrix  $A$  of size  $n \times n$  there exists a set  $\{e^i\}_{i=0}^{n-1}$  of eigenvectors that form a *basis* of  $\mathbb{R}^n$ . A basis is called *orthogonal* when all pairs of the basis vectors are orthogonal. A matrix  $A = (a_{ij}^j)$  is *symmetric* if for every  $i$  and  $j$  we have  $a_{ij}^j = a_{ji}^i$ .

We use the following two properties of symmetric matrices.

**Lemma 2.** *All eigenvalues of a symmetric matrix are real.*

**Lemma 3.** *Two eigenvectors of a symmetric matrix that correspond to different eigenvalues are orthogonal. Also every symmetric matrix of size  $n \times n$  is diagonalizable, which means that there exists an orthogonal basis of  $\mathbb{R}^n$  which consist of eigenvectors of this matrix.*

In this work we also encounter *irreducible* matrices. Among the several definitions, the following is the easiest to check for the non-negative matrices considered in this work. For each non-negative matrix  $A$  of size  $n \times n$  we can build a directed graph  $G_A$  by taking an empty graph on  $n$  vertices and adding for each non-negative component  $a_{ij}^j$  of  $A$  an edge from vertex  $i$  to vertex  $j$ . Then a matrix  $A$  is *irreducible* if and only if graph  $G_A$  is strongly connected.

For example, the transition matrix of an irreducible Markov chain (a chain such that each state is reachable from each other state) is irreducible.

A crucial role in this work is played by the Perron-Frobenius theorem [Mey00]. This theorem gives a series of properties of the irreducible matrices, among them we use the following four.

**Theorem 4** (Perron-Frobenius). *Any irreducible non-negative matrix  $A$  has the following properties.*

- The largest eigenvalue  $\lambda_0$  of  $A$  lies between the minimal and the maximal row sum of  $A$ .
- For every eigenvalue  $\lambda$  of  $A$  different from the largest eigenvalue  $\lambda_0$  we have  $|\lambda| < \lambda_0$ .
- The largest eigenvalue of  $A$  has a one-dimensional eigenspace.
- There exists an eigenvector which corresponds to the largest eigenvalue  $\lambda_0$  all components of which are strictly positive.

When talking about vector norms, we use the following notation. For any  $p \in (0, +\infty)$  and any vector  $x \in \mathbb{R}^n$ , we let

$$\|x\|_p = \left( \sum_{j=0}^{n-1} |x_j|^p \right)^{1/p}.$$

In this work we use only the Manhattan norm ( $p = 1$ ) and the Euclidean norm ( $p = 2$ ). We use the following properties of these norms.

**Lemma 5.** For all  $x \in \mathbb{R}^n$  we have

$$\|x\|_2 \leq \|x\|_1 \leq \sqrt{n} \|x\|_2.$$

The following lemma is often called *triangle inequality*

**Lemma 6.** For any norm  $\|\cdot\|$  and for every  $x, y$  and  $z = x + y$  we have

$$\|x\| - \|y\| \leq \|z\| \leq \|x\| + \|y\|$$

We use the following properties of the Euclidean norm.

**Lemma 7.** If vectors  $x^1, \dots, x^n$  are orthogonal, then for any values  $a_1, \dots, a_n \in \mathbb{R}$  we have

$$\left\| \sum_{i=1}^n a_i x^i \right\|_2 \leq \max_{i \in [1..n]} |a_i| \left\| \sum_{i=1}^n x^i \right\|_2$$

**Lemma 8.** If vectors  $x^1, \dots, x^n$  are orthogonal, then for any subset  $S \subset [1..n]$  we have

$$\left\| \sum_{i \in S} x^i \right\|_2 \leq \left\| \sum_{i=1}^n x^i \right\|_2$$

We also make a use of the orthogonal projection of vectors, which is defined as follows. Suppose we have vector  $x \in \mathbb{R}^n$  and it is decomposed into the sum of  $m$  orthogonal vectors  $\{x^i\}_{i=0}^{m-1}$  where  $m \leq n$ . Then  $x^i$  is the *orthogonal projection* of  $x$  to the linear span of  $x^i$ . To calculate precisely the norm of the projection, we use the following lemma.

**Lemma 9.** *If  $x^i$  is the orthogonal projection of  $x$ , then for any norm  $\|\cdot\|$  we have*

$$\|x^i\| = \frac{\langle x, \frac{x^i}{\|x^i\|} \rangle}{\langle \frac{x^i}{\|x^i\|}, \frac{x^i}{\|x^i\|} \rangle}.$$

We also encounter the *self-adjoint operators*. An operator  $A : \mathbb{R}^n \rightarrow \mathbb{R}^n$  is called *self-adjoint* if for all  $x \in \mathbb{R}^n$  and  $y \in \mathbb{R}^n$  we have  $\langle Ax, y \rangle = \langle x, Ay \rangle$ , where  $\langle \cdot, \cdot \rangle$  stands for the standard inner product. The operator in this space is self-adjoint if and only if its matrix is symmetric. The most important properties of self-adjoint operators are stated in the Hilbert-Schmidt theorem [RR04]. We use only one of them.

**Lemma 10.** *For any self-adjoint operator  $A : \mathbb{R}^n \rightarrow \mathbb{R}^n$  there exists an orthonormal basis of  $\mathbb{R}^n$  that consists of the eigenvectors of  $A$ .*

### 3.2 Absorbing Markov Chains<sup>4</sup>

Markov chains are a widely used tool for the runtime analysis of evolutionary algorithms (see, e.g., [Müh93, Suz95, Rud96]). In this work we only regard *absorbing* Markov chains. A Markov chain is called absorbing if there is a subset  $S'$  of the set of its states  $S$  such that

- (1) for every state  $s_1 \in S$  there exists a state  $s_2 \in S'$  such that there exists a path of transitions with positive probabilities from  $s_1$  to  $s_2$  (we call  $s_2$  an *absorbing* state) and
- (2) for every absorbing state  $s \in S'$  the probability to leave this state is zero.

The non-absorbing states (the states in  $S \setminus S'$ ) are called the *transient states*.

---

<sup>4</sup>In this subsection we use a standard notation for the absorbing Markov chains such as  $N$  for the fundamental matrix,  $P$  for the transition matrix and  $Q$  for the transient-to-transient transition matrix. In the rest of the paper for the reader's convenience we redefine these common and easy-to-remember symbols to denote the objects we work with most frequently.

Absorbing chains appear naturally in runtime analysis. When taking as states of the Markov chain the possible states of the algorithm, we can assume the optima to be absorbing. The runtime of the algorithm is the number of transitions in the chain until it reaches an absorbing state. We only regard absorbing Markov chains with exactly one absorbing state.

The standard way to compute the expected number of steps until an absorbing state is reached uses the *fundamental matrix*, which is built as follows. Let  $P$  be the transition matrix of an absorbing Markov chain, that is, the matrix where each element  $p_i^j$  is equal to the transition probability from state  $i$  to state  $j$ . Let  $Q$  be the square submatrix of  $P$  consisting only of the rows and columns which correspond to transient states of the chain. We call  $Q$  the *transient matrix* for brevity. Then the fundamental matrix  $N$  of this chain is defined as

$$N = \sum_{t=0}^{+\infty} Q^t = (I - Q)^{-1},$$

where  $I$  is the identity matrix of the same order as  $Q$ . Let  $\pi$  be a stochastic vector which represents the initial distribution over the transient states. Then the expected time until we reach an absorbing state is

$$E[T] = \pi N \mathbf{1} = \|\pi N\|_1,$$

where  $\mathbf{1}$  is a column vector of all ones.

However, working with the fundamental matrix is not convenient, since it might be hard to compute its elements precisely. Instead, in this paper we study the properties of the transient matrix  $Q$  and compute the expected time until the absorption as

$$E[T] = \left\| \sum_{t=0}^{+\infty} \pi Q^t \right\|_1 = \sum_{t=0}^{+\infty} \|\pi Q^t\|_1, \quad (1)$$

where the last equation is satisfied since all components of vectors  $\pi Q^t$  are non-negative. Another way to derive this equation for the expected runtime is to use the formula for the expectation of a non-negative integer-valued random variable, which is,

$$E[T] = \sum_{t=0}^{+\infty} \Pr[T \geq t].$$

Note that  $\Pr[T \geq t]$  is the probability that we are in a transient state in the start of iteration  $t$ , which is  $\|\pi Q^t\|_1$ . We show this approach to be much more fruitful, since after finding some properties of the spectrum of  $Q$  it allows us to use the decomposition of  $\pi$  into the sum of eigenvectors of  $Q$  to obtain precise estimates on the runtime.

### 3.3 Two Markov Chains

For the optimization process of our  $(1 + 1)$  EA we first observe that, since the unbiased operator with constant probability flips exactly one bit, the expected time to reach the plateau is  $O(n \log n)$ . Since the time for leaving the plateau (as shown in this paper) is  $\Omega(n^k)$ , we only consider the runtime of the algorithm after it has reached the plateau.

For this runtime analysis on the plateau we consider the plateau in two different ways. The first way is to regard a Markov chain that contains  $N + 1$  states, where  $N = \sum_{i=0}^{k-1} \binom{n}{k-i}$ . Each state represents one element of the plateau plus there is one absorbing state for the optimum. Note that  $N = \frac{n^k}{k!} + o(n^k)$ , since  $\binom{n}{j} = \frac{n^j}{j!}(1 + o(1))$  for all  $j \in [1..k]$ . The transition probability from transient state  $x$  to any state  $y$  is  $q_x^y = \Pr[\alpha = d] \binom{n}{d}^{-1}$ , where  $d$  is the Hamming distance between  $x$  and  $y$ . This implies that the transition probability from  $x$  to  $y$  is equal to the transition probability from  $y$  to  $x$  for any pair of the transient states. Therefore, the transient matrix is symmetric, which gives us the opportunity to use Lemma 2 and Lemma 3. We call this Markov chain the *individual chain*, denote its transient matrix by  $Q$  and call the space of real vectors of dimension  $N$  the *individual space*<sup>5</sup>, since the current state of the chain defines the current individual of the algorithm.

