

# An Exponential Lower Bound for the Runtime of the Compact Genetic Algorithm on Jump Functions

Benjamin Doerr

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

## ABSTRACT

In the first runtime analysis of an estimation-of-distribution algorithm (EDA) on the multimodal jump function class, Hasenöhl and Sutton (GECCO 2018) proved that the runtime of the compact genetic algorithm with suitable parameter choice on jump functions with high probability is at most polynomial (in the dimension) if the jump size is at most logarithmic (in the dimension), and is at most exponential in the jump size if the jump size is super-logarithmic. The exponential runtime guarantee was achieved with a hypothetical population size that is also exponential in the jump size. Consequently, this setting cannot lead to a better runtime.

In this work, we show that any choice of the hypothetical population size leads to a runtime that, with high probability, is at least exponential in the jump size. This result might be the first non-trivial exponential lower bound for EDAs that holds for arbitrary parameter settings.

## CCS CONCEPTS

• Theory of computation → Theory of randomized search heuristics.

## KEYWORDS

evolutionary algorithms, runtime analysis

### ACM Reference Format:

Benjamin Doerr. 2019. An Exponential Lower Bound for the Runtime of the Compact Genetic Algorithm on Jump Functions. In *Foundations of Genetic Algorithms XV (FOGA '19)*, August 27–29, 2019, Potsdam, Germany. ACM, New York, NY, USA, 9 pages. <https://doi.org/10.1145/3299904.3340304>

## 1 INTRODUCTION

Due to the inherent highly complex stochastic processes, the mathematical analysis of estimation-of-distribution algorithms (EDAs) is still in its early childhood. Whereas for classic evolutionary algorithms many deep results exist, see, e.g., [4, 25, 34], for EDAs even some of the most basic problems are not fully understood, such as the runtime of the compact genetic algorithm (cGA) on the ONEMAX benchmark function [16, 32, 37]. We direct the reader to the recent survey [28] for a complete picture of the state of the art in mathematical analyses of EDAs.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from [permissions@acm.org](mailto:permissions@acm.org).

FOGA '19, August 27–29, 2019, Potsdam, Germany

© 2019 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM ISBN 978-1-4503-6254-2/19/08...\$15.00

<https://doi.org/10.1145/3299904.3340304>

Given this state of the art, it is not surprising that the first runtime analysis of an EDA on a multimodal objective function appeared only very recently. In their GECCO 2018 paper, Hasenöhl and Sutton [23] analyze the optimization time of the cGA on the jump function class. Jump functions are simple unimodal functions except that they have a valley of low fitness of scalable size  $k$  around the global optimum. Hasenöhl and Sutton show [23, Theorem 3.3] that, for a sufficiently large constant  $C$  and any constant  $\varepsilon > 0$ , the cGA with hypothetical population size at least  $\mu \geq \max\{Cne^{4k}, n^{3.5+\varepsilon}\}$  with probability  $1 - o(1)$  finds the optimum of any jump function with jump size at most  $k = o(n)$  in  $O(\mu n^{1.5} \log n + e^{4k})$  generations (which is also the number of fitness evaluations, since the cGA evaluates two search points in each iteration).

We note that this runtime guarantee is polynomial (in the problem dimension  $n$ ) when  $k = O(\log n)$  and exponential (in  $k$ ) otherwise. Both parts of the result are remarkable when recalling that many classical evolutionary algorithms need time  $\Omega(n^k)$ .

For the polynomial part, the upper bound of order  $\mu n^{1.5} \log n$ , which is  $n^{5+\varepsilon}$  when choosing  $\mu$  optimally, was for  $k < \frac{1}{20} \ln n$  recently [12] improved to  $O(\mu \sqrt{n})$  for all  $\mu = \Omega(\sqrt{n} \log n) \cap \text{poly}(n)$ , which is  $O(n \log n)$  for the optimal choice  $\mu = \Theta(\sqrt{n} \log n)$ . Note that  $n \log n$  is the asymptotic runtime of many evolutionary algorithms, including the cGA with good parameter choices [16, 32, 37], on the simple unimodal ONEMAX problem. Hence this result shows that the cGA does not suffer significantly from the valley of low fitness around the optimum which is characteristic for jump functions (as long as this valley is not too wide, that is,  $k < \frac{1}{20} \ln n$ ). If we are willing to believe that  $\Omega(n \log n)$  is also a lower bound for the runtime of the cGA on these jump functions (which given the results for ONEMAX appears very plausible, but which seems hard to prove, see Section 4), then the result in [12] determines the precise asymptotic runtime of the cGA with optimal parameter choice for  $k < \frac{1}{20} \ln n$ .

What is left open by these two previous works is how good the exponential upper bound (for  $k$  super-logarithmic in  $n$ ) is. Since Hasenöhl and Sutton prove their exponential runtime guarantee only for a hypothetical population size  $\mu = \Omega(ne^{4k})$ , it is clear that they, in this setting, cannot have a sub-exponential runtime (for the sake of completeness, we shall make this elementary argument precise in Lemma 1). So the question remains if, by choosing a smaller hypothetical population size, one could have obtained a better runtime guarantee.

**Our main result** is a negative answer to this question. In Theorem 6 we show that, regardless of the hypothetical population size, the runtime of the cGA on a jump function with jump size  $k$  is at least exponential in  $k$  with high probability. Interestingly, not only our result is a uniform lower bound independent of the hypothetical population size, but our proof is also “uniform” in the sense that it

needs case distinctions neither w.r.t. the hypothetical population size nor w.r.t. the different reasons for the lower bound. Here we recall that the existing runtime analyses, see, e.g., again [16, 32, 37] find two reasons why an EDA can be inefficient. (i) The hypothetical population size is large and consequently it takes long to move the frequencies into the direction of the optimum. (ii) The hypothetical population size is small and thus, in the absence of a strong fitness signal, the random walk of the frequencies brings some frequencies close to the boundaries of the frequency spectrum; from there they are hard to move back into the game.

We avoid such potentially tedious case distinctions via an elegant drift argument on the sum of the frequencies. Ignoring some technicalities here, we show that, regardless of the hypothetical population size, the frequency sum overshoots a value of  $n - \frac{1}{4}k$  only after an expected number of  $2^{\Omega(k)}$  iterations. However, in an iteration where the frequency sum is below  $n - \frac{1}{4}k$ , the optimum is sampled only with probability  $2^{-\Omega(k)}$ . These two results give our main result.

As a **side result**, we show in Section 4 that a result like “ONEMAX is an easiest function with a unique global optimum for the cGA”, if true at all, cannot be proven along the same lines as the corresponding results for many mutation-based algorithms. This in particular explains why we and the previous works on this topic have not shown an  $\Omega(n \log n)$  lower bound for the runtime of the cGA on jump functions.

## 2 PRELIMINARIES

### 2.1 The Compact Genetic Algorithm<sup>1</sup>

The *compact genetic algorithm* (cGA) is an estimation-of-distribution algorithm (EDA) proposed by Harik, Lobo, and Goldberg [22] for the maximization of pseudo-Boolean functions  $\mathcal{F} : \{0, 1\}^n \rightarrow \mathbb{R}$ . Being a univariate EDA, it develops a probabilistic model described by a *frequency vector*  $f \in [0, 1]^n$ . This frequency vector describes a probability distribution on the search space  $\{0, 1\}^n$ . If  $X = (X_1, \dots, X_n) \in \{0, 1\}^n$  is a search point sampled according to this distribution—we write

$$X \sim \text{Sample}(f)$$

to indicate this—then we have  $\Pr[X_i = 1] = f_i$  independently for all  $i \in [1..n] := \{1, \dots, n\}$ . In other words, the probability that  $X$  equals some fixed search point  $y$  is

$$\Pr[X = y] = \prod_{i: y_i=1} f_i \prod_{i: y_i=0} (1 - f_i).$$

In each iteration, the cGA updates this probabilistic model as follows. It samples two search points  $x^1, x^2 \sim \text{Sample}(f)$ , computes the fitness of both, and defines  $(y^1, y^2) = (x^1, x^2)$  when  $x^1$  is at least as fit as  $x^2$  and  $(y^1, y^2) = (x^2, x^1)$  otherwise. Consequently,  $y^1$  is the better search point of the two (if not both have the same fitness). We then define a preliminary model by  $f' := f + \frac{1}{\mu}(y^1 - y^2)$ . This definition ensures that, when  $y^1$  and  $y^2$  differ in some bit position  $i$ , the  $i$ -th preliminary frequency moves by a step of  $\frac{1}{\mu}$  into the direction of  $y_i^1$ , which we hope to be the right direction since  $y^1$

is the better of the two search points. The *hypothetical population size*  $\mu$  is used to control how strong this update is.

To avoid a premature convergence, we ensure that the new frequency vector is in  $[\frac{1}{n}, 1 - \frac{1}{n}]^n$  by capping too small or too large values at the corresponding boundaries. More precisely, for all  $\ell \leq u$  and all  $r \in \mathbb{R}$  we define

$$\text{minmax}(\ell, r, u) := \max\{\ell, \min\{r, u\}\} = \begin{cases} \ell & \text{if } r < \ell \\ r & \text{if } r \in [\ell, u] \\ u & \text{if } r > u \end{cases}$$

and we lift this notation to vectors by reading it component-wise. Now the new frequency vector is  $\text{minmax}(\frac{1}{n}\mathbf{1}_n, f', (1 - \frac{1}{n})\mathbf{1}_n)$ .

