The Compact Genetic Algorithm Struggles on Cliff Functions

Frank Neumann
Optimisation and Logistics
School of Computer Science
University of Adelaide
Adelaide, Australia

Dirk Sudholt
Faculty of Computer Science and Mathematics
University of Passau
Passau, Germany

Carsten Witt
DTU Compute
Technical University of Denmark
Kongens Lyngby, Denmark

ABSTRACT

The compact genetic algorithm (cGA) is a non-elitist estimation of distribution algorithm which has shown to be able to deal with difficult multimodal fitness landscapes that are hard to solve by elitist algorithms. In this paper, we investigate the cGA on the Cliff function for which it has been shown recently that non-elitist evolutionary algorithms and artificial immune systems optimize it in expected polynomial time. We point out that the cGA faces major difficulties when solving the Cliff function and investigate its dynamics both experimentally and theoretically around the Cliff. Our experimental results indicate that the cGA requires exponential time for all values of the update strength K. We show theoretically that, under sensible assumptions, there is a negative drift when sampling around the location of the cliff. Experiments further suggest that there is a phase transition for K where the expected optimization time drops from $n^{\Theta(n)}$ to $2^{\Theta(n)}$.

CCS CONCEPTS

• **Theory of computation → Theory of randomized search heuristics**.

KEYWORDS

Estimation-of-distribution algorithms, compact genetic algorithm, evolutionary algorithms, running time analysis, theory.

ACM Reference Format:
Frank Neumann, Dirk Sudholt, and Carsten Witt. 2022. The Compact Genetic Algorithm Struggles on Cliff Functions. In *Genetic and Evolutionary Computation Conference (GECCO '22), July 9–13, 2022, Boston, MA, USA.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3512290.3528776

1 INTRODUCTION

Runtime analysis of evolutionary algorithms and other randomized search heuristics has provided a deep understanding of many working principles of these algorithms [5, 15]. The goal of these studies is to provide rigorous results of randomized search heuristics by analyzing them as a special class of randomized algorithms. This allows to use a wide range of tools such as concentration bounds and random walk arguments. A wide range of new methods for analyzing randomized search heuristics have been developed over the last 20 years. Starting with methods such as fitness based partitions for simple problems and elitist algorithms, more complex combinatorial optimization problem [22] (including NP-hard ones) and non elitist algorithms have been investigated.

Estimation of distribution algorithms (EDAs) [24] are a special class of randomized search heuristics that work with a probability distribution at each stage of the algorithm (instead of a set of solutions). This probability distribution is updated by reinforcing components that have shown to lead to solutions of good quality. EDAs have found a wide range of applications to problems such as military antenna design, multiobjective knapsack, and quadratic assignment (see [11] for an introduction and overview).

The theoretical runtime analysis concentrates on simple EDAs that capture their basic algorithmic properties [17]. The compact genetic algorithm (cGA) is such a simple EDA which has been studied in different runtime analyses. Following the seminal work by Droste [7] for the cGA in the mid 2000s, there has been a growing interest in studying the cGA and other EDAs over the last 8 years [2, 18, 27]. We refer the reader to Krejca and Witt [17] for a recent survey. These theoretical results focus on the working principles of the considered EDAs and especially discuss their difference to simple evolutionary algorithms such as the (1+1) EA. Several studies have shown that the update strength K, which determines the magnitude of changes to the probabilistic model, has a crucial impact on performance [6, 20, 26]. In [20] it was shown that the cGA optimizes OneMax efficiently, in expected time $O(\sqrt{n}K)$, if the update strength is sufficiently large, i. e. $K = \Omega(\sqrt{n}\log n)$. For $K = \Theta(\sqrt{n}\log n)$ this yields an upper bound of $O(n \log n)$ function evaluations. In [19, 20], the authors showed that for smaller values of K in $\Omega(\log^3 n)$ and $O(\sqrt{n}/(\log(n)\log\log n))$, the expected optimization time on OneMax is $\Omega(K^{1/3}n)$ in expectation and with high probability. Thus, in this so-called medium parameter regime the expected optimization time increases with K before dropping down to $O(n \log n)$ for $K \geq \Omega(\sqrt{n}\log n)$.