To define the second Markov chain, we first define the  $i$ -th level as the set of all search points that have exactly  $n - k + i$  one-bits. Then the plateau is the union of levels 0 to  $k - 1$  and the optimum is the only element of level  $k$ . Notice that the  $i$ -th level contains exactly  $\binom{n}{k-i}$  elements (search points). For every  $i, j \in [0..k]$  we have that for any element of the  $i$ -th level the probability to mutate to the  $j$ -th level is the same due to the unbiasedness of the operator. Therefore we can regard a Markov chain of  $k + 1$  states, where the  $i$ -th state ( $i \in [0..k]$ ) represents the elements of the  $i$ -th level. State  $k$  is an absorbing state. The transition probability from level  $i$  to level  $j$  is

$$p_i^j = \begin{cases} 0, & \text{if } i = k, j \neq k, \\ \sum_{m=0}^{k-j} \binom{k-i}{j-i+m} \binom{n-k+i}{m} \binom{n}{j-i+2m}^{-1} \Pr[\alpha = j - i + 2m], & \text{if } j > i, \\ \sum_{m=0}^{k-i} \binom{k-i}{m} \binom{n-k+i}{i-j+m} \binom{n}{i-j+2m}^{-1} \Pr[\alpha = i - j + 2m], & \text{if } j < i \text{ and } i \neq k, \\ 1 - \sum_{m=0, m \neq i}^k p_i^m, & \text{if } j = i, \end{cases} \quad (2)$$

---

<sup>5</sup>Note that the dimension of the individual space is equal to the number of transient states of the individual chain, not to the total number of states. Hence, matrix  $Q$  defines a linear operator on the individual space.

where we assume that  $n > 2k$  not to complicate the upper limit of sums. This assumption is justified by that we only consider constant  $k$  and we estimate the runtime with  $n$  tending to infinity. We notice the following useful property of these probabilities.

**Lemma 11.** *For all  $i, j \in [0..k-1]$  we have*

$$\binom{n}{k-i} p_i^j = \binom{n}{k-j} p_j^i.$$

*Proof.* Let  $L_s$  denote level  $s$  for all  $s \in [0..k-1]$ . Let also  $p_{x \rightarrow L_s}$  denote the probability to get from individual  $x$  to any individual in level  $s$ . Since for all individuals  $x$  in level  $i$  the probability  $p_{x \rightarrow L_j}$  is the same and equal to  $p_i^j$  and since there are  $\binom{n}{k-i}$  individuals in level  $i$ , we have

$$\binom{n}{k-i} p_i^j = \sum_{x \in L_i} p_{x \rightarrow L_j} = \sum_{x \in L_i} \sum_{y \in L_j} q_x^y = \sum_{x \in L_i} \sum_{y \in L_j} q_y^x = \sum_{y \in L_j} p_{y \rightarrow L_i} = \binom{n}{k-j} p_j^i.$$

□

We observe that the probability to gain  $\ell$  levels is  $O(n^{-\ell})$ .

**Lemma 12.** *For all  $i \in [0..k-1]$  and  $j \in [i+1..k]$ , we have  $p_i^j = O(n^{-(j-i)})$ .*

*Proof.* By Lemma 11 and since  $p_j^i \leq 1$  we have

$$p_i^j = \frac{\binom{n}{k-j} p_j^i}{\binom{n}{k-i}} \leq \frac{(n-k+i)!(k-i)!}{(n-k+j)!(k-j)!} = O(n^{-(j-i)}).$$

□

We call this Markov chain the *level chain* and we call the space of real vectors of length  $k$  the *level space*<sup>6</sup>. The level chain is illustrated in Fig. 2. The transient matrix  $P$  of the level chain has a size of  $k \times k$ . The matrix  $P$  (unlike  $Q$ ) is not symmetric. In our analysis we use the following property of the matrix  $P$ .

**Lemma 13.** *The sum of each row of  $P$  is  $1 - O(\frac{1}{n})$ .*

---

<sup>6</sup>As well as for the individual space, the dimension of the level space is equal to the number of the transient states of the level chain and matrix  $P$  defines a linear operator on this space.



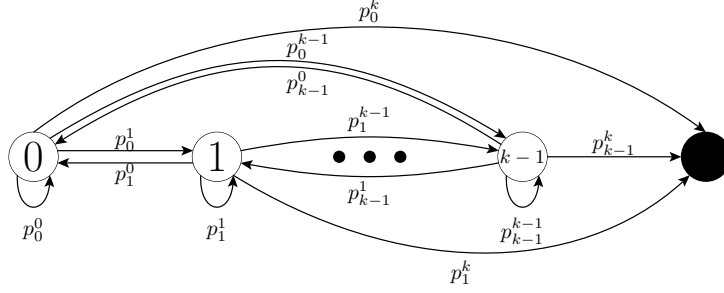


Figure 2: Illustration of the level chain. The black circle represents the optimum that is an absorbing state. The states  $[0..k-1]$  represent the levels of the plateau surrounding the optimum.

*Proof.* The sum of the  $i$ -th row of  $P$  is

$$\sum_{j=0}^{k-1} p_i^j = 1 - p_i^k, \quad (3)$$

since the sum of all the outgoing probabilities for each state in the original Markov chain is one. By Lemma 12 we have  $p_i^k = O(n^{-(k-i)}) = O(\frac{1}{n})$ .  $\square$

There is a natural mapping from the level space to the individual space. Every vector  $x = (x_0, \dots, x_{k-1})$  can be mapped to the vector  $\phi(x) = (y_0, \dots, y_{N-1})$ , where  $y_i = x_j / \binom{n}{k-j}$ , if the  $i$ -th element belongs to the  $j$ -th level. If  $x$  is a distribution over the levels, that is,  $x \in [0, 1]^k$  and  $\|x\|_1 = 1$ , then  $\phi(x)$  is the distribution over the elements of the plateau which is uniform on the levels and which has the same total mass on each level as  $x$ . This mapping has several useful properties.

**Lemma 14.**  $\phi$  is linear, that is, we have  $\phi(\alpha x + \beta y) = \alpha \phi(x) + \beta \phi(y)$  for all  $x, y \in \mathbb{R}^k$  and all  $\alpha, \beta \in \mathbb{R}$ .

This property follows directly from the definition of  $\phi$ .

**Lemma 15.** For all  $x \in \mathbb{R}^k$  we have  $\phi(xP) = \phi(x)Q$ .

*Proof.* In informal words, this property holds because both matrices  $P$  and  $Q$  represent the same operator, but in different spaces. Thus, the result of applying this operator to some vector and then switching the space is the same as performing these two actions in a reversed order.

For the formal proof, recall that *level*  $i$  is the set of all individuals in distance  $(k - i)$  from the optimum. We use the fact that for any individual  $m$  in level  $j$  we have

$$p_j^i = \sum_{\ell \in \text{level } i} q_m^\ell,$$

where  $q_m^\ell$  is the element of matrix  $Q$ , that is, the probability to obtain individual  $\ell$  from individual  $m$ . From this and from the definition of  $\phi$  we calculate the  $m$ -th element of  $\phi(xP)$ , assuming that individual  $m$  belongs to level  $j$ .

$$(\phi(xP))_m = \frac{(xP)_j}{\binom{n}{k-j}} = \frac{\sum_{i=0}^{k-1} x_i p_i^j}{\binom{n}{k-j}}.$$

By Lemma 11 we have  $\binom{n}{k-i} p_i^j = \binom{n}{k-j} p_j^i$ . Therefore,

$$(\phi(xP))_m = \sum_{i=0}^{k-1} \frac{x_i}{\binom{n}{k-i}} p_j^i = \sum_{i=0}^{k-1} \frac{x_i}{\binom{n}{k-i}} \sum_{\ell \in \text{level } i} q_m^\ell.$$

Recall that  $q_m^\ell = q_\ell^m$  for all  $\ell, m \in [0..N - 1]$ . Hence, we have

$$(\phi(xP))_m = \sum_{i=0}^{k-1} \frac{x_i}{\binom{n}{k-i}} \sum_{\ell \in \text{level } i} q_\ell^m = \sum_{\ell=0}^{N-1} (\phi(x))_\ell \cdot q_\ell^m = (\phi(x)Q)_m.$$

□

**Lemma 16.** *The spectrum  $\sigma(P)$  of the matrix  $P$  is a subset of the spectrum  $\sigma(Q)$  of the matrix  $Q$ . For any eigenvector  $x$  of the matrix  $P$  the vector  $\phi(x)$  is an eigenvector of  $Q$ .*

*Proof.* From Lemma 14 and Lemma 15 it follows that if  $x$  is an eigenvector of  $P$ , then  $\phi(x)$  is an eigenvector of  $Q$  with the same eigenvalue. Thus, every eigenvalue of  $P$  is an eigenvalue of  $Q$ . □

**Lemma 17.** *For all  $x \in \mathbb{R}^k$ , the Manhattan norm is invariant under  $\phi$ , that is,  $\|x\|_1 = \|\phi(x)\|_1$ .*

This follows from the fact that all components of  $\phi(x)$  that are from the same level have the same sign. Notice that an analogous property does not hold for the Euclidean norm  $\|\cdot\|_2$ .

Although the two Markov chains represent the same process and each of them contains all information about it, in our analysis we need to use both

of them simultaneously. We do not really work with the whole individual space, but only with its subspace  $\phi(S)$ , where  $S$  is the level space. Hence, it is natural to use the terms of the level space to simplify the computations and make them easier to understand. On the other hand, we cannot prove some essential facts about the operator  $\mathcal{P}$  represented by the matrix  $P$ , e.g., that there exists a basis of the level space which consists of the eigenvectors of  $\mathcal{P}$  (see Lemma 20). To prove them we have to switch to the individual space (or more precisely, to its subspace  $\phi(S)$ ) and use the properties of the self-adjoint operator  $\mathcal{Q}$  represented by the symmetric matrix  $Q$ . Therefore, both chains and their transient matrices are indispensable in our analysis.

### 3.4 The Spectrum of the Transient Matrix

The main result of this section is the following analysis of the eigenvalues of  $P$ , which builds on the interplay between the two Markov chains.

**Lemma 18.** *Let  $P$  be the transient matrix of the level chain. Then the following three properties hold.*

1. *All eigenvalues of  $P$  are real.*
2. *The largest eigenvalue  $\lambda_0$  of  $P$  satisfies  $\lambda_0 = 1 - O(1/n)$ .*
3. *Let  $\Pr[\alpha = 1] > 0$  and  $\Pr[\alpha = 1] = \omega(1/\sqrt[k-1]{n})$ . Then with  $c := \Pr[\alpha = 1]$  and with  $\varepsilon := \frac{c^{k-1}}{(k-1)2^k}$  any other eigenvalue  $\lambda' \neq \lambda_0$  of  $P$  satisfies  $|\lambda'| < 1 - \varepsilon$ .*

*Proof.* The fact that the eigenvalues are real follows from the facts that by Lemma 16 the spectrum of  $P$  is a subset of the spectrum of  $Q$  and that by Lemma 2 all eigenvalues of the symmetric matrix  $Q$  are real.

The largest eigenvalue  $\lambda_0$  of  $P$  is bounded by the minimal and the maximal row sum of  $P$  (see Theorem 4), which are both  $1 - O(1/n)$  by Lemma 13.