This iterative frequency development is pursued until some termination criterion is met. Since we aim at analyzing the time (number of iterations) it takes to sample the optimal solution (this is what we call the *runtime* of the cGA), we do not specify a termination criterion and pretend that the algorithm runs forever.

The pseudo-code for the cGA is given in Algorithm 1. We shall use the notation given there frequently in our proofs. For the frequency vector  $f_t$  obtained at the end of iteration  $t$ , we denote its  $i$ -th component by  $f_{i,t}$  or, when there is no risk of ambiguity, by  $f_{it}$ . We shall frequently argue with the sum of the frequencies, which can be written as  $\|f_t\|_1 := \sum_{i=1}^n |f_{it}|$  since the frequencies are non-negative. With a slight **abuse of notation**, we extend this common notation also to preliminary frequency vectors  $f'$  and thus write  $\|f'\|_1 := \sum_{i=1}^n f'_{it}$ , when there is not danger of confusion.

---

**Algorithm 1:** The compact genetic algorithm (cGA) to maximize a function  $\mathcal{F} : \{0, 1\}^n \rightarrow \mathbb{R}$ .

---

```

1  $t \leftarrow 0$ ;
2  $f_t = (\frac{1}{2}, \dots, \frac{1}{2}) \in [0, 1]^n$ ;
3 repeat
4    $x^1 \leftarrow \text{Sample}(f_t)$ ;
5    $x^2 \leftarrow \text{Sample}(f_t)$ ;
6   if  $\mathcal{F}(x^1) \geq \mathcal{F}(x^2)$  then  $(y^1, y^2) \leftarrow (x^1, x^2)$  else
7      $(y^1, y^2) \leftarrow (x^2, x^1)$ ;
8    $f'_{t+1} \leftarrow f_t + \frac{1}{\mu}(y^1 - y^2)$ ;
9    $f_{t+1} \leftarrow \text{minmax}(\frac{1}{n}\mathbf{1}_n, f'_{t+1}, (1 - \frac{1}{n})\mathbf{1}_n)$ ;
10   $t \leftarrow t + 1$ ;
11 until forever;
```

---

**Well-behaved frequency assumption:** For the hypothetical population size  $\mu$ , we take the common assumption that any two frequencies that can occur in a run of the cGA differ by a multiple of  $\frac{1}{\mu}$ . We call this the *well-behaved frequency assumption*. This assumption was implicitly already made in [22] by using even  $\mu$  in all experiments (note that the hypothetical population size is denoted by  $n$  in [22]). This assumption was made explicit in [16] by requiring  $\mu$  to be even. Both works do not use the frequencies boundaries  $\frac{1}{n}$  and  $1 - \frac{1}{n}$ , so an even value for  $\mu$  ensures well-behaved frequencies.

For the case with frequency boundaries, the well-behaved frequency assumption is equivalent to  $(1 - \frac{2}{n})$  being an even multiple of the update step size  $\frac{1}{\mu}$ . In this case,  $n_\mu = (1 - \frac{2}{n})\mu \in 2\mathbb{N}$  and the

<sup>1</sup>This and the subsequent subsection are (essentially) identical to the related sections in [12].

set of frequencies that can occur is

$$F := F_\mu := \left\{ \frac{1}{n} + \frac{i}{\mu} \mid i \in [0..n_\mu] \right\}.$$

This assumption was made, e.g., in the papers [20] (see the last paragraph of Section II.C) and [32] (see the paragraph following Lemma 2.1) as well as in the proof of Theorem 2 in [37].

**A trivial lower bound:** We finish this subsection on the cGA with the following very elementary remark, which shows that the cGA with hypothetical population size  $\mu$  with probability  $1 - \exp(-\Omega(n))$  has a runtime of at least  $\min\{\mu/4, \exp(\Theta(n))\}$  on any  $\mathcal{F} : \{0, 1\}^n \rightarrow \mathbb{R}$  with a unique global optimum. This shows, in particular, that the cGA with the parameter value  $\mu = \exp(\Omega(k))$  used to optimize jump functions with gap size  $k \in \omega(\log n) \cap o(n)$  in time  $\exp(O(k))$  in [23] cannot have a runtime better than exponential in  $k$ .

**LEMMA 1.** *Let  $\mathcal{F} : \{0, 1\}^n \rightarrow \mathbb{R}$  have a unique global optimum. The probability that the cGA generates the optimum of  $f$  in  $T = \min\{\mu/4, (1.3)^n\}$  iterations is at most  $\exp(-\Omega(n))$ .*

**PROOF.** By the definition of the cGA, the frequency vector  $f$  used in iteration  $t = 1, 2, 3, \dots$  satisfies  $f \in [\frac{1}{2} - \frac{t-1}{\mu}, \frac{1}{2} + \frac{t-1}{\mu}]^n$ . Consequently, the probability that a fixed one of the two search points which are generated in this iteration is the optimum, is at most  $(\frac{1}{2} + \frac{t-1}{\mu})^n$ . For  $t \leq \mu/4$ , this is at most  $(3/4)^n$ . Hence by a simple union bound, the probability that the optimum is generated in the first  $T = \min\{\mu/4, (1.3)^n\}$  iterations, is at most  $2T(3/4)^n = \exp(-\Omega(n))$ .  $\square$

## 2.2 Related Work

In all results described in this section, we shall assume that the hypothetical population size is at most polynomial in the problem size  $n$ , that is, that there is a constant  $c$  such that  $\mu \leq n^c$ .

The first to conduct a rigorous runtime analysis for the cGA was Droste in his seminal work [16]. He regarded the cGA without frequency boundaries, that is, he just took  $f_{t+1} := f'_{t+1}$  in our notation. He showed that this algorithm with  $\mu \geq n^{1/2+\varepsilon}$ ,  $\varepsilon > 0$  any positive constant, finds the optimum of the ONEMAX function defined by

$$\text{ONEMAX}(x) = \|x\|_1 = \sum_{i=1}^n x_i$$

for all  $x \in \{0, 1\}^n$  with probability at least  $1/2$  in  $O(\mu\sqrt{n})$  iterations [16, Theorem 8].

Droste also showed that this cGA for any objective function  $\mathcal{F}$  with unique optimum has an expected runtime of  $\Omega(\mu\sqrt{n})$  when conditioning on no premature convergence [16, Theorem 6]. It is easy to see that his proof of the lower bound can be extended to the cGA with frequency boundaries, that is, to Algorithm 1. For this, it suffices to deduce from his drift argument the result that the first time  $T_{n/4}$  that the frequency distance  $D = \sum_{i=1}^n (1 - f_{it})$  is less than  $n/4$  satisfies  $E[T_{n/4}] \geq \mu\sqrt{n}\frac{\sqrt{2}}{4}$ . Since the probability to sample the optimum from a frequency distance of at least  $n/4$  is at

most

$$\begin{aligned} \prod_{i=1}^n f_{it} &= \prod_{i=1}^n (1 - (1 - f_{it})) \leq \prod_{i=1}^n \exp(-(1 - f_{it})) \\ &= \exp\left(-\sum_{i=1}^n (1 - f_{it})\right) \leq \exp(-n/4), \end{aligned}$$

the algorithm with high probability does not find the optimum before time  $T_{n/4}$ .

Around ten years after Droste's work, Sudholt and Witt [37] showed that the  $O(\mu\sqrt{n})$  upper bound also holds for the cGA with frequency boundaries. There (but the same should be true for the cGA without boundaries) a hypothetical population size of  $\mu = \Omega(\sqrt{n} \log n)$  suffices (recall that Droste required  $\mu = \Omega(n^{1/2+\varepsilon})$ ). The technically biggest progress with respect to upper bounds most likely lies in the fact that the analysis in [37] also holds for the expected optimization time, which means that it also includes the rare case that frequencies reach the lower boundary. We refer to Section 2.3 of [12] for a detailed discussion of the relation of expectations and tail bounds for runtimes of EDAs, including a method to transform EDAs with with-high-probability guarantees into those with guarantees on the expected runtime). Sudholt and Witt also show that the cGA with frequency boundaries with high probability (and thus also in expectation) needs at least  $\Omega(\mu\sqrt{n} + n \log n)$  iterations to optimize ONEMAX. While the  $\mu\sqrt{n}$  lower bound could have been also obtained with methods similar to Droste's, the innocent-looking  $\Omega(n \log n)$  bound is surprisingly difficult to prove.

Not much is known for hypothetical population sizes below the order of  $\sqrt{n}$ . It is clear that then the frequencies will reach the lower boundary of the frequency range, so working with a non-trivial lower boundary like  $\frac{1}{n}$  is necessary to prevent premature convergence. The recent lower bound  $\Omega(\mu^{1/3}n)$  valid for  $\mu = O(\frac{\sqrt{n}}{\log n \log \log n})$  of [32] indicates that already a little below the  $\sqrt{n}$  regime significantly larger runtimes occur, but with no upper bounds this regime remains largely not understood.