Other studies have unveiled remarkable advantages of EDAs. Their ability to learn good solution components, coupled with a slow adaptation of the probabilistic model, makes EDAs higly robust with respect to noisy fitness evaluations [9]. Furthermore, their ability to sample with a large sampling variance implies that they are good at exploring the search space. This has been shown rigorously for the Jump function, a multimodal function of unitation (i. e. the fitness only depends on the number of ones) where evolutionary algorithms typically need to make a large jump. With the right choice of the update strength, the cGA is able to optimize Jump efficiently, if the size of the jump is not too large [4, 10, 28].

In this work we consider the runtime of the cGA on a multimodal function. Cliff is a function of unitation with the difficulty that

GECCO '22, July 9–13, 2022, Boston, MA, USA

ACM ISBN 978-1-4503-9237-2/22/07…$15.00
https://doi.org/10.1145/3512290.3528776

inferior solutions need to be accepted in order to advance towards the global optimum (unless the algorithm jumps to the optimum directly). In other words, algorithms need to be able to "jump down" a cliff in the fitness landscape (see Section 2 for a definition and Figure 1 for an illustration of Cliff). It was originally proposed by Jägersküpper and Storch [14] to show the advantages of non-elitism in evolutionary algorithms. They showed that a simple (1,λ) EA that generates λ offspring independently and picks the best offspring to replace the parent optimizes Cliff in expected $O(n^{25})$ evaluations. Hevia Fajardo and Sudholt [12] showed that this time is in fact in $O(n^\eta \log^2 n)$ and $\omega(n^{\eta-\varepsilon})$ for every constant $\varepsilon > 0$, where $\eta \approx 3.976770136$.

The same paper [12] also showed that a (1,λ) EA with a self-adjusting offspring population size λ can optimize Cliff in expected $O(n)$ generations and $O(n \log n)$ expected function evaluations. The same time bound $O(n \log n)$ was shown earlier for other non-elitist algorithms: hyperheuristics that have a certain probability of accepting every offspring [21] and for evolutionary algorithms using ageing [1].

The Cliff function has a similar structure to Jump with a jump length of $n/3$ and the sets of local optima are identical for both functions. However, when overcoming those local optima, Cliff shows a gradient pointing towards the global optimum whereas Jump has a gradient leading back towards local optima. The gradient structure for Cliff is hence more benign than that for Jump.

Based on the aforementioned positive results for non-elitist evolutionary algorithms on Cliff, and the positive results for the cGA on Jump, one might expect that the non-elitist cGA is also effective on Cliff, if the update strength is chosen just right.

The main contribution of this paper is to show, theoretically and empirically, that this is not the case. In particular, the cGA does not seem to benefit much from the benign gradients past the set of local optima, that is, past the top of the cliff.

By examining the behavior of the cGA when sampling around the cliff, we show in Section 3 that, under some conditions, the probabilistic model experiences a negative drift and tends to move away from the optimal distribution. This happens when the cGA tends to sample one offspring at the top of the cliff and one offspring at the bottom of the cliff, and the former offspring is reinforced. This negative drift prevents the probabilistic model to overcome the region around the cliff, leading to exponential times.

Our negative drift bound uses novel arguments for the analysis of the cGA by approximating conditional sampling distributions in the cGA, conditional on whether the offspring lie on the same side or on different sides of the cliff, by truncated normal distributions. However, this novel approach is not fully rigorous as it is based on the assumption that the sampling variance is always super-constant.

We conjecture that the variance typically stabilizes to super-constant values and continue to prove exponential lower bounds on the expected optimization time in Section 5 under conjectured lower and upper bounds on the variance. We justify our conjecture in Section 4 by reviewing related work by Lengler et al. [20] on OneMax, where such variance bounds were proven rigorously, and explain which parts of their analysis can be translated to Cliff and where this approach breaks down. We instead present empirical data on the sampling variance to support our conjecture.