It remains to show that the absolute values of all other eigenvalues are less than  $1 - \varepsilon$  for  $\varepsilon = \frac{c^{k-1}}{(k-1)2^k}$ , which requires more work. To prove this statement we perform a precise analysis of the characteristic polynomial of  $P$ .

Recall that the spectrum of  $P$  is the set of the roots of its characteristic polynomial

$$\chi_P(\lambda) = \det(P - \lambda I) = \sum_{\sigma \in S_k} \text{sgn}(\sigma) \prod_{i=0}^{k-1} (P - \lambda I)_{i, \sigma(i)},$$

where  $S_k$  is the set of all permutations of the set  $[0..k-1]$  and  $\text{sgn}(\sigma)$  denotes the *signature* of permutation  $\sigma$  (that is,  $+1$  if it can be obtained from the identity permutation in even number of element swaps, and  $-1$  otherwise). Note that for all permutations except the identity the product in the sum contains at least one factor  $(P - \lambda I)_{i,j}$  with  $j > i$  and this element satisfies  $(P - \lambda I)_{i,j} = p_i^j = O(1/n)$  by Lemma 12. The other factors of the product are either  $p_{i'}^{j'}$  or  $(p_{i'}^{j'} - \lambda)$  for some  $i', j'$ , therefore every product where  $\sigma$  is not the identity is a polynomial in  $\lambda$  with coefficients which are  $O(1/n)$ . Thus, the characteristic polynomial can be written as

$$\chi_P(\lambda) = \prod_{i=0}^{k-1} (p_i^i - \lambda) + \beta(\lambda), \quad (4)$$

where  $\beta(\lambda)$  is some polynomial in  $\lambda$  with coefficients that are all  $O(1/n)$ . For this reason the derivative  $\beta'(\lambda)$  will also be  $O(1/n)$  for all  $\lambda \in [-1, 1]$ , where we recall that all asymptotics are for  $n \rightarrow \infty$  (and, e.g., not for any limit behavior of  $\lambda$ ).

To prove that for all eigenvalues  $\lambda' \neq \lambda_0$  we have  $\lambda' < 1 - \varepsilon$  we need to prove that there is no more than one root of the characteristic polynomial in  $[1 - \varepsilon, 1]$ . To do so it suffices to prove that  $\chi_P(\lambda)$  is strictly monotonic in this segment.

Consider  $\lambda \geq 1 - \varepsilon$ . This implies that  $\lambda \geq 1 - \frac{c}{2}$ . For every  $i \neq 0$  we have  $p_i^i \leq 1 - \Pr[\alpha = 1] = 1 - c$ . Thus, for every  $i \neq 0$  and any  $\lambda \geq 1 - \varepsilon$  we have

$$(p_i^i - \lambda) \leq -c/2. \quad (5)$$

By (4), the derivative of  $\chi_P(\lambda)$  can be written as

$$\chi'_P(\lambda) = (p_0^0 - \lambda) \left( \prod_{i=1}^{k-1} (p_i^i - \lambda) \right)' - \prod_{i=1}^{k-1} (p_i^i - \lambda) + \beta'(\lambda). \quad (6)$$

Recall that  $\varepsilon = \frac{c^{k-1}}{(k-1)2^k}$ . For all  $\lambda \geq 1 - \varepsilon$  we have

$$\begin{aligned} p_0^0 - \lambda &\leq 1 - \left( 1 - \frac{c^{k-1}}{(k-1)2^k} \right) = \frac{c^{k-1}}{(k-1)2^k}, \\ \left| \left( \prod_{i=1}^{k-1} (p_i^i - \lambda) \right)' \right| &= \left| - \sum_{i=1}^{k-1} \prod_{\substack{j \in [1..k-1] \\ j \neq i}} (p_j^j - \lambda) \right| \leq (k-1), \\ \left| \prod_{i=1}^{k-1} (p_i^i - \lambda) \right| &\geq \frac{c^{k-1}}{2^{k-1}}, \end{aligned}$$

where the last inequality follows from (5). Furthermore, from (5) it also follows that for  $i \neq 0$  we have  $(p_i^i - \lambda) < 0$ . Thus,

$$\text{sign} \left( \left( \prod_{i=1}^{k-1} (p_i^i - \lambda) \right)' \right) = \text{sign} \left( - \sum_{i=1}^{k-1} \prod_{\substack{j \in [1..k-1] \\ j \neq i}} (p_j^j - \lambda) \right) = (-1)^{k-1}. \quad (7)$$

Consequently, we have two cases.

Case 1: When  $p_0^0 - \lambda \geq 0$ , we have

$$\begin{aligned} & \left| (p_0^0 - \lambda) \left( \prod_{i=1}^{k-1} (p_i^i - \lambda) \right)' - \prod_{i=1}^{k-1} (p_i^i - \lambda) \right| \\ & \geq \left| \prod_{i=1}^{k-1} (p_i^i - \lambda) \right| - \left| (p_0^0 - \lambda) \left( \prod_{i=1}^{k-1} (p_i^i - \lambda) \right)' \right| \\ & \geq \left| \frac{c}{2} \right|^{k-1} - \frac{c^{k-1}}{(k-1)2^k} \sum_{i=1}^{k-1} \left| \prod_{\substack{j \in [1..k-1] \\ j \neq i}} (p_j^j - \lambda) \right| \\ & \geq \frac{c^{k-1}}{2^{k-1}} - (k-1) \frac{c^{k-1}}{(k-1)2^k} = \frac{c^{k-1}}{2^k}. \end{aligned}$$

Hence, we have

$$|\chi'_P(\lambda)| = \frac{c^{k-1}}{2^k} + O(1/n) = \frac{\omega(1/n^{\frac{1}{k-1}})^{k-1}}{2^k} + O(1/n) = \omega(1/n).$$

Case 2: When  $p_0^0 - \lambda < 0$ , since by (5)  $(p_i^i - \lambda) < 0$  for all  $i \neq 0$ , we have

$$\left| (p_0^0 - \lambda) \left( \prod_{i=1}^{k-1} (p_i^i - \lambda) \right)' - \prod_{i=1}^{k-1} (p_i^i - \lambda) \right| \geq \left| \prod_{i=1}^{k-1} (p_i^i - \lambda) \right| \geq \frac{c^{k-1}}{2^{k-1}}.$$

Therefore,

$$|\chi'_P(\lambda)| = \omega(1/n).$$

For  $n$  large enough, this together with (6) and (7) implies that  $\chi'_P(\lambda)$  has the same sign as  $(-1)^{k-1}$  for every  $\lambda \in [1 - \varepsilon, 1]$ . Thus, there can be only one root of characteristic polynomial in this segment.

To rule out that there is a negative eigenvalue  $\lambda$  with  $|\lambda| > 1 - \varepsilon$ , we notice that for  $\lambda < -\frac{1}{2}$  and for every  $i$  we have  $(p_i^i - \lambda) > \frac{1}{2}$ . Therefore,  $|\chi_P(\lambda)| > \left(\frac{1}{2}\right)^k - o(1)$ , and thus there are no roots that are less than  $-\frac{1}{2}$  when  $n$  is large enough.

This finally shows that for all eigenvalues  $\lambda' \neq \lambda_0$  we have  $|\lambda'| < 1 - \min\left(\frac{1}{2}, \frac{c^{k-1}}{(k-1)2^k}\right) = 1 - \frac{c^{k-1}}{(k-1)2^k}$ .

□

## 4 Runtime Analysis

In this section, we prove our main result, which determines the runtime of the  $(1 + 1)$  EA on the PLATEAU function.

**Theorem 19.** *Consider the  $(1 + 1)$  EA using any unbiased mutation operator such that the probability to flip exactly one bit is at least  $\Pr[\alpha = 1] = \omega\left(n^{-\frac{1}{2k-2}}\right)$ . Let  $T$  denote the runtime of this algorithm starting on an arbitrary search point of the plateau of the PLATEAU $_k$  function. Then*

$$E[T] = \frac{n^k}{\Pr[1 \leq \alpha \leq k]k!}(1 + o(1)),$$

$$\Pr[T > t] = (1 \pm o(1)) \left(1 - \frac{k! \Pr[1 \leq \alpha \leq k]}{n^k}(1 \pm o(1))\right)^t + r(t),$$

where  $|r(t)| \leq \sqrt{N}(1 - \varepsilon)^t$ ,  $\varepsilon = \frac{(\Pr[\alpha=1])^{k-1}}{(k-1)2^k}$ , and  $N = \frac{n^k}{k!}(1 \pm o(1))$ . All asymptotic notation refers to  $n \rightarrow \infty$  and is independent of  $t$ .

We start with a few preparatory results. Recall that by Theorem 4 the largest eigenvalue of a positive matrix has a one-dimensional eigenspace. Also this theorem asserts that both left and right eigenvectors that correspond to the largest eigenvalue have all components with the same sign and they do not have any zero component. Let  $\pi^*$  be such a left eigenvector with positive components for  $P$  and let it be normalized in such way that  $\|\pi^*\|_1 = 1$ . We view  $\pi^*$  as distribution over the levels of the plateau and call it the *conditional stationary distribution of  $P$*  since it does not change in one iteration under the condition that the algorithm does not find the optimum. Also let  $u = (u_0, \dots, u_{k-1})$  be the probability distribution in the level space such that  $\phi(u)$  is the uniform distribution in the individual space. Hence

$$u_i = \binom{n}{k-i} / \sum_{j=0}^{k-1} \binom{n}{k-j} = \binom{n}{k-i} N^{-1}$$

for all  $i \in [0..k-1]$ . Our next target is showing that  $\pi^*$  and  $u$  are asymptotically equal. For this, we need the following basis of the level space.

**Lemma 20.** *There exists a basis  $\{e^i\}_{i=0}^{k-1}$  of the level space with the following properties.*

1.  $\pi^* = e^0$ .
2.  $e^i$  is an eigenvector of  $P$  for all  $i \in [0..k-1]$ .
3. The  $\phi(e^i)$  are orthogonal in the individual space.

*Proof.* Let  $S$  be the level space and  $S_{\text{ind}}$  be the individual space. Then  $\phi(S)$  is a subspace of  $S_{\text{ind}}$  with  $\dim \phi(S) = k$ , since the kernel of  $\phi$  is trivial.