We refer the reader to the recent survey [28] for more results on the runtime of the cGA on classic unimodal test functions like LEADINGONES and BINVAL. Interestingly, nothing was known for multimodal functions before the recent work of Hasenöhl and Sutton [23] on jump functions, which we discussed already in the introduction.

The general topic of lower bounds on runtimes of EDAs remains largely little understood. Apart from the lower bounds for the cGA on ONEMAX discussed above, the following is known. Krejca and Witt [29] prove a lower bound for the UMDA on ONEMAX, which is of a similar flavor as the lower bound for the cGA of Sudholt and Witt [37]: For  $\lambda = (1 + \beta)\mu$ , where  $\beta > 0$  is a constant, and  $\lambda$  polynomially bounded in  $n$ , the expected runtime of the UMDA on ONEMAX is  $\Omega(\mu\sqrt{n} + n \log n)$ . For the binary value function BINVAL, Droste [16] and Witt [41] together give a lower bound of  $\Omega(\min\{n^2, \mu n\})$  for the runtime of the cGA. Apart from these sparse results, we are not aware of any lower bounds for EDAs. Of course, the black-box complexity of the problem is a lower bound for any black-box algorithm, hence also for EDAs, but these bounds are often lower than the true complexity of a given algorithm. For

example, the black-box complexities of ONEMAX, LEADINGONES, and jump functions with jump size  $k \leq \frac{1}{2}n - n^\varepsilon$ ,  $\varepsilon > 0$  any constant, are  $\Theta(\frac{n}{\log n})$  [2, 18],  $\Theta(n \log \log n)$  [1], and  $\Theta(\frac{n}{\log n})$  [5], respectively.

To complete the picture, we briefly describe some typical runtimes of evolutionary algorithms on jump functions. We recall that the  $n$ -dimensional jump function with jump size  $k \geq 1$  is defined by

$$\text{JUMP}_{nk}(x) = \begin{cases} \|x\|_1 + k & \text{if } \|x\|_1 \in [0..n-k] \cup \{n\}, \\ n - \|x\|_1 & \text{if } \|x\|_1 \in [n-k+1..n-1]. \end{cases}$$

Hence for  $k = 1$ , we have a fitness landscape isomorphic to the one of ONEMAX, but for larger values of  $k$  there is a fitness valley (“gap”)

$$G_{nk} := \{x \in \{0, 1\}^n \mid n-k < \|x\|_1 < n\}$$

consisting of the  $k-1$  highest sub-optimal fitness levels of the ONEMAX function. This valley is hard to cross for evolutionary algorithms using standard-bit mutation with mutation rate  $\frac{1}{n}$  since with very high probability they need to generate the optimum from one of the local optima, which in a single application of the mutation operator happens only with probability less than  $n^{-k}$ . For this reason the  $(1+1)$  EA has a runtime of at least  $n^k$  [17]. By using larger mutation rates or a heavy-tailed mutation operator, a  $k^{\Theta(k)}$  runtime improvement can be obtained [15], but the runtime remains  $\Omega(n^k)$  for  $k$  constant.

Asymptotically better runtimes can be achieved when using crossover, though this is harder than expected. The first work in this direction [27], among other results, could show that a simple  $(\mu+1)$  genetic algorithm using uniform crossover with rate  $p_c = O(\frac{1}{kn})$  obtains an  $O(\mu n^2 k^3 + 2^{2k} p_c^{-1})$  runtime when the population size is at least  $\mu = \Omega(k \log n)$ . A shortcoming of this result, already noted by the authors, is that it only applies to uncommonly small crossover rates. Using a different algorithm that first always applies crossover and then mutation, a runtime of  $O(n^{k-1} \log n)$  was achieved by Dang et al. [9, Theorem 2]. For  $k \geq 3$ , the logarithmic factor in the runtime can be removed by using a higher mutation rate. With additional diversity mechanisms, the runtime can be further reduced up to  $O(n \log n + 4^k)$ , see [8]. In the light of this last result, the insight stemming from the previous work [23] and ours is that the cGA apparently without further modifications supplies the necessary diversity to obtain a runtime of  $O(n \log n + 2^{O(k)})$ .

With a three-parent majority vote crossover, among other results, a runtime of  $O(n \log n)$  could be obtained via a suitable island model for all  $k = O(n^{1/2-\varepsilon})$  [19]. Via a hybrid genetic algorithm using as variation operators only local search and a deterministic voting crossover, an  $O(n)$  runtime was obtained in [38].

Finally, we note that runtimes of  $O(n \binom{n}{k})$  and  $O(k \log(n) \binom{n}{k})$  were shown for the  $(1+1)$  IA<sup>hyp</sup> and the  $(1+1)$  Fast-IA artificial immune systems, respectively [6, 7].

### 2.3 Technical Ingredients of the Main Proof

We now collect a few elementary tools that will be used on our analysis. The first is well-known and the next two are from [12], so it is only the last one which we could not find in the literature.

The following estimate seems well-known (e.g., it was used in [26] without proof or reference). Gießen and Witt [21, Lemma 3]

give a proof via estimates of binomial coefficients and the binomial identity. A more elementary proof can be found in [10, Lemma 10.37].

LEMMA 2. Let  $X \sim \text{Bin}(n, p)$ . Let  $k \in [0..n]$ . Then

$$\Pr[X \geq k] \leq \binom{n}{k} p^k.$$

The next estimate was essentially proven in the extended version of [12, Lemma 1]. The only small difference here is that in the first inequality, we used a slightly different Chernoff bound, which also allows deviation parameters  $\Delta$  which are greater than one. We therefore omit the proof.

LEMMA 3. Let  $f \in [0, 1]^n$ ,  $D := n - \|f\|_1$ ,  $D^- \leq D \leq D^+$ ,  $x \sim \text{Sample}(f)$ , and  $d(x) := n - \|x\|_1$ . Then for all  $\Delta \geq 0$  and  $\delta \in [0, 1]$ , we have

$$\Pr[d(x) \geq (1 + \Delta)D^+] \leq \exp(-\frac{1}{3} \min\{\Delta^2, \Delta\}D^+),$$

$$\Pr[d(x) \leq (1 - \delta)D^-] \leq \exp(-\frac{1}{2} \delta^2 D^-).$$

To estimate the influence from capping the frequencies into the interval  $[\frac{1}{n}, 1 - \frac{1}{n}]$ , the following elementary result was shown in the extended version of [12]. We here recall the common definition that for an  $n$ -dimensional vector  $x$  and a subset  $L \subseteq [1..n]$  of its index set,  $x|_L$  denotes the restriction of  $x$  to  $L$ , that is, the vector  $(x_\ell)_{\ell \in L}$ .

LEMMA 4. Let  $P = 2\frac{1}{n}(1 - \frac{1}{n})$ . Let  $t \geq 0$ . Using the notation given in Algorithm 1, consider iteration  $t+1$  of a run of the cGA started with a fixed frequency vector  $f_t \in [\frac{1}{n}, 1 - \frac{1}{n}]^n$ .

- (i) Let  $L := \{i \in [1..n] \mid f_{it} = \frac{1}{n}\}$ ,  $\ell = |L|$ , and  $M := \{i \in L \mid x_i^1 \neq x_i^2\}$ . Then  $|M| \sim \text{Bin}(\ell, P)$  and  $\|f_{t+1}\|_1 - \|f'_{t+1}\|_1 \leq \|(f_{t+1})|_L\|_1 - \|(f'_{t+1})|_L\|_1 \leq \frac{1}{\mu}|M| \leq \frac{1}{\mu} \text{Bin}(n, \frac{2}{n})$ .
- (ii) Let  $L := \{i \in [1..n] \mid f_{it} = 1 - \frac{1}{n}\}$ ,  $\ell = |L|$ , and  $M := \{i \in L \mid x_i^1 \neq x_i^2\}$ . Then  $|M| \sim \text{Bin}(\ell, P)$  and  $\|f'_{t+1}\|_1 - \|f_{t+1}\|_1 \leq \|(f'_{t+1})|_L\|_1 - \|(f_{t+1})|_L\|_1 \leq \frac{1}{\mu}|M| \leq \frac{1}{\mu} \text{Bin}(n, \frac{2}{n})$ .

To argue that the cGA makes at least some small progress, we shall use the following blunt estimate for the probability that two bit strings  $x, y \sim \text{Sample}(f)$  sampled from the same product distribution have a different distance from the all-ones string (and, by symmetry, from any other string, but this is a statement which we do not need here).

LEMMA 5. Let  $n \in \mathbb{N}$ ,  $m \in [\frac{n}{2}..n]$ , and  $f \in [\frac{1}{n}, 1 - \frac{1}{n}]^m$ . Let  $x^1, x^2 \sim \text{Sample}(f)$  be independent. Then  $\Pr[\|x^1\|_1 \neq \|x^2\|_1] \geq \frac{1}{16}$ .