In Section 6 we provide experiments on the runtime of the cGA on Cliff. The parameter landscape for the update strength K shows a highly complex behavior. Our data suggests that the expected optimization time slowly increases from $2^{\Theta(n)}$ to $n^{\Theta(n)}$ as K grows, before dropping sharply to $2^{\Theta(n)}$ again. We give possible explanations for these effects and finish with a list of open problems.

2 PRELIMINARIES

The cGA is defined in Algorithm 1. It uses a univariate probabilistic model of *frequencies* $p_{t,1}, p_{t,2}, \dots, p_{t,n} \in [0,1]$, which is used to sample new search points. The i-th frequency $p_{t,i}$ represents the probability of setting the i-th bit to 1 in iteration t. In every iteration, the cGA samples two search points x and y in this way. We shall refer to these as *offspring*, using the language of evolutionary computation. It then sorts x and y such that $f(x) \geq f(y)$ and reinforces x in the probabilistic model. This is done by inspecting the bits at position i and increasing $p_{t,i}$ if $x_i = 1$ and $y_i = 0$ and decreasing $p_{t,i}$ if $x_i = 0$ and $y_i = 1$. The aim is to increase the likelihood of sampling the bit value of the better offspring in the future. If both offspring have the same bit value, the frequency $p_{t,i}$ is unchanged. Frequencies are changed by $\pm 1/K$ and K is called the *update strength* of the cGA. Small values of K imply large values of $1/K$ and hence large changes. This means that novel information has a large impact on the probabilistic model. Large values of K imply small changes to the probabilistic model, such that the probabilistic model is adapted gradually, and information from many past samples is stored in the frequencies.

Frequencies are always capped to the interval $[1/n, 1-1/n]$ such that the probability of sampling any particular search point is always at least $(1/n)^n > 0$. We refer to $1/n$ as the *lower border* and to $1-1/n$ as the *upper border*. Throughout the paper we tacitly assume that K is in the set $\mathcal{K} := \{i(1/2 - 1/n) \mid i \in \mathbb{N}\}$ so that the state space of frequencies is restricted to $p_{i,t} \in \{1/n, 1/n + 1/K, \dots, 1/2, \dots, 1 - 1/n - 1/K, 1 - 1/n\}$.

As common in theoretical runtime analysis, we define the *optimization time* as the number of function evaluations required to sample a global optimum for the first time. Since the cGA makes two evaluations in every iteration, the optimization time is twice the number of iterations needed to sample a global optimum.

Algorithm 1: Compact Genetic Algorithm (cGA)

```
t ← 0;
p_{t,1} ← p_{t,2} ← ··· ← p_{t,n} ← 1/2;
while termination criterion not met do
    for i ∈ {1,...,n} do
        x_i ← 1 with prob. p_{t,i}, x_i ← 0 with prob. 1 − p_{t,i};
    for i ∈ {1,...,n} do
        y_i ← 1 with prob. p_{t,i}, y_i ← 0 with prob. 1 − p_{t,i};
    if f(x) < f(y) then swap x and y;
    for i ∈ {1,...,n} do
        if x_i > y_i then p_{t+1,i} ← p_{t,i} + 1/K;
        if x_i < y_i then p_{t+1,i} ← p_{t,i} − 1/K;
        if x_i = y_i then p_{t+1,i} ← p_{t,i};
        p_{t+1,i} ← max{min{p_{t+1,i}, 1 − 1/n}, 1/n};
    t ← t + 1;
```

The function Cliff is a function of unitation, that is, it only depends on the number of ones in a bit string x, denoted as $|x|_1$. Then Cliff is defined as:

$$\textsc{Cliff}(x) := \begin{cases} |x|_1 & \text{if } |x|_1 \leq 2n/3 \\ |x|_1 - n/3 + 1/2 & \text{otherwise.} \end{cases}$$

See Figure 1 for an illustration. We refer to the region of search points with at most $2n/3$ ones as the *first slope*, and all remaining search points as the *second slope*. The only global optimum is the all-ones string 1^n with a fitness of $2n/3 + 1/2$. All search points with $2n/3$ ones are local optima at the top of the cliff. Note that all search points on the second slope are strictly worse than all search points at the top of the cliff, except for the global optimum.