Consider the operator  $\mathcal{Q}$  that is represented by the matrix  $Q$ . It is a self-adjoint operator on  $S_{\text{ind}}$ , since its matrix is symmetric. Moreover, this operator maps  $\phi(S)$  into  $\phi(S)$ , since for all  $x \in S$  by Lemma 15 we have  $\mathcal{Q}(\phi(x)) = \phi(x)Q = \phi(xP)$ . Therefore,  $\mathcal{Q}$  is a self-adjoint operator on  $\phi(S)$ . Thus, by Lemma 10 there exists an orthonormal basis  $f^0, \dots, f^{k-1}$  of  $\phi(S)$  that consists of eigenvectors of  $\mathcal{Q}$ . Let  $e^i = \phi^{-1}(f^i)$  for all  $i \in [0..k-1]$ . By Lemma 16 the  $e^i$  are eigenvectors of  $P$ . By the linearity of  $\phi$  (Lemma 14) they are linearly independent, hence they form a basis of  $S$ .

By assuming that  $e^0$  corresponds to the largest eigenvalue and multiplying  $e^0$  by a suitable scalar, we also satisfy the first property of the lemma.  $\square$

We use the basis from Lemma 20 to prove that  $\phi(\pi^*)$  is very close to the uniform distribution.

**Lemma 21.** *If  $\Pr[\alpha = 1] = \omega\left(n^{-\frac{1}{2k-2}}\right)$ , then for all  $j \in [0..k-1]$ , we have  $\pi_j^* = u_j(1 + \gamma_j)$ , where  $|\gamma_j| \leq \gamma$  for some  $\gamma = o(1)$ .*

*Proof.* Figure 3 illustrates the relation of the terms used in this proof to make it easier to follow. Lemma 20, there exist unique  $c_0, c_1, \dots, c_{k-1} \in \mathbb{R}$  such that

$$u = \sum_{i=0}^{k-1} c_i e^i.$$

If we transfer this decomposition into the individuals space (using the linearity of  $\phi$ , see Lemma 14), we obtain

$$U = \sum_{i=0}^{k-1} U^i,$$





We also compute the denominator of (8) as

$$\|U^0\|_1 = \sum_{m=0}^{N-1} |U_m^0| = \sum_{m=0}^{N-1} \frac{1}{N} (1 - \beta_m) = 1 - \sum_{m=0}^{N-1} \frac{1}{N} \beta_m.$$

Putting this into (8), we obtain

$$\Pi_m^* = U_m \frac{1 - \beta_m}{1 - \sum_{m=0}^{N-1} \frac{1}{N} \beta_m}. \quad (9)$$

In the remainder of the proof we aim at bounding  $|\beta_m|$  from above by some  $\beta = o(1)$ . Then (9) gives

$$\frac{1 - \beta_m}{1 - \sum_{m=0}^{N-1} \frac{1}{N} \beta_m} \in \left[ \frac{1 - \beta}{1 + \beta}, \frac{1 + \beta}{1 - \beta} \right] \subset \left[ 1 - \frac{2\beta}{1 - \beta}, 1 + \frac{2\beta}{1 - \beta} \right],$$

hence defining  $\gamma = \frac{2\beta}{1 - \beta} = o(1)$  proves the lemma.

To find the desired  $\beta$  with  $|\beta_m| < \beta = o(1)$ , we regard the vector  $U - UQ$ . On the one hand, its elements are very small. For all  $m \in [0..N-1]$ , we have

$$(U - UQ)_m = U_m - \sum_{\ell=0}^{N-1} q_\ell^m U_\ell = \frac{1}{N} \left( 1 - \sum_{\ell=0}^{N-1} q_\ell^m \right) = \frac{q_m^N}{N},$$

where  $q_\ell^m$  is the probability to go from individual  $\ell$  to individual  $m$  (recall that  $q_\ell^m = q_m^\ell$ ) and  $q_m^N$  is the probability to leave the plateau from the  $m$ -th individual. The Euclidean norm of this vector is also very small.

$$\begin{aligned} \|U - UQ\|_2 &= \sqrt{\sum_{m=0}^{N-1} \left( \frac{q_m^N}{N} \right)^2} \\ &= \frac{1}{N} \sqrt{\sum_{i=0}^{k-1} \binom{n}{k-i} \left( \Pr[\alpha = k-i] / \binom{n}{k-i} \right)^2}. \end{aligned}$$

To bound the sum under the root we notice that it is maximized when we maximize  $\Pr[\alpha = 1]$  (since we consider only constant  $k$ , we assume that  $n$  is large enough so that  $n > 2k$ ). Let this probability be equal to 1, then we have only one non-zero summand and hence,

$$\|U - UQ\|_2 \leq \frac{1}{N} \sqrt{\binom{n}{1}^{-1}} = \frac{1}{\sqrt{n}N}.$$

On the other hand, if we recall that the  $U^i$  are eigenvectors, then we have

$$U - UQ = \sum_{i=0}^{k-1} U^i - \sum_{i=0}^{k-1} \lambda_i U^i = \sum_{i=0}^{k-1} (1 - \lambda_i) U^i.$$

As the  $U^i$  are orthogonal, for every  $i \in [0..k-1]$  we have

$$\|(1 - \lambda_i)U^i\|_2 \leq \|U - UQ\|_2 \leq \frac{1}{\sqrt{nN}}.$$

Since the absolute value of every component of a vector cannot be larger than its Euclidean norm, for all  $m \in [0..N-1]$  we conclude that

$$|(1 - \lambda_i)U_m^i| \leq \|(1 - \lambda_i)U^i\|_2 \leq \frac{1}{\sqrt{nN}}.$$

Recall that by Lemma 18 we have

$$(1 - \lambda_i) > \varepsilon > \frac{(\Pr[\alpha = 1])^{k-1}}{(k-1)2^k}$$

for all  $i \neq 0$ . Consequently,

$$\begin{aligned} |\beta_m| &= \left| N \sum_{i=1}^{k-1} U_m^i \right| \leq N \sum_{i=1}^{k-1} |U_m^i| \leq N \sum_{i=1}^{k-1} \frac{1}{\sqrt{nN}(1 - \lambda_i)} \\ &\leq \frac{(k-1)}{\sqrt{n\varepsilon}} =: \beta. \end{aligned}$$

Since we have  $\Pr[\alpha = 1] = \omega(n^{-\frac{1}{2k-2}})$ , we conclude that

$$\varepsilon = \frac{\left(\omega(n^{-\frac{1}{2k-2}})\right)^{k-1}}{(k-1)2^k} = \omega(1/\sqrt{n}).$$

Thus,

$$\beta = \frac{(k-1)}{\sqrt{n\varepsilon}} = o(1)$$

as desired. □

We are now in the position to prove our main result.

*Proof of Theorem 19.* To prove the theorem we first estimate the probability that the runtime  $T$  is greater than  $t$  by  $\Pr[T > t] = \|\pi P^t\|_1$ , where  $\pi$  is the initial distribution over the levels of the plateau (see Section 3.2). Then by (1) we estimate the expected runtime as  $E[T] = \sum_{t=1}^{+\infty} \|\pi P^{t-1}\|_1$ .

To analyse  $\|\pi P^t\|_1$ , we decompose  $\pi$  into a sum of eigenvectors of  $P$  using the basis  $e^0, \dots, e^{k-1}$  from Lemma 20. Let  $\pi^0, \dots, \pi^{k-1}$  be scalar multiples of  $e^0, \dots, e^{k-1}$  such that

$$\pi = \sum_{i=0}^{k-1} \pi^i. \quad (10)$$

Using the triangle inequalities (Lemma 6) we obtain

$$\|\pi^0 P^t\|_1 - \left\| \sum_{i=1}^{k-1} \pi^i P^t \right\|_1 \leq \|\pi P^t\|_1 \leq \|\pi^0 P^t\|_1 + \left\| \sum_{i=1}^{k-1} \pi^i P^t \right\|_1.$$

Since the  $\pi^i$  are the eigenvectors of  $P$ , we have

$$\lambda_0^t \|\pi^0\|_1 - \left\| \sum_{i=1}^{k-1} \lambda_i^t \pi^i \right\|_1 \leq \|\pi P^t\|_1 \leq \lambda_0^t \|\pi^0\|_1 + \left\| \sum_{i=1}^{k-1} \lambda_i^t \pi^i \right\|_1. \quad (11)$$

Now we estimate the “error term”  $\left\| \sum_{i=1}^{k-1} \lambda_i^t \pi^i \right\|_1$  of these bounds. First, by Lemma 17 and by Lemma 14, we have

$$\left\| \sum_{i=1}^{k-1} \lambda_i^t \pi^i \right\|_1 = \left\| \phi \left( \sum_{i=1}^{k-1} \lambda_i^t \pi^i \right) \right\|_1 = \left\| \sum_{i=1}^{k-1} \lambda_i^t \phi(\pi^i) \right\|_1. \quad (12)$$

Using Lemma 5 and Lemma 7 we estimate

$$\left\| \sum_{i=1}^{k-1} \lambda_i^t \phi(\pi^i) \right\|_1 \leq \sqrt{N} \left\| \sum_{i=1}^{k-1} \lambda_i^t \phi(\pi^i) \right\|_2 \leq \sqrt{N} \max_{i \in [1..k-1]} (|\lambda_i^t|) \left\| \sum_{i=1}^{k-1} \phi(\pi^i) \right\|_2. \quad (13)$$

By Lemma 18 we have  $\max_{i \in [1..k-1]} (|\lambda_i^t|) \leq (1 - \varepsilon)^t$ , where  $\varepsilon = \frac{(\Pr[\alpha=1])^{k-1}}{(k-1)2^k}$ . Hence, by Lemma 8 and Lemma 5, we conclude

$$\sqrt{N} \max_{i \in [1..k-1]} (|\lambda_i^t|) \left\| \sum_{i=1}^{k-1} \phi(\pi^i) \right\|_2 \leq \sqrt{N} (1 - \varepsilon)^t \|\phi(\pi)\|_1 = \sqrt{N} (1 - \varepsilon)^t. \quad (14)$$

Finally, by (11), (12), (13), and (14) we obtain that

$$\|\pi P^t\|_1 = \lambda_0^t \|\pi^0\|_1 + r(t), \quad (15)$$

where  $|r(t)| \leq \sqrt{N}(1 - \varepsilon)^t$ .