PROOF. For all  $v \in \mathbb{R}^m$  and  $a, b \in [1..m]$  with  $a \leq b$  we use the abbreviation  $v_{[a..b]} := \sum_{i=a}^b v_i$ . By symmetry, we can assume that  $f_{[1..m]} \leq \frac{m}{2}$ . Without loss of generality, we may further assume that  $f_i \leq f_{i+1}$  for all  $i \in [1..m-1]$ . We have  $f_{[m/4]} \leq \frac{2}{3}$  as otherwise

$$f_{[1..m]} \geq f_{[m/4]+1..m]} > \frac{2}{3}(n - \lfloor m/4 \rfloor) \geq \frac{2}{3} \cdot \frac{3}{4}m = \frac{m}{2},$$

contradicting our assumption.

Let  $\ell$  be minimal such that  $S = f_{[1..\ell]} \geq \frac{1}{8}$ . Since  $\ell \leq \frac{n}{8} \leq \frac{m}{4}$ , we have  $f_\ell \leq \frac{2}{3}$  and thus  $S \leq \frac{1}{8} + \frac{2}{3} = \frac{19}{24}$ .



For  $j \in \{0, 1\}$  let  $q_j = \Pr[x_{[1..\ell]}^1 = j] = \Pr[x_{[1..\ell]}^2 = j]$ . We compute

$$\begin{aligned} \Pr[\|x^1\|_1 \neq \|x^2\|_1] &\geq \Pr[x_{[1..\ell]}^1 = x_{[1..\ell]}^2 \wedge x_{[\ell+1..n]}^1 \neq x_{[\ell+1..n]}^2] \\ &\quad + \Pr[x_{[1..\ell]}^1 \neq x_{[1..\ell]}^2 \wedge x_{[\ell+1..n]}^1 = x_{[\ell+1..n]}^2] \\ &= \Pr[x_{[1..\ell]}^1 = x_{[1..\ell]}^2] \Pr[x_{[\ell+1..n]}^1 \neq x_{[\ell+1..n]}^2] \\ &\quad + \Pr[x_{[1..\ell]}^1 \neq x_{[1..\ell]}^2] \Pr[x_{[\ell+1..n]}^1 = x_{[\ell+1..n]}^2] \\ &\geq \min\{\Pr[x_{[1..\ell]}^1 = x_{[1..\ell]}^2], \Pr[x_{[1..\ell]}^1 \neq x_{[1..\ell]}^2]\} \\ &\geq \min\{q_0^2 + q_1^2, 2q_0q_1\} = 2q_0q_1, \end{aligned}$$

the latter by the inequality of the arithmetic and geometric mean. Using Bernoulli's inequality, we estimate coarsely

$$\begin{aligned} q_0 &= \prod_{i=1}^{\ell} (1 - f_i) \geq 1 - f_{[1..\ell]}, \\ q_1 &= \sum_{i=1}^{\ell} f_i \prod_{j \in [1..\ell] \setminus \{i\}} (1 - f_j) \geq f_{[1..\ell]} (1 - f_{[1..\ell]}). \end{aligned}$$

From the concavity of  $z \mapsto z(1 - z)^2$  in  $[0, 1]$ , we obtain

$$\begin{aligned} 2q_0q_1 &\geq 2 \min\{z(1 - z)^2 \mid z \in [\tfrac{1}{8}, \tfrac{19}{24}]\} \\ &= 2 \min\{z(1 - z)^2 \mid z \in [\tfrac{1}{8}, \tfrac{19}{24}]\} = 2 \tfrac{19}{24} (\tfrac{5}{25})^2 \geq \tfrac{1}{16}. \end{aligned}$$

□

### 3 MAIN RESULT

We now state precisely our main result, explain the central proof ideas, and state the formal proof.

**THEOREM 6.** *There are constants  $\alpha_1, \alpha_2 > 0$  such that for any  $n$  sufficiently large and any  $k \in [1..n]$ , regardless of the hypothetical population size  $\mu$ , the runtime of the cGA on  $JUMP_{nk}$  with probability  $1 - \exp(-\alpha_1 k)$  is at least  $\exp(\alpha_2 k)$ . In particular, the expected runtime is exponential in  $k$ .*

We note that we intentionally prove a runtime bound that holds with high probability. The reason is that EDAs may with small probability reach states from which they find it very hard to reach the optimum. Such a situation could lead to a very high expected runtime even when the EDA with high probability is very efficient. For that reason, lower bounds that hold with high probability are particularly desirable for EDAs. Needless to say, a lower bound that holds with high probability immediately implies an asymptotically identical lower bound on the expected runtime.

To prove this result, we will regard the stochastic process  $D_t := n - \|f_t\|_1$ , that is, the difference of the sum of the frequencies from the ideal value  $n$ . Our general argument is that this process with probability  $1 - \exp(-\Omega(k))$  stays above  $\frac{1}{4}k$  for  $\exp(\Omega(k))$  iterations. In each iteration where  $D_t \geq \frac{1}{4}k$ , the probability that the optimum is sampled, is only  $\exp(-\Omega(k))$ . Hence there is a  $T = \exp(\Omega(k))$  such that with probability  $1 - \exp(-\Omega(k))$ , the optimum is not sampled in the first  $T$  iterations.

The heart of the proof is an analysis of the process  $(D_t)$ . It is intuitively clear that once the process is below  $k$ , then often the two search points sampled in one iteration both lie in the gap

region, which gives a positive drift (that is, a decrease of the average frequency). To turn this drift away from the target (a small  $D_t$  value) into an exponential lower bound for the runtime, we consider the process  $Y_t = \exp(c \min\{\frac{1}{2}k - D_t, \frac{1}{4}k\})$ , that is, an exponential rescaling of  $D_t$ . Such a rescaling has recently also been used in [3]. We note that the usual way to prove exponential lower bounds is the negative drift theorem of Oliveto and Witt [35]. We did not immediately see how to use it for our purposes, though, since in our process we do not have very strong bounds on the one-step differences. E.g., when  $D_t = \frac{1}{2}k$ , then the underlying frequency vector may be such that  $D_{t+1} \geq D_t + \sqrt{k}$  happens with constant probability.

We shall show that the process  $Y_t$  has at most a constant point-wise drift, more precisely, that

$$E[Y_{t+1} - Y_t \mid Y_t = y] \leq 2 \quad (1)$$

holds for all  $y < Y_{\max} := \exp(\frac{c}{4}k)$ . From this statement, the additive drift theorem of He and Yao [24] (see also [31]) would immediately show that the expected time to reach a  $D_t$  value of  $\frac{k}{4}$  or less is at least exponential in  $k$ . However, since we aim at a runtime bound that holds with high probability, we take a different (and, in fact, more elementary) route. We regard the process  $(\tilde{Y}_t)$  which is identical to  $Y$  until  $Y$  first reaches  $Y_{\max}$  and then stays constant at  $Y_{\max}$ . This process satisfies  $E[\tilde{Y}_{t+1} - \tilde{Y}_t] \leq 2$  for all times  $t$ . From this and  $\tilde{Y}_0 = Y_0 < 1$  we obtain  $E[\tilde{Y}_t] \leq 1 + 2t$ . Hence for  $T = \exp(\Omega(k))$  sufficiently small, we have

$$\frac{E[\tilde{Y}_T]}{Y_{\max}} = \exp(-\Omega(k))$$

and a simple Markov bound argument is enough to show that  $\Pr[\tilde{Y}_T = Y_{\max}] = \exp(-\Omega(k))$ . Note that  $\tilde{Y}_T < Y_{\max}$  is equivalent to  $Y_t < Y_{\max}$  for all  $t \in [0..T]$ .

The main work in the following proof is showing (1). The difficulty here is hidden in a small detail. When  $D_t \in [\frac{1}{4}k, \frac{3}{4}k]$ , and this is the most interesting case (case 2 in the formal proof), then we have  $\|f'_{t+1}\|_1 \leq \|f_t\|_1$  whenever the two search points sampled lie in the gap region, and hence with probability  $1 - \exp(-\Omega(k))$ ; from Lemma 5 we obtain, in addition, a true decrease, that is,  $\|f'_{t+1}\|_1 \leq \|f_t\|_1 - \frac{1}{\mu}$ , with constant probability. This progress of  $f'_{t+1}$  over  $f_t$  would be perfectly fine for our purposes. Hence the true difficulty arises from the capping of the frequencies into the interval  $[\frac{1}{n}, 1 - \frac{1}{n}]$ , that is, from the fact that the new frequency vector is  $f_{t+1} := \minmax(\frac{1}{n} \mathbf{1}_n, f'_{t+1}, (1 - \frac{1}{n}) \mathbf{1}_n)$ . This appears to be a minor problem, among others, because only a capping at the lower bound  $\frac{1}{n}$  can have an adverse effect on our process, and there are at most  $O(k)$  frequencies sufficiently close to the lower boundary. Things become difficult due to the exponential scaling, which can let rare events still have a significant influence on the expected change of the process.

We now make these arguments precise and prove Theorem 6.

**PROOF.** Since we are aiming at an asymptotic statement, we can assume in the following that  $n$  is sufficiently large.

To ease the presentation of the main part of the proof, let us first give a basic argument for the case of small  $k$  and then assume that  $k \geq w(n)$  for some function  $w : \mathbb{N} \rightarrow \mathbb{N}$  with  $\lim_{n \rightarrow \infty} w(n) = \infty$ .