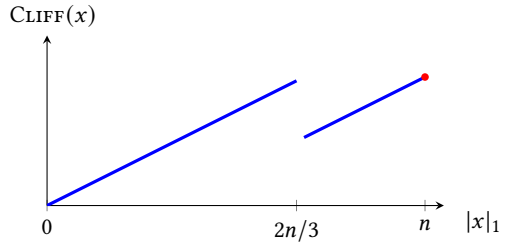

Figure 1: Illustration of Cliff

When analyzing the cGA on functions of unitation (e. g., OneMax as analyzed in [20, 26] and Jump as analyzed in [4, 28]), one is interested in the number of one-bits sampled in an offspring. This random value follows a Poisson-binomial distribution with the frequencies $(p_{t,1}, \dots, p_{t,n})$ as underlying success probabilities. In particular, the following two quantities play a key role in bounding the progress of the cGA towards the optimum:

(1) the *potential* $P_t := \sum_{i=1}^n p_{t,i}$ equals the expected value of the Poisson-binomial distribution, i. e., the expected number of one-bits sampled in an offspring,
(2) the *sampling variance* $V_t := \sum_{i=1}^n p_{t,i}(1-p_{t,i})$ is the variance in the number of one-bits.

The following negative drift theorem will be used in Section 5 to analyze the one-step change of potential $\Delta_t := P_{t+1} - P_t$.

Theorem 2.1 (Negative Drift with Scaling, cf. [23]). *Let $(X_t)_{t\geq 0}$ be a stochastic process, adapted to a filtration $\mathcal{F}_t$, over some state space $S \subseteq \mathbb{R}$. Suppose there exist an interval $[a, b] \subseteq \mathbb{R}$ and, possibly depending on $\ell := b - a$, a drift bound $\varepsilon := \varepsilon(\ell) > 0$ as well as a scaling factor $r := r(\ell) > 0$ such that for all $t \geq 0$ the following three conditions hold:*

(1) $\mathrm{E}(X_{t+1} - X_t \mid \mathcal{F}_t\,;\, a < X_t < b) \geq \varepsilon$,
(2) $\Pr(|X_{t+1} - X_t| \geq jr \mid \mathcal{F}_t\,;\, a < X_t) \leq e^{-j}$ *for* $j \in \mathbb{N}_0$,
(3) $1 \leq r^2 \leq \varepsilon\ell/(132 \log(r/\varepsilon))$.

Then for the first hitting time $T^* := \min\{t \geq 0 \colon X_t \leq a \mid X_0 \geq b\}$ it holds that $\Pr\left(T^* \leq e^{\varepsilon\ell/(132r^2)} \mid \mathcal{F}_0\right) = O(e^{-\varepsilon\ell/(132r^2)})$.

To verify the second condition of the negative drift theorem in our concrete analysis, we will use the following lemma dealing with Chernoff-type bounds depending on the variance. The lemma goes back to [13]. We present a version given in [3, Theorem 1.10.14].