To estimate the expected runtime we put (15) into (1) and obtain

$$\begin{aligned} E[T] &= \sum_{t=1}^{+\infty} \|\pi P^{t-1}\|_1 = \sum_{t=1}^{+\infty} (\lambda_0^{t-1} \|\pi^0\|_1 + r(t-1)) \\ &= \sum_{t=1}^{+\infty} \lambda_0^{t-1} \|\pi^0\|_1 + \sum_{t=1}^{+\infty} r(t-1) = \frac{\|\pi^0\|_1}{1 - \lambda_0} + \sum_{t=0}^{+\infty} r(t). \end{aligned} \quad (16)$$

By (14) we have

$$\left| \sum_{t=0}^{+\infty} r(t) \right| \leq \sum_{t=0}^{+\infty} |r(t)| \leq \sum_{t=1}^{+\infty} \sqrt{N}(1 - \varepsilon)^t = \frac{(1 - \varepsilon)\sqrt{N}}{\varepsilon} \leq \frac{\sqrt{N}}{\varepsilon}.$$

From the assumptions of the theorem we have

$$\varepsilon = \frac{(\Pr[\alpha = 1])^{k-1}}{(k-1)2^k} = \frac{\omega(1/\sqrt{n})}{(k-1)2^k} = \omega(1/\sqrt{n}),$$

and hence

$$\left| \sum_{t=0}^{+\infty} r(t) \right| = o(\sqrt{nN}).$$

It remains to estimate  $\|\pi^0\|_1$  and  $\lambda_0$ . Recall that by Lemma 17 we have  $\|\pi^0\|_1 = \|\phi(\pi^0)\|_1$ . From the linearity of  $\phi$  (see Lemma 14) and (10) we obtain  $\phi(\pi) = \sum_{i=0}^{k-1} \phi(\pi^i)$ . Since  $\pi^0, \dots, \pi^{k-1}$  are scalar multiples of  $e^0, \dots, e^{k-1}$  and all  $\phi(e^i)$  are orthogonal, we have a decomposition of  $\phi(\pi)$  into a sum of orthogonal vectors. Therefore, by Lemma 9, we have

$$\|\phi(\pi^0)\|_1 = \frac{\langle \phi(\pi), \phi(e^0) \rangle}{\langle \phi(e^0), \phi(e^0) \rangle}. \quad (17)$$

By Lemma 21, the components of  $e^0$  are almost equal to the components of the vector  $u$  of the uniform distribution in the level space<sup>7</sup>. Transferring

---

<sup>7</sup>Note that  $e^0$  is the same vector as  $\pi^*$  in Lemma 21. However we now refer to this vector as  $e^0$  to underline that we consider it as a basis vector of the level space, while in Lemma 21 we referred to it as  $\pi^*$  since we considered it as a vector of the probabilistic distribution over the states of the level chain.

this result to the individual space, we have  $(\phi(e^0))_m = \frac{1}{N}(1 + \gamma_j)$  for all  $j \in [0..k-1]$  and all individuals  $m$  that belong to level  $j$ . This and the fact that by Lemma 17 we have  $\|\phi(\pi)\|_1 = \|\pi\|_1 = 1$  allow to calculate the inner products in (17) and obtain

$$\|\pi^0\|_1 = \frac{\sum_{j=0}^{k-1} \sum_{m \in \text{level } j} (\phi(\pi))_m \frac{1+\gamma_j}{N}}{\sum_{j=0}^{k-1} \sum_{m \in \text{level } j} \left(\frac{1+\gamma_j}{N}\right)^2} = 1 + o(1). \quad (18)$$

We compute  $(1 - \lambda_0)$  in the following way. First, since  $\|\pi^*\|_1 = 1$  and  $\pi^*$  is an eigenvector of  $P$ , we have

$$1 - \lambda_0 = 1 - \|\lambda_0 \pi^*\|_1 = 1 - \|\pi^* P\|_1 = 1 - \sum_{i=0}^{k-1} \left| \sum_{j=0}^{k-1} \pi_j^* p_j^i \right|.$$

Since by Theorem 4 all components of  $\pi^*$  are positive and the components of  $P$  are non-negative, this simplifies to

$$1 - \sum_{i=0}^{k-1} \left| \sum_{j=0}^{k-1} \pi_j^* p_j^i \right| = 1 - \sum_{i=0}^{k-1} \sum_{j=0}^{k-1} \pi_j^* p_j^i = 1 - \sum_{j=0}^{k-1} \pi_j^* \sum_{i=0}^{k-1} p_j^i.$$

By the definition of  $P$ , the sum of the  $j$ -th row of  $P$  is equal to  $(1 - p_j^k)$ , and we have  $\sum_{i=0}^{k-1} \pi_i^* = \|\pi^*\|_1 = 1$ . Hence,

$$1 - \sum_{j=0}^{k-1} \pi_j^* \sum_{i=0}^{k-1} p_j^i = 1 - \sum_{j=0}^{k-1} \pi_j^* (1 - p_j^k) = \sum_{j=0}^{k-1} \pi_j^* p_j^k.$$

By Lemma 21 and (2) we have

$$\begin{aligned} \sum_{j=0}^{k-1} \pi_j^* p_j^k &= \sum_{j=0}^{k-1} \binom{n}{k-j} N^{-1} (1 + \gamma_j) \binom{n}{k-j}^{-1} \Pr[\alpha = k-j] \\ &= \frac{1}{N} \sum_{j=1}^k \Pr[\alpha = j] (1 + \gamma_{k-j}) = \frac{1}{N} \Pr[1 \leq \alpha \leq k] (1 + o(1)). \end{aligned}$$

Thus, we obtain

$$\lambda_0 = 1 - \frac{1}{N} \Pr[1 \leq \alpha \leq k] (1 + o(1)). \quad (19)$$

By substituting  $\lambda_0$  and  $\|\pi^0\|_1$  into (16) and (15) with their values from (19) and (18) and recalling that  $N = \frac{n^k}{k!} (1 + o(1))$ , we prove the theorem.  $\square$

We also underline that  $r(t)$  in the tail bounds on the runtime distribution is negligible, as soon as  $t = \omega(\sqrt{n} \log(n))$ , that is, far before the algorithm finds the optimum.

## 5 Corollaries

We now exploit Theorem 19 to analyze how the choice of the mutation operator influences the runtime.

By the runtime in this section we mean the runtime of the algorithm when it starts from an arbitrary individual, which is not necessarily on the plateau. However, without proof we notice that the  $(1 + 1)$  EA with any mutation operator considered in this section reaches the plateau in an expected number of  $O(n \log(n))$  iterations from any starting individual, which is significantly less than the time which it spends on the plateau. Therefore, the time to leave the plateau coincides with the total runtime precisely apart from lower order terms. Since, by our main result, the time to leave the plateau depends only on the probability to flip between 1 and  $k$  bits, determining the runtimes in this section is an easy task.

We first observe that for all unbiased operators with constant probability to flip exactly one bit, the expected optimization time is  $\Theta(N)$ , where we recall that the size  $N$  of the plateau is

$$N = \sum_{i=0}^{k-1} \binom{n}{n-k+i} = (1 \pm o(1)) \frac{n^k}{k!}.$$

Hence all these mutation operators lead to asymptotically the same runtime of  $\Theta(n^k)$ . The interesting aspect thus is how the leading constant changes.

### 5.1 Randomized Local Search and Variants

When taking such a more precise look at the runtime, that is, including the leading constant, then the best runtime, obviously, is obtained from mutation operators which flip always between 1 and  $k$  bits. This includes variants of randomized local search which also flip more than one bit, see, e.g., [GW03, NW07, DDY20], as long as they do not flip more than  $k$  bits, but most prominently the classic randomized local search heuristic, which always flips a single random bit. Note that the latter uniformly for all  $k$  (and including the case  $k = 1$  not regarded in this work) is among the most effective algorithms.

## 5.2 Standard $(1 + 1)$ EA

The classic mutation operator in evolutionary computation is *standard bit mutation*, where each bit is flipped independently with some probability (“mutation rate”)  $\gamma/n$ , where  $\gamma$  usually is a constant. We call the  $(1 + 1)$  EA which uses the standard bit mutation the *standard  $(1 + 1)$  EA*.

**Theorem 22.** *Let  $\gamma$  be some arbitrary positive constant and  $k \geq 2$ . Then the standard  $(1 + 1)$  EA with mutation rate  $\gamma/n$  optimizes  $\text{PLATEAU}_k$  in an expected number of*

$$E[T] = (1 + o(1)) \frac{n^k}{k! e^{-\gamma} \sum_{i=1}^k \frac{\gamma^i}{i!}}$$

*iterations. This time is asymptotically minimal for  $\gamma = \sqrt[k]{k!} \approx k/e$ .*

*Proof.* For the standard bit mutation with mutation rate  $\gamma/n$ , the probability to flip exactly one bit is

$$n \frac{\gamma}{n} \left(1 - \frac{\gamma}{n}\right)^{n-1} \geq \gamma e^{-\gamma} (1 - o(1)),$$

which is at least some positive constant as long as  $\gamma$  is a constant. Thus, we can apply Theorem 19 and obtain

$$\begin{aligned} E[T] &= (1 \pm o(1)) \frac{N}{\Pr[1 \leq \alpha \leq k]} \\ &= (1 \pm o(1)) N \left( \sum_{i=1}^k \binom{n}{i} \left(\frac{\gamma}{n}\right)^i \left(1 - \frac{\gamma}{n}\right)^{n-i} \right)^{-1} \\ &= (1 \pm o(1)) \frac{n^k}{k!} \left( \sum_{i=1}^k \frac{\gamma^i}{i!} e^{-\gamma} \right)^{-1}. \end{aligned}$$

Consider  $d(\gamma) = e^{-\gamma} \sum_{i=1}^k \frac{\gamma^i}{i!}$ . In order to minimize  $E[T]$ , we have to maximize  $d(\gamma)$ . Now  $\gamma \mapsto d(\gamma)$  is a smooth continuous function, so its maximal value for  $\gamma \in [0, +\infty)$  can only be at  $\gamma = 0$ , for  $\gamma \rightarrow +\infty$ , or in the zeros of its derivative. We have  $d(0) = \lim_{\gamma \rightarrow \infty} d(\gamma) = 0$ . The derivative is

$$\begin{aligned} d'(\gamma) &= \left( e^{-\gamma} \sum_{i=1}^k \frac{\gamma^i}{i!} \right)' = e^{-\gamma} \sum_{i=1}^k \frac{i \gamma^{i-1}}{i!} - e^{-\gamma} \sum_{i=1}^k \frac{\gamma^i}{i!} \\ &= e^{-\gamma} \left( \sum_{i=0}^{k-1} \frac{\gamma^i}{i!} - \sum_{i=1}^k \frac{\gamma^i}{i!} \right) = e^{-\gamma} \left( 1 - \frac{\gamma^k}{k!} \right). \end{aligned}$$

Hence the only value of  $\gamma$  with  $d'(\gamma) = 0$  is  $\gamma = \sqrt[k]{k!}$ . For this value we have  $d(\sqrt[k]{k!}) > 0$ , so this defines the unique optimal mutation rate. Finally, by Stirling's formula  $k! \approx \sqrt{2\pi k} \left(\frac{k}{e}\right)^k$  we have  $\sqrt[k]{k!} \approx (2\pi k)^{\frac{1}{2k}} \frac{k}{e} \approx \frac{k}{e}$ .  $\square$

### 5.3 Fast $(1 + 1)$ EA

The *fast*  $(1 + 1)$  EA recently proposed in [DLMN17] is simply a  $(1 + 1)$  EA that uses standard bit mutation with a random mutation rate  $\gamma/n$  with  $\gamma \in [1..n/2]$  chosen according to a power-law distribution. More precisely, for a parameter  $\beta > 1$  which is assumed to be a constant (independent of  $n$ ), we have

$$\Pr[\gamma = i] = 0$$

for every  $i > n/2$  and  $i = 0$ , and

$$\Pr[\gamma = i] = i^{-\beta} / H_{n/2, \beta}$$

otherwise, where  $H_{n/2, \beta} := \sum_{i=1}^{n/2} i^{-\beta}$  is a generalized harmonic number.