We first note that with probability  $f_{it}^2 + (1 - f_{it})^2 \geq \frac{1}{2}$ , the two search points  $x^1$  and  $x^2$  generated in the  $t$ -th iteration agree on the  $i$ -th bit, which in particular implies that  $f_{i,t+1} = f_{it}$ . Hence with probability at least  $2^{-T}$ , this happens for the first  $T$  iterations, and thus  $f_{it} = \frac{1}{2}$  for all  $t \in [0..T]$ . Let us call such a bit position  $i$  *sleepy*.

Note that the events of being sleepy are independent for all  $i \in [1..n]$ . Hence, taking  $T = \lfloor \frac{1}{2} \log_2 n \rfloor$ , we see that the number  $X$  of sleepy positions has an expectation of  $E[X] \geq n2^{-T} \geq \sqrt{n}$ , and by a simple Chernoff bound, we have  $\Pr[X \geq \frac{1}{2}\sqrt{n}] \geq 1 - \exp(-\Omega(\sqrt{n}))$ .

Conditional on having at least  $\frac{1}{2}\sqrt{n}$  sleepy bit positions, the probability that a particular search point sampled in the first  $T$  iterations is the optimum is at most  $2^{-\frac{1}{2}\sqrt{n}}$ . By a simple union bound argument, the probability that at least one of the search points generated in the first  $T$  iterations is the optimum is at most  $2T2^{-\frac{1}{2}\sqrt{n}} = \exp(-\Omega(\sqrt{n}))$ . In summary, we have that with probability at least  $1 - \exp(-\Omega(\sqrt{n}))$ , the runtime of the cGA on any function with unique optimum (and in particular any jump function) is greater than  $T = \frac{1}{2} \log_2 n$ . This implies the claim of this theorem for any  $k \leq C \log \log n$ , where  $C$  is a sufficiently small constant, and, as discussed above,  $n$  is sufficiently large.

With this, we can now safely assume that  $k = \omega(1)$ . For the case that  $k \geq \frac{n}{320}$ , we will need slightly modified calculations. To keep this proof readable, we hide (but treat nevertheless) this case as follows. We consider a run of the cGA on a jump function  $\text{JUMP}_{nk'}$  with  $k' \in [1..n] \cap \omega(1)$  arbitrary and we let  $k := \min\{k', \frac{n}{320}\}$ .

Let  $D_t := n - \|f_t\|_1 = n - \sum_{i=1}^n f_{it}$  be the distance of the sum of the frequencies from the ideal value  $n$ .

Our intuition (which will be made precise) is that the process ( $D_t$ ) finds it hard to go significantly below  $k$  because there we will typically sample individuals in the gap, which lead to a decrease of the sum of frequencies (when the two individuals have different distances from the optimum). To obtain an exponential lower bound on the runtime, we suitably rescale the process by defining, for a sufficiently small constant  $c$ ,

$$Y_t = \min\{\exp(c(\frac{1}{2}k - D_t)), \exp(\frac{1}{4}ck)\} = \exp(c \min\{\frac{1}{2}k - D_t, \frac{1}{4}k\}).$$

Observe that  $Y_t$  attains its maximal value  $Y_{\max} = \exp(\frac{1}{4}ck)$  precisely when  $D_t \leq \frac{1}{4}k$ . Also,  $Y_t \leq 1$  for  $D_t \geq \frac{1}{2}k$ .

To argue that we have  $D_t > \frac{1}{4}k$  for a long time, we now show that for all  $y < Y_{\max}$  the drift  $E[Y_{t+1} - Y_t \mid Y_t = y]$  is at most constant. To this aim, we condition on a fixed value of  $f_t$ , which also determines  $D_t$ . We treat separately the two cases that  $D_t \geq \frac{3}{4}k$  and that  $\frac{3}{4}k > D_t > \frac{1}{4}k$ .

**Case 1:** Assume first that  $D_t \geq \frac{3}{4}k$ . By Lemma 3, with probability  $1 - \exp(-\Omega(D_t)) \geq 1 - \exp(-\Omega(k))$ , the two search points  $x^1, x^2$  sampled in iteration  $t+1$  both satisfy

$$\|x^i\|_1 - \|f_t\|_1 = |d(x^i) - D_t| < \frac{1}{18}D_t \leq \frac{1}{6}(D_t - \frac{1}{2}k). \quad (2)$$

Here and in the following, when writing  $\Omega(k)$  we mean that there is a positive constant  $C$ , independent of  $n$ ,  $k$ , and  $c$ , such that the expression is at least  $Ck$ . Let us call  $A$  the event described in (2). In this case, we argue as follows. We use the notation  $\Sigma[v] := \sum_{i=1}^n v_i$  to denote the sum of the elements of an  $n$ -dimensional vector  $v$  and we recall further that, with a slight abuse of notation, we defined  $\|f'\|_1 := \Sigma[f']$  for intermediate frequency vectors  $f'$ . Let

$\{y^1, y^2\} = \{x^1, x^2\}$  such that  $\mathcal{F}(y^1) \geq \mathcal{F}(y^2)$ . Then

$$\begin{aligned} \|f'_{t+1}\|_1 &= \sum \left[ f_t + \frac{1}{\mu}(y^1 - y^2) \right] \\ &= \sum \left[ f_t + \frac{1}{\mu}(y^1 - f_t) - \frac{1}{\mu}(y^2 - f_t) \right] \\ &\leq \sum \left[ f_t \right] + \frac{1}{\mu} \left| \sum [y^1 - f_t] \right| + \frac{1}{\mu} \left| \sum [y^2 - f_t] \right| \\ &= \|f_t\|_1 + \frac{1}{\mu} \left| \|x^1\|_1 - \|f_t\|_1 \right| + \frac{1}{\mu} \left| \|x^2\|_1 - \|f_t\|_1 \right| \\ &\leq n - D_t + 2 \frac{1}{\mu} \frac{1}{6} (D_t - \frac{1}{2}k) \\ &\leq n - D_t + 2 \frac{1}{6} (D_t - \frac{1}{2}k) \\ &= n - \frac{2}{3}D_t - \frac{1}{6}k \leq n - \frac{2}{3} \cdot \frac{3}{4}k - \frac{1}{6}k \leq n - \frac{2}{3}k. \end{aligned}$$

We still need to consider the possibility that  $f_{i,t+1} > f'_{i,t+1}$  for some  $i \in [1..n]$ . By Lemma 4, not conditioning on  $A$ , we have that  $\|f_{t+1}\|_1 - \|f'_{t+1}\|_1 \leq \frac{1}{\mu} \text{Bin}(\ell, P) \leq \text{Bin}(\ell, P)$  for some  $\ell \in [1..n]$  and  $P = 2 \frac{1}{n} (1 - \frac{1}{n})$ .

Let us call  $B$  the event that  $\|f_{t+1}\|_1 - \|f'_{t+1}\|_1 < \frac{1}{6}k$ . Note that  $A \cap B$  implies  $\|f_{t+1}\|_1 < n - \frac{1}{2}k$  and thus  $Y_{t+1} \leq 1$ . By Lemma 2 and the estimate  $\binom{a}{b} \leq (\frac{ea}{b})^b$ , we have

$$\Pr[\neg B] \leq \left( \frac{\ell}{\frac{1}{6}k} \right)^{P^{k/6}} \leq \left( \frac{12e\ell}{kn} \right)^{k/6} \leq k^{-\Omega(k)}.$$

We conclude that the event  $A \cap B$  holds with probability  $1 - \exp(-\Omega(k))$ ; in this case  $Y_t \leq 1$  and  $Y_{t+1} \leq 1$ . In all other cases, we bluntly estimate  $Y_{t+1} - Y_t \leq Y_{\max}$ . This gives

$$E[Y_{t+1} - Y_t] \leq (1 - \exp(-\Omega(k))) \cdot 1 + \exp(-\Omega(k)) Y_{\max}.$$

By choosing the constant  $c$  in the definition of ( $Y_t$ ) sufficiently small and taking  $n$  sufficiently large, we have  $E[Y_{t+1} - Y_t] \leq 2$ .

**Case 2:** Assume now that  $\frac{3}{4}k > D_t > \frac{1}{4}k$ . Let  $x^1, x^2$  be the two search points sampled in iteration  $t+1$ . By Lemma 3 again, we have  $k > n - \|x^i\|_1 > 0$  with probability  $1 - \exp(-\Omega(k))$  for both  $i \in \{1, 2\}$ . Let us call this event  $A$ . Note that if  $A$  holds, then both offspring lie in the gap region. Consequently,  $\|y^1\|_1 \leq \|y^2\|_1$  and thus  $\|f'_{t+1}\|_1 \leq \|f_t\|_1$ .

Let  $L = \{i \in [1..n] \mid f_{it} = \frac{1}{n}\}$ ,  $\ell = |L|$ , and  $M = \{i \in L \mid x^1_i \neq x^2_i\}$  as in Lemma 4. Note that by definition,  $D_t \geq (1 - \frac{1}{n})\ell$ , hence from  $D_t < \frac{3}{4}k$  and  $n \geq 4$  we obtain  $\ell < k$ .