Lemma 2.2. *Let $X_1, \dots, X_n$ be independent random variables. Let b be such that $X_i \leq \mathrm{E}(X_i) + b$ for all $i = 1, \dots, n$. Let $X = \sum_{i=1}^n X_i$. Let $\sigma^2 = \sum_{i=1}^n \mathrm{Var}(X_i) = \mathrm{Var}(X)$. Then, for all $\lambda \geq 0$,*

$$\Pr(X \geq \mathrm{E}(X) + \lambda) \leq e^{-(1/3)\min\{\lambda^2/\sigma^2, \lambda/b\}}.$$

As a simple consequence, we obtain the following corollary:

Corollary 2.3. *Consider the cGA on an arbitrary fitness function Then for all $t \geq 0$ and $\lambda > 0$ it holds that*

$$\Pr(|P_{t+1} - P_t| \geq \lambda/K) \leq 2e^{-(1/3)\min\{\lambda^2/V_t, \lambda\}}.$$

To see that the corollary follows, we argue in the same way as in [28], where jump functions were considered: the absolute value of the one-step change in potential is no larger than the absolute difference in the number of one-bits of the two individuals sampled, scaled down by $1/K$. This holds since each bit sampled 1 in the fitter offspring and 0 in the other offspring contributes a $+1/K$ to the change of potential (or nothing, in case the frequency is capped at the upper border) and no less than $-1/K$ in the opposite case. The factor 2 accounts for the two possible orderings of offspring.

3 NEGATIVE DRIFT AROUND THE CLIFF

We will under certain assumptions prove that the potential of the cGA cannot overcome the cliff region efficiently since there is a negative drift in the potential. The intuition is as follows:

The initial potential is $n/2$ and, as long as the potential is significantly less than $2n/3$, the cGA is very unlikely to sample search points on the second slope of Cliff. If that does not happen, the fitness landscape is the same as on OneMax. Hence, using the results from [26], the potential P_t will steadily increase towards the location of the cliff, i. e. $2n/3$.

However, when the potential P_t has increased to roughly $2n/3$, i. e., the expected number of ones sampled is close to the cliff, it is relatively likely that the cGA samples search points on both the first slope and the second slope of cliff. In particular, if the sampling variance V_t is large and $P_t = 2n/3$, the sampling distribution is similar to a normal distribution with mean $2n/3$ and the given variance. Hence, we are confronted with an approximately symmetric distribution. Then the probability of sampling the two offspring on both sides of the cliff becomes roughly $(1/2) \cdot (1/2) + (1/2) \cdot (1/2) = 1/2$ by counting the two opposite events of sampling the first offspring on the first slope and the second one on the second slope and vice versa. By a similar argumentation, also the probability of sampling both offspring on the same slope will approach $1/4 + 1/4 = 1/2$.

We will analyze the drift, i. e., expected change of potential, under event M of sampling on two different slopes and its complement. The key observation is that under M, the offspring with the smaller number of one-bits will have roughly $2n/3 - \sqrt{V_t}$ one-bits in expectation and the other offspring will have roughly $2n/3 + \sqrt{V_t}$ one-bits in expectation by properties of truncated normal distributions that arise under M. Since the offspring on the first slope will be fitter and reinforced in the frequency update, this corresponds to an expected decrease in potential of $(2n/3 - \sqrt{V_t} - (2n/3 + \sqrt{V_t}))/K = -2\sqrt{V_t}/K$.

Under $\overline{M}$, both offspring are on the same slope and their expected difference in one-bits is no larger than the variance $\sqrt{V_t}$, again by simple analyses of truncated normal distributions. Taking these two cases of roughly identical probability together, the total drift

becomes $(1/2)(-2\sqrt{V_t}+\sqrt{V_t})/K = -\sqrt{V_t}/(2K)$. This argumentation can be made rigorous not only when $P_t = 2n/3$, but for roughly all $P_t \in [2n/3 - \sqrt{V_t}, 2n/3 + \sqrt{V_t}]$, as the following lemma shows. We will use this result when applying a negative drift theorem (Theorem 2.1) in Section 5.

Theorem 3.1. *Assume $V_t = \omega(1)$. Let $\varepsilon > 0$ be an arbitrary constant. Then conditioned on $P_t \in [2n/3 - (\alpha(n))^{1/2-\varepsilon}, 2n/3]$, it holds that $\mathrm{E}(\Delta_t \mid P_t) = -\Omega(\sqrt{V_t}/K)$.*

Before we proceed with the proof, we collect well-known properties of the expected value $\mathrm{E}(X \mid X \le t)$ of a truncated normal distribution and show that $t - \mathrm{E}(X \mid X \le t)$, i. e., the distance of this expected value from the truncation parameter t, increases when the truncation condition becomes weaker, i. e., when t grows.