**Theorem 23.** *For  $k \geq 2$  the expected runtime of the fast  $(1 + 1)$  EA on  $\text{PLATEAU}_k$  is  $C_{kn} \frac{n^k}{k!}$ , where  $C_{kn} := \frac{H_{n/2, \beta}}{H_{k, \beta}}(1 + o(1))$  can be bounded by constants, namely  $C_{kn} \in \left[ \frac{\frac{1}{\beta-1} - o(1)}{H_{k, \beta}}, \frac{\frac{1}{\beta-1} + 1}{H_{k, \beta}} \right]$ .*

*Proof.* From the definition of the fast  $(1 + 1)$  EA we have

$$\sum_{i=1}^k \Pr[\gamma = i] = \frac{\sum_{i=1}^k i^{-\beta}}{\sum_{i=1}^{n/2} i^{-\beta}} = \frac{H_{k, \beta}}{H_{n/2, \beta}}.$$

Since  $\beta > 1$  and  $k$  are constants,  $H_{k, \beta}$  is a constant as well. We estimate  $H_{n/2, \beta}$  through the corresponding integral.

$$\frac{1}{\beta-1} + 1 = \int_1^{+\infty} x^{-\beta} dx + 1 \geq H_{n/2, \beta} \geq \int_1^{n/2} x^{-\beta} dx = \frac{1 - (n/2)^{1-\beta}}{\beta-1} = \frac{1 - o(1)}{\beta-1}.$$

Notice that  $\Pr[\alpha = 1] = \frac{H_{1, \beta}}{H_{n/2, \beta}} = (H_{n/2, \beta})^{-1}$  is at least some constant. Thus, Theorem 19 gives an expected runtime of  $E[T] = \frac{H_{n/2, \beta}}{H_{k, \beta}} N(1 + o(1))$ , which we can estimate by

$$\frac{\frac{1-o(1)}{\beta-1}}{H_{k, \beta}} \frac{n^k}{k!} \leq E[T] \leq \frac{\frac{1}{\beta-1} + 1}{H_{k, \beta}} \frac{n^k}{k!} (1 + o(1)).$$

$\square$



## 5.4 Hyper-Heuristics

Hyper-heuristics are randomized search heuristics that combine, in a suitable and again usually randomized fashion, simple low-level heuristics. Despite many success stories in applications, their theoretical understanding is still very low and only the last few years have seen some first results. These exclusively regard simple  $(1 + 1)$  type hill-climbers which choose between different mutation operators as low-level heuristics. We now regard the hyper-heuristics discussed in [AL14] argue that for some of these, our method is applicable, whereas for others it is not clear how to do this.

Like almost all previous theoretical works, we regard as available low-level mutation operators one-bit flips (flipping a bit chosen uniformly at random) and two-bit flips (flipping two bits chosen uniformly at random from all 2-sets of bit positions). Hence the  $(1 + 1)$  hill-climber with this a hyper-heuristic selection between these two operators starts with a random search point and then repeats generating a new search point by applying one of the mutation operators (chosen according to the hyper-heuristic) and accepting the new search point if it has an at least as good fitness as the parent.

The most elementary hyper-heuristic called *simple random* in each iteration simply chooses one of the two available mutation operators with equal probability  $1/2$ . This compound mutation operator (choosing one randomly and applying it) still is a unary unbiased mutation operator, so our main result (Theorem 19) is readily applicable and gives the following result.

**Theorem 24.** *Consider the  $(1 + 1)$  hill-climber using the simple random hyper-heuristic to decide between the one-bit flip and the two-bit flip mutation operator. When started on an arbitrary point of the plateau, its runtime  $T$  on the  $\text{PLATEAU}_k$ ,  $k \geq 2$ , function satisfies*

$$E[T] = \frac{n^k}{k!}(1 \pm o(1)).$$

The more interesting hyper-heuristic *random gradient* in the first iteration chooses a random low-level heuristic. In each further iteration, it chooses the same low-level heuristic as in the previous iteration, if this has ended with a fitness gain, and it chooses again a random low-level heuristic otherwise. This way of performing mutation obviously cannot be described via a single unary unbiased operator. However, once the algorithm has reached the plateau, it can. The reason is that from that point on and until the optimum is found, no further improvements are found. Consequently, the algorithm reverts to the one using the *simple random* approach.

**Corollary 25.** *Consider the  $(1 + 1)$  hill-climber using the random gradient hyper-heuristic to decide between the one-bit flip and the two-bit flip mutation*

operator. When started on an arbitrary point of the plateau, its runtime  $T$  on the  $\text{PLATEAU}_k$ ,  $k \geq 2$ , function satisfies

$$E[T] = \frac{n^k}{k!}(1 + o(1)).$$

For two other common hyper-heuristics, we currently do not see how to apply our methods. The *permutation* heuristic initially fixes a permutation of the low-level heuristics and then repeatedly uses them in this order. The *greedy* heuristic uses, in each iteration, all available hyper-heuristics in parallel and proceeds with the best offspring produced (if it is at least as good as the parent). While we are optimistic that these heuristics lead to asymptotically the same runtimes as the two heuristics just analyzed, we cannot prove this since our main result is not applicable.

It has been observed in [LOW17] that, due to the generally low probability of finding an improvement, better results are obtained when the *random gradient* heuristic is used with a longer learning period, that is, the randomly chosen low-level heuristic is repeated for a phase of  $\tau$  iterations. If an improvement is found, a new phase with the same low-level heuristic is started. Otherwise, the next phase starts with a random operator. This idea was extended in [DLOW18] so that now a phase was called successful if within  $\tau$  iterations a certain number  $\sigma$  of improvements were obtained. This mechanism was more stable and allowed a self-adjusting choice of the previously delicate parameter  $\tau$ . Again, for these hyper-heuristics our results are not applicable.

## 5.5 Comparison for Concrete Values

Since the leading constants computed above, in their general form, are hard to compare, we now provide in Table 1 a few explicit values for specific algorithm parameters and plateau sizes.

## 6 Conclusion

In this paper we developed a new method to analyze the runtime of evolutionary algorithms on plateaus. This method does not depend on the particular mutation operator used by the EA as long as there is a sufficiently large probability to flip a single random bit. We performed a very precise analysis on the particular class of plateau functions, but we are optimistic that similar methods can be applied for the analysis of other plateaus. For example, Lemmas 18, 20 and 21 remain true for those plateaus of the function  $\text{XDIVK}$

Algorithm		$k = 2$	$k = 4$	$k = 6$
Random Local Search		1	1	1
(1 + 1) EA with standard bit mutation	Mutation rate $1/n$	1.812	1.591	1.582
	Mutation rate $k/(en)$	2.074	1.328	1.027
Fast Genetic Algorithm	$\beta = 1.5$	1.930	1.563	1.428
	$\beta = 2$	1.316	1.155	1.103
(1 + 1) EA with hyperheuristics	simple random	1	1	1
	random gradient	1	1	1

Table 1: Comparison of the leading constant in the expected runtime of the evolutionary algorithms with different mutation operators on the  $\text{PLATEAU}_k$  function, that is, the constant  $c$ , such that the expected runtime is  $c \frac{n^k}{k!} (1 - o(1))$ .

(that is defined as  $\lfloor \text{ONEMAX}(x)/k \rfloor$  for some parameter  $k$ ) that are in a constant Hamming distance from the optimum (and these are the plateaus which contribute most to the runtime). That said, the proof of Lemma 18 would need to be adapted to these plateaus different from the one of our plateau function. We are optimistic that this can be done, but leave it as an open problem for now.

The inspiration for our analysis method stems from the observation that the algorithm spends a relatively long time on the plateau. So regardless of the initial distribution on the plateau, the distribution of the individual converges to the conditional stationary distribution long before the algorithm leaves the plateau. This indicates that our method is less suitable to analyze how evolutionary algorithms leave plateaus which are easy to leave, but such plateaus usually present not bigger problems in optimization.

Overall, we are optimistic that our main analysis method, switching between the level chain and the individual chain, which might be the first attempt to devise a general analysis method for EAs on plateaus, will find further applications.

While our analysis method can deal with a large class of  $(1 + 1)$ -type hill-climbers, it is currently less clear how to analyze population-based algorithms. A series of works [HY01, HY04, JJW05, Wit06, JS07, CHS<sup>+</sup>09, RS14, DK15, ADFH18, ADY19, DDE15, DD18] analyzing the runtime of various versions of the  $(\mu + \lambda)$  EA,  $(\mu, \lambda)$  EA, and  $(1 + (\lambda, \lambda))$  GA on ONEMAX show that these algorithms quickly reach the plateau of the  $\text{PLATEAU}$  function,

but it is currently not clear how to extend our method to get sharp runtime estimates also for the part of the process on the plateau. Likewise, it is not clear how our methods can be extended to algorithms that dynamically change their parameters [DD20], because here in most cases the relevant state of the algorithm not only consists of the current search point(s). For hyper-heuristics, we could show two elementary results, but again, as discussed in Section 5.4, for most hyper-heuristics our general result cannot be applied. By analogy with jump functions, crossover-based algorithms should be efficient on plateaus, especially the ones using different diversity mechanisms [DFK<sup>+</sup>18]. However, these algorithms are more complicated, and even on jump functions there are no asymptotically tight bounds on their runtime. This suggests that studying their behavior on plateaus might be even more complicated. Given that plateaus of constant fitness appear frequently in optimization problems, we feel that the open questions discussed in this paragraph are worth pursuing in the near future.

## Acknowledgements

This work was financially supported by the National Center for Cognitive Research of ITMO University and by a public grant as part of the Investissements d’avenir project, reference ANR-11-LABX-0056-LMH, LabEx LMH.

## A Appendix

Since both a reviewer of our submission to the GECCO 2018 theory track and (after rejection of the former) a reviewer of PPSN 2018 claimed that our result is already wrong for the small case  $k = 2$  and a specific mutation operator, to clarify the situation and to avoid similar problems in future reviewing processes, we analyze now the case  $k = 2$  in full generality by elementary means. This proves the reviewers’ claims wrong and shows that the case  $k = 2$  can be solved by regarding a simple system of equations, whose unique solution agrees with our main results.