Let  $B_0$  be the event that  $|M| = 0$ , that is,  $x^1_L = x^2_L$ . Note that in this case,

$$\begin{aligned} \|f_{t+1}\|_1 &\leq \|f'_{t+1}\|_1 = \|f_t + \frac{1}{\mu}(y^1 - y^2)\|_1 \\ &= \|f_t\|_1 + \frac{1}{\mu}(\|y^1\|_1 - \|y^2\|_1). \end{aligned}$$

By Lemma 4, Bernoulli's inequality, and  $\ell \leq k$ , we have

$$\Pr[B_0] = (1 - 2 \frac{1}{n} (1 - \frac{1}{n}))^\ell \geq 1 - \frac{2\ell}{n} \geq 1 - \frac{2k}{n}.$$

Since  $\ell < k \leq \frac{n}{320} < \frac{n}{2}$ , by Lemma 5, we have  $\|x^1_{[n] \setminus L}\|_1 \neq \|x^2_{[n] \setminus L}\|_1$  with probability at least  $\frac{1}{16}$ . This event, called  $C$  in the following, is independent of  $B_0$ . We have

$$\begin{aligned} \Pr[A \cap B_0 \cap C] &\geq \Pr[B_0 \cap C] - \Pr[\bar{A}] \\ &\geq (1 - \frac{2k}{n}) \frac{1}{16} - \exp(-\Omega(k)). \end{aligned}$$

If  $A \cap B_0 \cap C$  holds, then  $\|f_{t+1}\|_1 \leq \|f_t\|_1 - \frac{1}{\mu}$ . If  $A \cap B_0 \cap \bar{C}$  holds, then we still have  $\|f_{t+1}\|_1 \leq \|f_t\|_1$ .

Let us now, for  $j \in [1..\ell]$ , denote by  $B_j$  the event that  $|M| = j$ , that is, that  $x_L^1$  and  $x_L^2$  differ in exactly  $j$  bits. By Lemma 4 again, we have  $\Pr[B_j] = \Pr[\text{Bin}(\ell, P) = j]$ .

The event  $A \cap B_j$  implies  $\|f_{t+1}\|_1 \leq \|f'_t\|_1 + \frac{j}{\mu} \leq \|f_t\|_1 + \frac{j}{\mu}$  and occurs with probability  $\Pr[A \cap B_j] \leq \Pr[B_j] = \Pr[\text{Bin}(\ell, P) = j]$ .

Taking these observations together, we compute

$$\begin{aligned} E[Y_{t+1}] &= \Pr[\bar{A}] E[Y_{t+1} | \bar{A}] \\ &\quad + \sum_{j=1}^{\ell} \Pr[A \cap B_j] E[Y_{t+1} | A \cap B_j] \\ &\quad + \Pr[A \cap B_0 \cap \bar{C}] E[Y_{t+1} | A \cap B_0 \cap \bar{C}] \\ &\quad + \Pr[A \cap B_0 \cap C] E[Y_{t+1} | A \cap B_0 \cap C] \\ &\leq \exp(-\Omega(k)) Y_{\max} \\ &\quad + \sum_{j=1}^{\ell} \Pr[\text{Bin}(\ell, P) = j] Y_t \exp\left(\frac{cj}{\mu}\right) \\ &\quad + \Pr[\text{Bin}(\ell, P) = 0] Y_t \\ &\quad - \left(\frac{1}{16} \left(1 - \frac{2k}{n}\right) - \exp(-\Omega(k))\right) Y_t (1 - \exp(-\frac{c}{\mu})). \end{aligned} \quad (3)$$

We note that the second and third term amount to  $Y_t E[\exp(\frac{cZ}{\mu})]$ , where  $Z \sim \text{Bin}(\ell, P)$ . Writing  $Z = \sum_{i=1}^{\ell} Z_i$  as a sum of  $\ell$  independent binary random variables with  $\Pr[Z_i = 1] = P$ , we obtain

$$E[\exp(\frac{cZ}{\mu})] = \prod_{i=1}^{\ell} E[\exp(\frac{cZ_i}{\mu})] = (1 - P + P \exp(\frac{c}{\mu}))^{\ell}.$$

By assuming  $c \leq 1$  and using the elementary estimate  $e^x \leq 1 + 2x$  valid for  $x \in [0, 1]$ , see, e.g., Lemma 4.2(b) in [10], we have

$$1 - P + P \exp(\frac{c}{\mu}) \leq 1 + 2P(\frac{c}{\mu}).$$

Hence with  $P \leq \frac{2}{n}$ ,  $\mu \geq 1$ , and  $\ell \leq \frac{n}{320}$ , we obtain

$$E[\exp(\frac{cZ}{\mu})] \leq (1 + 2P(\frac{c}{\mu}))^{\ell} \leq \exp(2P(\frac{c}{\mu})\ell) \leq \exp(\frac{4c}{320\mu}) \leq 1 + \frac{c}{40\mu},$$

again by using  $e^x \leq 1 + 2x$ . The second and third term of (3) thus add up to at most  $(1 + \frac{c}{40\mu}) Y_t$ .

In the first term of (3), we again assume that  $c$  is sufficiently small to ensure that  $\exp(-\Omega(k)) Y_{\max} = \exp(-\Omega(k)) \exp(\frac{1}{4}ck) \leq 1$ . Recalling that  $k \leq \frac{n}{320}$  and assuming  $k$  sufficiently large (since  $k = \omega(1)$  and  $n$  is large), we finally estimate in the last term  $\frac{1}{16} \left(1 - \frac{2k}{n}\right) - \exp(-\Omega(k)) \geq \frac{1}{20}$  and, more interestingly,  $1 - \exp(-\frac{c}{\mu}) \geq \frac{c}{\mu} \left(1 - \frac{1}{e}\right)$  using the estimate  $e^{-x} \leq 1 - x \left(1 - \frac{1}{e}\right)$  valid for all  $x \in [0, 1]$ , which stems simply from the convexity of the exponential function.

With these estimates we obtain

$$E[Y_{t+1}] \leq 1 + \left(1 + \frac{c}{40\mu}\right) Y_t - \frac{1}{20} \left(1 - \frac{1}{e}\right) \frac{c}{\mu} Y_t \leq 1 + Y_t$$

and thus  $E[Y_{t+1} - Y_t] \leq 1$ .

In summary, we have now shown that for all  $y < Y_{\max}$  and at all times  $t$  the process  $(Y_t)$  satisfies  $E[Y_{t+1} - Y_t | Y_t = y] \leq 2$ . We note that  $Y_0 \leq 1$  with probability one. For the sake of the argument, let us artificially modify the process from the point on when it has reached a state of at least  $Y_{\max}$ . So we define  $(\tilde{Y}_t)$  by setting  $\tilde{Y}_t = Y_t$ , if  $Y_t < Y_{\max}$  or if  $Y_t \geq Y_{\max}$  and  $Y_{t-1} < Y_{\max}$ , and  $\tilde{Y}_t = Y_{t-1}$  otherwise. In other words,  $(\tilde{Y}_t)$  is a copy of  $(Y_t)$  until it

reaches a state of at least  $Y_{\max}$  and then does not move anymore. With this trick, we have  $E[\tilde{Y}_{t+1} - \tilde{Y}_t] \leq 2$  for all  $t$ .

A simple induction and the initial condition  $\tilde{Y}_0 \leq 1$  shows that  $E[\tilde{Y}_t] \leq 2t + 1$  for all  $t$ . In particular, for  $T = \frac{1}{2} \exp(\frac{1}{8}ck) - 1$ , we have  $E[Y_T] \leq \exp(\frac{1}{8}ck)$  and, by Markov's inequality,

$$\Pr[\tilde{Y}_T \geq Y_{\max}] \leq \frac{\exp(\frac{1}{8}ck)}{Y_{\max}} = \exp(-\frac{1}{8}ck).$$

Hence with probability  $1 - \exp(-\frac{1}{8}ck)$ , we have  $\tilde{Y}_T < Y_{\max}$ . We now condition on this event. By construction of  $(\tilde{Y}_t)$ , we have  $Y_t < Y_{\max}$ , equivalently  $D_t > \frac{1}{4}k$ , for all  $t \in [0..T]$ . If  $D_t > \frac{1}{4}k$ , then the probability that a sample generated in this iteration is the optimum, is at most

$$\begin{aligned} \prod_{i=1}^n f_{it} &= \prod_{i=1}^n (1 - (1 - f_{it})) \\ &\leq \prod_{i=1}^n \exp(-(1 - f_{it})) \\ &= \exp(-(n - \|f_t\|_1)) \\ &= \exp(-D_t) \leq \exp(-\frac{1}{4}k). \end{aligned}$$

Assuming  $c \leq 1$  again, we see that the probability that the optimum is generated in one of the first  $T$  iterations, is at most  $2T \exp(-\frac{1}{4}k) \leq \exp(\frac{1}{8}ck) \exp(-\frac{1}{4}k) = \exp(-\frac{1}{8}ck)$ . This shows the claim.  $\square$

#### 4 AN $\Omega(n \log n)$ LOWER BOUND?

With the exponential lower bound proven in the previous section, the runtime of the cGA on jump functions is well understood, except that the innocent looking lower bound  $\Omega(n \log n)$ , matching the corresponding upper bound for  $k < \frac{1}{20} \ln n$ , is still missing. Since Sudholt and Witt [37] have proven an  $\Omega(n \log n)$  lower bound for the simple unimodal function ONEMAX, which for many EAs is known to be one of the easiest functions with unique global optimum [11, 13, 36, 39], it would be very surprising if this lower bound would not hold for jump functions as well.