We start by describing the reviewers’ incorrect concerns. The reviewer of our submission to the GECCO 2018 theory track wrongfully considers the following a counterexample. Suppose  $k = 2$  and suppose the mutation operator flips either 1 or 2 randomly chosen bits, each with probability  $\frac{1}{2}$ . Then our theorem gives an expected runtime of  $E[T] = n^2/2(1 + o(1))$ , but the reviewer claims that the expected runtime is  $n^2(1 + o(1))$ , referring to own calculations not provided.

The reviewer of PPSN 2018 suggested a more general counterexample. She or he considers again  $k = 2$  and the mutation operator that flips one randomly chosen bit with probability  $p_1$  (where  $p_1 = \Omega(1)$ ) and it flips two randomly chosen bits with probability  $p_2$ . The reviewer claims, again without giving details, that the expected runtime of the described algorithm on the  $\text{PLATEAU}_2$  function is  $\frac{n^2}{2p_1+p_2}(1 + o(1))$ , while our Theorem 19 gives an expected runtime of  $E[T] = \frac{n^2}{2(p_1+p_2)}(1 + o(1))$ .

To cover both examples and possible future ones, we consider the case  $k = 2$  for a general unbiased mutation operator (with probability to flip exactly one bit of  $\Omega(1)$ ). For brevity and to match the notation of the latter reviewer we define  $p_i$  as the probability that the mutation operator flips exactly  $i$  bits (that was denoted as  $\Pr[\alpha = i]$  in the main part of the paper). We recall from the body of the paper that there is a one-to-one correspondence between unary unbiased mutation operators and vectors  $p = (p_0, p_1, \dots, p_n) \in [0, 1]^{n+1}$  with  $\|p\|_1 = 1$ .

**Lemma 26.** *If  $p_1 = \Omega(1)$ , then the runtime of the  $(1 + 1)$  EA with an arbitrary unbiased mutation operator as described above optimizing the  $n$ -dimensional  $\text{PLATEAU}_2$  function is  $\frac{n^2}{2(p_1+p_2)}(1 + o(1))$ .*

*Proof.* To find the expected runtime of the algorithm on the plateau we consider the level chain. It contains two states (for level 0 and level 1), but additionally to find the expected runtime we now include the optimum into the chain as a new state called level 2.

By  $p_i^j$  we denote the transition probabilities between levels for  $i \in [0..1]$  and  $j \in [0..2]$ . We define  $T_i$  as the runtime of the algorithm if it starts on level  $i$  for  $i \in [0..1]$ . The following system of equations follows from elementary Markov chain theory.

$$\begin{aligned} E[T_0] &= 1 + p_0^0 E[T_0] + p_0^1 E[T_1]. \\ E[T_1] &= 1 + p_1^0 E[T_0] + p_1^1 E[T_1]. \end{aligned}$$

By elementary transformations we obtain the following equivalent system.

$$\begin{aligned} E[T_0] &= \frac{p_1^0 + p_0^1 + p_1^2}{p_1^2 p_0^1 + p_1^0 p_0^2 + p_1^2 p_0^2}. \\ E[T_1] &= \frac{p_1^0 + p_0^1 + p_0^2}{p_1^2 p_0^1 + p_1^0 p_0^2 + p_1^2 p_0^2}. \end{aligned} \tag{20}$$

To evaluate the right-hand sides, we first compute all the transition probabilities.

- $p_0^1$  is the probability to either flip one zero-bit or to flip both zero-bits and one one-bit, that is,

$$p_0^1 = p_1 \frac{2}{n} + p_3 \frac{6}{n(n-1)} = p_1 \frac{2}{n} + o(1/n).$$

Recall that  $p_1$  is considered as some positive constant.

- $p_0^2$  is the probability to flip both zero-bits, that is,

$$p_0^2 = p_2 \frac{2}{n(n-1)}.$$

- $p_1^0$  is the probability to either flip one one-bit or to flip two one-bits and the only zero-bit, that is,

$$p_1^0 = p_1 \frac{n-1}{n} + p_3 \frac{3}{n} = p_1 + o(1).$$

- $p_1^2$  is the probability to flip the only zero-bit, that is,

$$p_1^2 = p_1 \frac{1}{n}.$$

- In other cases the mutation operator generates either an individual with the same number of one-bits or an individual from outside the plateau, so the algorithm does not accept it. Therefore,  $p_0^0 = 1 - p_0^1 - p_0^2$  and  $p_1^1 = 1 - p_1^0 - p_1^2$ .

We compute the numerators and denominator in the right-hand sides of (20).

- $p_1^0 + p_0^1 + p_1^2 = p_1 + o(1) + p_1 \frac{2}{n} + o(1/n) + p_1 \frac{1}{n} = p_1 + o(1).$
- $p_1^0 + p_0^1 + p_0^2 = p_1 + o(1) + p_1 \frac{2}{n} + o(1/n) + p_2 \frac{2}{n(n-1)} = p_1 + o(1).$
- $p_1^2 p_0^1 + p_1^0 p_0^2 + p_1^2 p_0^2 = (p_1)^2 \frac{2}{n^2} + p_1 p_2 \frac{2}{n^2} + o(1/n^2).$

This gives the desired values for the expected runtimes.

$$E[T_0] = \frac{p_1 + o(1)}{p_1(p_1 + p_2) \frac{2}{n^2} + o(1/n^2)} = \frac{n^2}{2(p_1 + p_2)}(1 + o(1)),$$

$$E[T_1] = \frac{p_1 + o(1)}{p_1(p_1 + p_2)\frac{2}{n^2} + o(1/n^2)} = \frac{n^2}{2(p_1 + p_2)}(1 + o(1)).$$

So independently on the starting state we have precisely the same expected runtime (apart from the lower order terms ignored in both cases) as obtained through Theorem 19.

□

## References

- [AD18] Denis Antipov and Benjamin Doerr. Precise runtime analysis for plateaus. In *Parallel Problem Solving from Nature, PPSN XV, Part II*, pages 117–128. Springer, 2018.
- [ADFH18] Denis Antipov, Benjamin Doerr, Jiefeng Fang, and Tangi Hetet. Runtime analysis for the  $(\mu + \lambda)$  EA optimizing OneMax. In *Genetic and Evolutionary Computation Conference, GECCO 2018*, pages 1459–1466. ACM, 2018.
- [ADY19] Denis Antipov, Benjamin Doerr, and Quentin Yang. The efficiency threshold for the offspring population size of the  $(\mu, \lambda)$  EA. In *Genetic and Evolutionary Computation Conference, GECCO 2019*, pages 1461–1469. ACM, 2019.
- [AL14] Fawaz Alanazi and Per Kristian Lehre. Runtime analysis of selection hyper-heuristics with classical learning mechanisms. In *Congress on Evolutionary Computation, CEC 2014*, pages 2515–2523. IEEE, 2014.
- [BDK16] Maxim Buzdalov, Benjamin Doerr, and Mikhail Kever. The unrestricted black-box complexity of jump functions. *Evolutionary Computation*, 24:719–744, 2016.
- [BDN10] Süntje Böttcher, Benjamin Doerr, and Frank Neumann. Optimal fixed and adaptive mutation rates for the LeadingOnes problem. In *Parallel Problem Solving from Nature, PPSN XI*, pages 1–10. Springer, 2010.
- [BFH<sup>+</sup>09] Dima Brockhoff, Tobias Friedrich, Nils Hebbinghaus, Christian Klein, Frank Neumann, and Eckart Zitzler. On the effects of adding objectives to plateau functions. *IEEE Transactions on Evolutionary Computation*, 13:591–603, 2009.

- [CDEL18] Dogan Corus, Duc-Cuong Dang, Anton V. Eremeev, and Per Kristian Lehre. Level-based analysis of genetic algorithms and other search processes. *IEEE Transactions on Evolutionary Computation*, 22:707–719, 2018.
- [CHS<sup>+</sup>09] Tianshi Chen, Jun He, Guangzhong Sun, Guoliang Chen, and Xin Yao. A new approach for analyzing average time complexity of population-based evolutionary algorithms on unimodal problems. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 39:1092–1106, 2009.
- [COY17] Dogan Corus, Pietro S. Oliveto, and Donya Yazdani. On the runtime analysis of the opt-IA artificial immune system. In *Genetic and Evolutionary Computation Conference, GECCO 2017*, pages 83–90, 2017.
- [COY18] Dogan Corus, Pietro S. Oliveto, and Donya Yazdani. Artificial immune systems can find arbitrarily good approximations for the NP-hard partition problem. In *Parallel Problem Solving from Nature, PPSN XV, Part II*, pages 16–28. Springer, 2018.
- [DD18] Benjamin Doerr and Carola Doerr. Optimal static and self-adjusting parameter choices for the  $(1+(\lambda, \lambda))$  genetic algorithm. *Algorithmica*, 80:1658–1709, 2018.
- [DD20] Benjamin Doerr and Carola Doerr. Theory of parameter control for discrete black-box optimization: Provable performance gains through dynamic parameter choices. In Benjamin Doerr and Frank Neumann, editors, *Theory of Evolutionary Computation: Recent Developments in Discrete Optimization*, pages 271–321. Springer, 2020. Also available at <https://arxiv.org/abs/1804.05650>.
- [DDE15] Benjamin Doerr, Carola Doerr, and Franziska Ebel. From black-box complexity to designing new genetic algorithms. *Theoretical Computer Science*, 567:87–104, 2015.
- [DDK14] Benjamin Doerr, Carola Doerr, and Timo Kötzing. The unbiased black-box complexity of partition is polynomial. *Artificial Intelligence*, 216:275–286, 2014.
- [DDK15] Benjamin Doerr, Carola Doerr, and Timo Kötzing. Unbiased black-box complexities of jump functions. *Evolutionary Computation*, 23:641–670, 2015.