Unfortunately, we do not see any easy way to prove such a lower bound. We strongly believe that the proof of [37] can be extended to also include jump functions, but since this proof is truly complicated, we shy away from taking such an effort to prove a result that would be that little surprising. We instead argue here why the usual “ONEMAX is the easiest case” argument fails. While we would not say that it is not a valuable research goal to extend the proof of [37] to jump functions, we would much prefer if someone could prove a general  $\Omega(n \log n)$  lower bound for all functions with unique global optimum (or disprove this statement).

The true reason why ONEMAX is the easiest optimization problem for many evolutionary algorithms, implicit in all such proofs and explicit in [11], is that when comparing a run of the evolutionary algorithm on ONEMAX and some other function  $\mathcal{F}$  with unique global optimum, then at all times the Hamming distance between the current-best solution and the optimum in the ONEMAX process is stochastically dominated by the one of the other process. This follows by induction and a coupling argument from the following key insight (here formulated for the  $(1 + 1)$  EA only).

LEMMA 7. Let  $\mathcal{F} : \{0, 1\}^n \rightarrow \mathbb{R}$  be some function with unique global optimum  $x^*$  and let ONEMAX be the  $n$ -dimensional ONEMAX function with unique global optimum  $y^* = (1, \dots, 1)$ . Let  $x, y \in \{0, 1\}^n$  such that  $H(x, x^*) \geq H(y, y^*)$ , where  $H(\cdot, \cdot)$  denotes the Hamming distance. Consider one iteration of the  $(1 + 1)$  EA optimizing  $\mathcal{F}$ , started with  $x$  as parent individual, and denote by  $x'$  the parent in the next iteration. Define  $y'$  analogously for ONEMAX and  $y$ . Then  $H(x', x^*) \geq H(y', y^*)$ .

As a side remark, note that the lemma applied in the special case  $\mathcal{F} = \text{ONEMAX}$  shows that the intuitive rule “the closer a search point is to the optimum, the shorter is the optimization time when starting from this search point” holds for optimizing ONEMAX via the  $(1 + 1)$  EA.

We now show that a statement like Lemma 7 is not true for the cGA. Since the states of a run of the cGA are the frequency vectors  $f$ , the natural extension of the Hamming distance quality measure above is the  $\ell_1$ -distance  $d(f, x^*) = \|f - x^*\|_1 = \sum_{i=1}^n |f_i - x_i^*|$ .

Consider now a run of the cGA on an  $n$ -dimensional ( $n$  even for simplicity) jump function  $\mathcal{F}$  with jump size  $k \leq n/4$ . Consider one iteration starting with the frequency vector  $f = \frac{1}{2}\mathbf{1}_n$ . For comparison, consider one iteration of the cGA optimizing ONEMAX, starting with a frequency vector  $g \in [0, 1]^n$  such that half the entries of  $g$  are equal to  $\frac{1}{n} + \frac{1}{\mu}$  and the other half equals  $1 - \frac{1}{n} - \frac{1}{\mu}$ . Let us take  $\mu = n$  for simplicity. Note that both  $\mathcal{F}$  and ONEMAX have the same unique global optimum  $x^* = y^* = (1, \dots, 1)$ .

We obviously have  $d(f, x^*) \geq d(g, y^*)$ , since both numbers are equal to  $\frac{n}{2}$ . Let  $f', g'$  be the frequency vectors after one iteration. Since with probability  $1 - \exp(-\Omega(n))$ , both search points sampled in the jump process have between  $\frac{n}{4}$  and  $\frac{3n}{4n}$  ones, their jump fitnesses equal their ONEMAX fitnesses. Consequently, we may apply Lemma 5 from [16] and see that  $E[d(f', x^*)] \leq \frac{n}{2} - \Omega(\frac{1}{\mu}\sqrt{n})$ . For the ONEMAX process, however, denoting the two search points generated in this iteration by  $x^1$  and  $x^2$ , we see

$$\begin{aligned} E[\|g - g'\|_1]^2 &\leq E[\|g - g'\|_1^2] = E\left[\left(\frac{1}{\mu} \sum_{i=1}^n (x_i^1 - x_i^2)\right)^2\right] \\ &= \frac{1}{\mu^2} \text{Var}\left[\sum_{i=1}^n (x_i^1 - x_i^2)\right] = O(1) \end{aligned}$$

and hence  $E[d(g', y^*)] \geq d(g, y^*) - O(1) = \frac{n}{2} - O(1)$ . Consequently, we cannot have  $d(f', x^*) \geq d(g', y^*)$ .

We note that a second imaginable domination result is also not true. Assume, for simplicity, that we optimize a function  $\mathcal{F}$  with unique global maximum equal to  $(1, \dots, 1)$  and the function ONEMAX via the cGA with same parameter setting. If  $f \leq g$  (component-wise), and  $f'$  is the frequency vector resulting from one iteration optimizing  $\mathcal{F}$  starting with  $f$  and  $g'$  is the frequency vector resulting from one iteration optimizing ONEMAX starting with  $g$ , then in general we do not have  $f'_i \leq g'_i$  for all  $i \in [1..n]$ .

As counter-example, let  $f = (\frac{1}{2}, \frac{1}{n}, \dots, \frac{1}{n})$ , but now  $g = \frac{1}{2}\mathbf{1}_n$ . Clearly,  $f \leq g$ . We now consider the results of one iteration of the cGA, always with the ONEMAX function as the objective. When performing one iteration of the cGA on ONEMAX started with  $f$ , and denoting the two samples by  $x^1$  and  $x^2$  and their quality difference in all but the first bit by  $\Delta = \|x_{[2..n]}^1\|_1 - \|x_{[2..n]}^2\|_1$ , then the

resulting frequency vector  $f'$  satisfies

$$\begin{aligned} \Pr[f'_1 = \tfrac{1}{2} + \tfrac{1}{\mu}] &= \Pr[x_1^1 \neq x_1^2] \left( \tfrac{1}{2} \Pr[\Delta \notin \{-1, 0\}] + \Pr[\Delta \in \{-1, 0\}] \right) \\ &= \Pr[x_1^1 \neq x_1^2] \left( \tfrac{1}{2} + \tfrac{1}{2} \Pr[\Delta \in \{-1, 0\}] \right). \end{aligned} \quad (4)$$

Since

$$\begin{aligned} \Pr[\Delta \in \{-1, 0\}] &\geq \Pr[\|x_{[2..n]}^1\|_1 = \|x_{[2..n]}^2\|_1 = 0] \\ &= \left(1 - \frac{1}{n}\right)^{2(n-1)} \geq \frac{1}{e^2}, \end{aligned}$$

we have  $\Pr[f'_1 = \frac{1}{2} + \frac{1}{\mu}] \geq \frac{1}{4} + \frac{1}{4e^2}$ .

When starting the iteration with  $g$ , the resulting frequency vector  $g'$  satisfies an equation analogous to (4), but now  $\Delta$  is the difference of two binomial distributions with parameters  $n - 1$  and  $\frac{1}{2}$ . Hence, we have  $\Pr[\Delta \in \{-1, 0\}] = O(n^{-1/2})$ , see, e.g., [10, Lemma 4.13] for this elementary estimate, and thus  $\Pr[g'_1 = \frac{1}{2} + \frac{1}{\mu}] = \frac{1}{4} + o(1)$ , disproving that  $f'_1 \leq g'_1$ .

In summary, the richer mechanism of building a probabilistic model of the search space in the cGA (as opposed to using a population in EAs) makes it hard to argue that ONEMAX is the easiest function for the cGA. This, in particular, has the consequence that lower bounds for the runtime of the cGA on ONEMAX cannot be easily extended to other functions with a unique global optimum.

## 5 CONCLUSION

The main result of this work is an  $\exp(\Omega(k))$  lower bound for the runtime of the cGA on jump functions with jump size  $k$ , regardless of the hypothetical population size  $\mu$ . This in particular shows that the result of Hasenöhrl and Sutton [23] cannot be improved by running the cGA with a hypothetical population size that is sub-exponential in  $k$ . An interesting question is to what extent similar lower bounds hold for other EDAs. Natural candidates for such an investigation could be the significant-based cGA [14], which is not one of the standard EDAs, but algorithmically relatively close to the classic cGA, or the UMDA [33], for which a decent theoretical understanding exists, in particular, on ONEMAX [29, 30, 40].

What is noteworthy in our proof is that it does not require a distinction between the different cases that frequencies reach boundary values or not (as in, e.g., the highly technical lower bound proof for ONEMAX in [37]). It seems to be an interesting direction for future research to find out to what extent such an approach can be used also for other lower bound analyses.

As a side result, we observed that two natural domination arguments that could help showing that ONEMAX is the easiest function for the cGA are not true. For this reason, the natural lower bound of  $\Omega(n \log n)$  remains unproven. Proving it, or even better, proving that  $\Omega(n \log n)$  is a lower bound for the runtime of the cGA on any function  $\mathcal{F} : \{0, 1\}^n \rightarrow \mathbb{R}$  with a unique global optimum, remain challenging open problems.

## ACKNOWLEDGMENTS

This work was supported by a public grant as part of the Investissement d’avenir project, reference ANR-11-LABX-0056-LMH, LabEx LMH, in a joint call with Gaspard Monge Program for optimization, operations research and their interactions with data sciences.



## REFERENCES

- [1] Peyman Afshani, Manindra Agrawal, Benjamin Doerr, Carola Doerr, Kasper Green Larsen, and Kurt Mehlhorn. 2019. The query complexity of a permutation-based variant of Mastermind. *Discrete Applied Mathematics* 260 (2019), 28–50.
- [2] Gautham Anil and R. Paul Wiegand. 2009. Black-box search by elimination of fitness functions. In *Foundations of Genetic Algorithms, FOGA 2009*. ACM, 67–78.
- [3] Denis Antipov, Benjamin Doerr, and Quentin Yang. 2019. The efficiency threshold for the offspring population size of the  $(\mu, \lambda)$  EA. In *Genetic and Evolutionary Computation Conference, GECCO 2019*. ACM. To appear.
- [4] Anne Auger and Benjamin Doerr (Eds.). 2011. *Theory of Randomized Search Heuristics*. World Scientific Publishing.
- [5] Maxim Buzdalov, Benjamin Doerr, and Mikhail Kever. 2016. The unrestricted black-box complexity of jump functions. *Evolutionary Computation* 24 (2016), 719–744.
- [6] Dogan Corus, Pietro Simone Oliveto, and Donya Yazdani. 2017. On the runtime analysis of the Opt-IA artificial immune system. In *Genetic and Evolutionary Computation Conference, GECCO 2017*. ACM, 83–90.
- [7] Dogan Corus, Pietro Simone Oliveto, and Donya Yazdani. 2018. Fast artificial immune systems. In *Parallel Problem Solving from Nature, PPSN 2018*. Springer, 67–78.
- [8] Duc-Cuong Dang, Tobias Friedrich, Timo Kötzing, Martin S. Krejca, Per Kristian Lehre, Pietro Simone Oliveto, Dirk Sudholt, and Andrew M. Sutton. 2016. Escaping local optima with diversity mechanisms and crossover. In *Genetic and Evolutionary Computation Conference, GECCO 2016*. ACM, 645–652.
- [9] Duc-Cuong Dang, Tobias Friedrich, Timo Kötzing, Martin S. Krejca, Per Kristian Lehre, Pietro Simone Oliveto, Dirk Sudholt, and Andrew M. Sutton. 2018. Escaping local optima using crossover with emergent diversity. *IEEE Transactions on Evolutionary Computation* 22 (2018), 484–497.
- [10] Benjamin Doerr. 2018. Probabilistic Tools for the Analysis of Randomized Optimization Heuristics. *CoRR* abs/1801.06733 (2018). arXiv:1801.06733
- [11] Benjamin Doerr. 2019. Analyzing randomized search heuristics via stochastic domination. *Theoretical Computer Science* 773 (2019), 115–137.
- [12] Benjamin Doerr. 2019. 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*. ACM. To appear.
- [13] Benjamin Doerr, Daniel Johannsen, and Carola Winzen. 2012. Multiplicative drift analysis. *Algorithmica* 64 (2012), 673–697.
- [14] Benjamin Doerr and Martin S. Krejca. 2018. Significance-based estimation-of-distribution algorithms. In *Genetic and Evolutionary Computation Conference, GECCO 2018*. ACM, 1483–1490.
- [15] Benjamin Doerr, Huu Phuoc Le, Régis Makhlara, and Ta Duy Nguyen. 2017. Fast genetic algorithms. In *Genetic and Evolutionary Computation Conference, GECCO 2017*. ACM, 777–784.
- [16] Stefan Droste. 2006. A rigorous analysis of the compact genetic algorithm for linear functions. *Natural Computing* 5 (2006), 257–283.
- [17] Stefan Droste, Thomas Jansen, and Ingo Wegener. 2002. On the analysis of the  $(1+1)$  evolutionary algorithm. *Theoretical Computer Science* 276 (2002), 51–81.
- [18] Stefan Droste, Thomas Jansen, and Ingo Wegener. 2006. Upper and lower bounds for randomized search heuristics in black-box optimization. *Theory of Computing Systems* 39 (2006), 525–544.
- [19] Tobias Friedrich, Timo Kötzing, Martin S. Krejca, Samadhi Nallaperuma, Frank Neumann, and Martin Schirneck. 2016. Fast building block assembly by majority vote crossover. In *Genetic and Evolutionary Computation Conference, GECCO 2016*. ACM, 661–668.
- [20] Tobias Friedrich, Timo Kötzing, Martin S. Krejca, and Andrew M. Sutton. 2017. The compact genetic algorithm is efficient under extreme Gaussian noise. *IEEE Transactions on Evolutionary Computation* 21 (2017), 477–490.
- [21] Christian Gießen and Carsten Witt. 2017. The interplay of population size and mutation probability in the  $(1 + \lambda)$  EA on OneMax. *Algorithmica* 78 (2017), 587–609.
- [22] Georges R. Harik, Fernando G. Lobo, and David E. Goldberg. 1999. The compact genetic algorithm. *IEEE Transactions on Evolutionary Computation* 3 (1999), 287–297.
- [23] Václav Hasenöhl and Andrew M. Sutton. 2018. On the runtime dynamics of the compact genetic algorithm on jump functions. In *Genetic and Evolutionary Computation Conference, GECCO 2018*. ACM, 967–974.
- [24] Jun He and Xin Yao. 2001. Drift analysis and average time complexity of evolutionary algorithms. *Artificial Intelligence* 127 (2001), 51–81.
- [25] Thomas Jansen. 2013. *Analyzing Evolutionary Algorithms - The Computer Science Perspective*. Springer. 1–236 pages.
- [26] Thomas Jansen, Kenneth A. De Jong, and Ingo Wegener. 2005. On the choice of the offspring population size in evolutionary algorithms. *Evolutionary Computation* 13 (2005), 413–440.
- [27] Thomas Jansen and Ingo Wegener. 2002. The analysis of evolutionary algorithms—a proof that crossover really can help. *Algorithmica* 34 (2002), 47–66.
- [28] Martin S. Krejca and Carsten Witt. 2018. Theory of Estimation-of-Distribution Algorithms. *CoRR* abs/1806.05392 (2018). arXiv:1806.05392
- [29] Martin S. Krejca and Carsten Witt. 2019. Lower bounds on the run time of the Univariate Marginal Distribution Algorithm on OneMax. *Theoretical Computer Science* (2019). <https://doi.org/10.1016/j.tcs.2018.06.004> To appear.
- [30] Per Kristian Lehre and Phan Trung Hai Nguyen. 2017. Improved runtime bounds for the univariate marginal distribution algorithm via anti-concentration. In *Genetic and Evolutionary Computation Conference, GECCO 2017*. ACM, 1383–1390.
- [31] Johannes Lengler. 2017. Drift Analysis. *CoRR* abs/1712.00964 (2017). arXiv:1712.00964
- [32] Johannes Lengler, Dirk Sudholt, and Carsten Witt. 2018. Medium step sizes are harmful for the compact genetic algorithm. In *Genetic and Evolutionary Computation Conference, GECCO 2018*. ACM, 1499–1506.
- [33] Heinz Mühlenbein and Gerhard Paass. 1996. From recombination of genes to the estimation of distributions I. Binary parameters. In *Parallel Problem Solving from Nature, PPSN 1996*. Springer, 178–187.
- [34] Frank Neumann and Carsten Witt. 2010. *Bioinspired Computation in Combinatorial Optimization – Algorithms and Their Computational Complexity*. Springer.
- [35] Pietro Simone Oliveto and Carsten Witt. 2012. Erratum: Simplified Drift Analysis for Proving Lower Bounds in Evolutionary Computation. *CoRR* abs/1211.7184 (2012).
- [36] Dirk Sudholt. 2013. A new method for lower bounds on the running time of evolutionary algorithms. *IEEE Transactions on Evolutionary Computation* 17 (2013), 418–435.
- [37] Dirk Sudholt and Carsten Witt. 2019. On the choice of the update strength in estimation-of-distribution algorithms and ant colony optimization. *Algorithmica* 81 (2019), 1450–1489.
- [38] Darrell Whitley, Swetha Varadarajan, Rachel Hirsch, and Anirban Mukhopadhyay. 2018. Exploration and exploitation without mutation: solving the jump function in  $\Theta(n)$  time. In *Parallel Problem Solving from Nature, PPSN 2018, Part II*. Springer, 55–66.
- [39] Carsten Witt. 2013. Tight bounds on the optimization time of a randomized search heuristic on linear functions. *Combinatorics, Probability & Computing* 22 (2013), 294–318.
- [40] Carsten Witt. 2017. Upper bounds on the runtime of the univariate marginal distribution algorithm on OneMax. In *Genetic and Evolutionary Computation Conference, GECCO 2017*. ACM, 1415–1422.
- [41] Carsten Witt. 2018. Domino convergence: why one should hill-climb on linear functions. In *Genetic and Evolutionary Computation Conference, GECCO 2018*. ACM, 1539–1546.