- [DDL19] Benjamin Doerr, Carola Doerr, and Johannes Lengler. Self-adjusting mutation rates with provably optimal success rules. In *Genetic and Evolutionary Computation Conference, GECCO 2019*, pages 1479–1487. ACM, 2019.
- [DDY20] Benjamin Doerr, Carola Doerr, and Jing Yang. Optimal parameter choices via precise black-box analysis. *Theoretical Computer Science*, 801:1–34, 2020.
- [DFK<sup>+</sup>16] Duc-Cuong Dang, Tobias Friedrich, Timo Kötzing, Martin S. Krejca, Per Kristian Lehre, Pietro S. Oliveto, Dirk Sudholt, and Andrew M. Sutton. Emergence of diversity and its benefits for crossover in genetic algorithms. In *Parallel Problem Solving from Nature, PPSN XIV*, pages 890–900. Springer, 2016.
- [DFK<sup>+</sup>18] Duc-Cuong Dang, Tobias Friedrich, Timo Kötzing, Martin S. Krejca, Per Kristian Lehre, Pietro S. Oliveto, Dirk Sudholt, and Andrew M. Sutton. Escaping local optima using crossover with emergent diversity. *IEEE Transactions on Evolutionary Computation*, 22:484–497, 2018.
- [DFW11] Benjamin Doerr, Mahmoud Fouz, and Carsten Witt. Sharp bounds by probability-generating functions and variable drift. In *Genetic and Evolutionary Computation Conference, GECCO 2011*, pages 2083–2090. ACM, 2011.
- [DHN07] Benjamin Doerr, Nils Hebbinghaus, and Frank Neumann. Speeding up evolutionary algorithms through asymmetric mutation operators. *Evolutionary Computation*, 15:401–410, 2007.
- [DJK08] Benjamin Doerr, Tomas Jansen, and Christian Klein. Comparing global and local mutations on bit strings. In *Genetic and Evolutionary Computation Conference, GECCO 2008*, pages 929–936. ACM, 2008.
- [DJS<sup>+</sup>13] Benjamin Doerr, Thomas Jansen, Dirk Sudholt, Carola Winzen, and Christine Zarges. Mutation rate matters even when optimizing monotonic functions. *Evolutionary Computation*, 21:1–27, 2013.
- [DJW02] Stefan Droste, Thomas Jansen, and Ingo Wegener. On the analysis of the (1+1) evolutionary algorithm. *Theoretical Computer Science*, 276:51–81, 2002.

- [DJW12] Benjamin Doerr, Daniel Johannsen, and Carola Winzen. Multiplicative drift analysis. *Algorithmica*, 64:673–697, 2012.
- [DK15] Benjamin Doerr and Marvin Künnemann. Optimizing linear functions with the  $(1 + \lambda)$  evolutionary algorithm—different asymptotic runtimes for different instances. *Theoretical Computer Science*, 561:3–23, 2015.
- [DK19] Benjamin Doerr and Timo Kötzing. Multiplicative up-drift. In *Genetic and Evolutionary Computation Conference, GECCO 2019*, pages 1470–1478. ACM, 2019.
- [DKLW13] Benjamin Doerr, Timo Kötzing, Johannes Lengler, and Carola Winzen. Black-box complexities of combinatorial problems. *Theoretical Computer Science*, 471:84–106, 2013.
- [DL16] Duc-Cuong Dang and Per Kristian Lehre. Runtime analysis of non-elitist populations: from classical optimisation to partial information. *Algorithmica*, 75:428–461, 2016.
- [DLMN17] Benjamin Doerr, Huu Phuoc Le, Régis Makhlara, and Ta Duy Nguyen. Fast genetic algorithms. In *Genetic and Evolutionary Computation Conference, GECCO 2017*, pages 777–784. ACM, 2017.
- [DLOW18] Benjamin Doerr, Andrei Lissovoi, Pietro S. Oliveto, and John A. Warwicker. On the runtime analysis of selection hyper-heuristics with adaptive learning periods. In *Genetic and Evolutionary Computation Conference Companion, GECCO 2018*, pages 1015–1022. ACM, 2018.
- [Doe19a] Benjamin Doerr. An exponential lower bound for the runtime of the compact genetic algorithm on jump functions. In *Foundations of Genetic Algorithms, FOGA 2019*, pages 25–33. ACM, 2019.
- [Doe19b] Benjamin Doerr. A tight runtime analysis for the cGA on jump functions: EDAs can cross fitness valleys at no extra cost. In *Genetic and Evolutionary Computation Conference, GECCO 2019*, pages 1488–1496. ACM, 2019.
- [FHN09] Tobias Friedrich, Nils Hebbinghaus, and Frank Neumann. Comparison of simple diversity mechanisms on plateau functions. *Theoretical Computer Science*, 410:2455–2462, 2009.

- [FHN10] Tobias Friedrich, Nils Hebbinghaus, and Frank Neumann. Plateaus can be harder in multi-objective optimization. *Theoretical Computer Science*, 411:854–864, 2010.
- [FKK<sup>+</sup>16] Tobias Friedrich, Timo Kötzing, Martin S. Krejca, Samadhi Nallaperuma, Frank Neumann, and Martin Schirneck. Fast building block assembly by majority vote crossover. In *Genetic and Evolutionary Computation Conference, GECCO 2016*, pages 661–668. ACM, 2016.
- [GKS99] Josselin Garnier, Leila Kallel, and Marc Schoenauer. Rigorous hitting times for binary mutations. *Evolutionary Computation*, 7:173–203, 1999.
- [GW03] Oliver Giel and Ingo Wegener. Evolutionary algorithms and the maximum matching problem. In *Symposium on Theoretical Aspects of Computer Science, STACS 2003*, pages 415–426. Springer, 2003.
- [HPR<sup>+</sup>18] Hsien-Kuei Hwang, Alois Panholzer, Nicolas Rolin, Tsung-Hsi Tsai, and Wei-Mei Chen. Probabilistic analysis of the (1+1)-evolutionary algorithm. *Evolutionary Computation*, 26:299–345, 2018.
- [HS18] Václav Hasenöhrl and Andrew M. Sutton. On the runtime dynamics of the compact genetic algorithm on jump functions. In *Genetic and Evolutionary Computation Conference, GECCO 2018*, pages 967–974. ACM, 2018.
- [HW19] Hsien-Kuei Hwang and Carsten Witt. Sharp bounds on the runtime of the (1+1) EA via drift analysis and analytic combinatorial tools. In *Foundations of Genetic Algorithms, FOGA 2019*, pages 1–12. ACM, 2019.
- [HY01] Jun He and Xin Yao. Drift analysis and average time complexity of evolutionary algorithms. *Artificial Intelligence*, 127:51–81, 2001.
- [HY04] Jun He and Xin Yao. A study of drift analysis for estimating computation time of evolutionary algorithms. *Natural Computing*, 3:21–35, 2004.

- [JJW05] Thomas Jansen, Kenneth A. De Jong, and Ingo Wegener. On the choice of the offspring population size in evolutionary algorithms. *Evolutionary Computation*, 13:413–440, 2005.
- [JS07] Jens Jägersküpper and Tobias Storch. When the plus strategy outperforms the comma strategy and when not. In *Foundations of Computational Intelligence, FOCI 2007*, pages 25–32. IEEE, 2007.
- [JW01] Thomas Jansen and Ingo Wegener. Evolutionary algorithms - how to cope with plateaus of constant fitness and when to reject strings of the same fitness. *IEEE Transactions on Evolutionary Computation*, 5:589–599, 2001.
- [JW02] Thomas Jansen and Ingo Wegener. The analysis of evolutionary algorithms—a proof that crossover really can help. *Algorithmica*, 34:47–66, 2002.
- [LOW17] Andrei Lissovoi, Pietro S. Oliveto, and John A. Warwicker. On the runtime analysis of generalised selection hyper-heuristics for pseudo-Boolean optimisation. In *Genetic and Evolutionary Computation Conference, GECCO 2017*, pages 849–856. ACM, 2017.
- [LW12] Per Kristian Lehre and Carsten Witt. Black-box search by unbiased variation. *Algorithmica*, 64:623–642, 2012.
- [LW14] Per Kristian Lehre and Carsten Witt. Concentrated hitting times of randomized search heuristics with variable drift. In *International Symposium on Algorithms and Computation, ISAAC 2014*, pages 686–697. Springer, 2014.
- [MB17] Vladimir Mironovich and Maxim Buzdalov. Evaluation of heavy-tailed mutation operator on maximum flow test generation problem. In *Genetic and Evolutionary Computation Conference Companion, GECCO 2017*, pages 1423–1426. ACM, 2017.
- [Mey00] Carl D. Meyer, editor. *Matrix Analysis and Applied Linear Algebra*. Society for Industrial and Applied Mathematics, 2000.
- [Müh93] Heinz Mühlenbein. Evolutionary algorithms: Theory and applications. In *Local Search in Combinatorial Optimization*. Wiley, 1993.

- [NSW09] Frank Neumann, Dirk Sudholt, and Carsten Witt. Analysis of different MMAS ACO algorithms on unimodal functions and plateaus. *Swarm Intelligence*, 3:35–68, 2009.
- [NW07] Frank Neumann and Ingo Wegener. Randomized local search, evolutionary algorithms, and the minimum spanning tree problem. *Theoretical Computer Science*, 378(1):32–40, 2007.
- [RR04] Michael Renardy and Robert C. Rogers. *An Introduction to Partial Differential Equations*. Texts in Applied Mathematics. Springer, 2004.
- [RS14] Jonathan E. Rowe and Dirk Sudholt. The choice of the offspring population size in the  $(1, \lambda)$  evolutionary algorithm. *Theoretical Computer Science*, 545:20–38, 2014.
- [Rud96] Günter Rudolph. Convergence of evolutionary algorithms in general search spaces. In *International Conference on Evolutionary Computation*, pages 50–54. IEEE, 1996.
- [Sud13] Dirk Sudholt. A new method for lower bounds on the running time of evolutionary algorithms. *IEEE Transactions on Evolutionary Computation*, 17:418–435, 2013.
- [Sud20] Dirk Sudholt. Analysing the robustness of evolutionary algorithms to noise: Refined runtime bounds and an example where noise is beneficial. *Algorithmica*, 2020. To appear.
- [Suz95] Joe Suzuki. A markov chain analysis on simple genetic algorithms. *IEEE Transactions on Systems, Man, and Cybernetics*, 25:655–659, 1995.
- [Vit00] Paul Vitányi. A discipline of evolutionary programming. *Theoretical Computer Science*, 241:3–23, 2000.
- [Weg01] Ingo Wegener. Theoretical aspects of evolutionary algorithms. In *Automata, Languages and Programming, ICALP 2001*, pages 64–78. Springer, 2001.
- [Wit06] Carsten Witt. Runtime analysis of the  $(\mu + 1)$  EA on simple pseudo-Boolean functions. *Evolutionary Computation*, 14:65–86, 2006.

- [Wit13] Carsten Witt. Tight bounds on the optimization time of a randomized search heuristic on linear functions. *Combinatorics, Probability and Computing*, 22:294–318, 2013.
- [WVHM18] Darrell Whitley, Swetha Varadarajan, Rachel Hirsch, and Anirban Mukhopadhyay. Exploration and exploitation without mutation: solving the jump function in  $\Theta(n)$  time. In *Parallel Problem Solving from Nature, PPSN XV, Part II*, pages 55–66. Springer, 2018.