Lemma 3.2. *Given a normally distributed random variable X with mean μ and variance σ, we have for all $t \in \mathbb{R}$ that*

$$\mathrm{E}(X \mid X \le t) = \mu - \sigma\frac{\phi((t-\mu)/\sigma)}{\Phi((t-\mu)/\sigma)}$$

$$\text{and} \quad \mathrm{E}(X \mid X \ge t) = \mu + \sigma\frac{\phi((t-\mu)/\sigma)}{\Phi((t-\mu)/\sigma)},$$

where ϕ and Φ denote the density and cumulative distribution function of the standard normal distribution, respectively. Moreover, the function $t - \mathrm{E}(X \mid X \le t)$ is monotone increasing in t.

Proof. The first two claims relate to the expected value of the so-called truncated normal distribution and are well known in the literature (e. g., p. 156 in [16]).

For the final claim, we consider w. l. o. g. a standard normally distributed random variable Z and write for arbitrary $x \in \mathbb{R}$

$$x - \mathrm{E}(Z \mid Z \le x) = x + \frac{\phi(x)}{\Phi(x)}. \tag{1}$$

The function $\frac{\phi(x)}{\Phi(x)}$ is known as the *inverse Mills ratio* in the literature and known to have a derivative of at least -1 (see [25], who shows that the derivative of $\frac{\phi(x)}{1-\Phi(x)} = \frac{\phi(-x)}{\Phi(-x)}$ is at most 1). Hence, the derivative of (1) is at least 0 and the final claim follows. □

Proof of Theorem 3.1. By assumption, we have that $V_t = \omega(1)$ for all $t \ge 0$. Hence, by the generalized central limit theorem (Chapter XV.6 in [8]) the number of one-bits sampled in each offspring, which follows a Poisson-binomial distribution with mean P_t and variance V_t, converges in distribution to a normal distribution with mean P_t and variance V_t. More precisely, let $X = |x|_1$ for an arbitrary offspring sampled with current frequency vector of potential P_t and variance V_t and let $X' \sim N(P_t, \sqrt{V_t})$. Then for all $t \in \mathbb{R}$, $\Pr(X \le t) = (1 \pm o(1))\Pr(X' \le t)$. Often, we will pretend that $X \sim N(P_t, \sqrt{V_t})$ and omit $1 - o(1)$ factors stemming from the normal approximation.

We will decompose the drift according to three events for the location of the two offspring of the cGA:

L) Both offspring have at most $2n/3$ one-bits, i. e., lie both on the first (**l**eft) slope.
R) Both offspring have at least $2n/3 + 1$ one-bits, i. e., lie both on the second (**r**ight) slope.
M) One offspring has at most $2n/3$ one-bits and one at least $2n/3 + 1$ one-bits, i. e., there is an offspring on each slope (the **m**ixed case).

Obviously, by the law of total probability,

$$\mathrm{E}(\Delta_t \mid P_t) = \mathrm{E}(\Delta_t \mid P_t; L)\Pr(L \mid P_t) + \mathrm{E}(\Delta_t \mid P_t; R)\Pr(R \mid P_t) + \mathrm{E}(\Delta_t \mid P_t; M)\Pr(M \mid P_t).$$

For readability, we may omit the condition on the random P_t in the following. Let $p_R := \Pr(X > 2n/3 \mid P_t)$, then $\Pr(R) = p_R^2$, $\Pr(L) = (1-p_R)^2$ and $\Pr(M) = 2p_R(1-p_R)$. Hence,

$$\mathrm{E}(\Delta_t \mid P_t) = \mathrm{E}(\Delta_t \mid L)(1-p_R)^2 + \mathrm{E}(\Delta_t \mid R)p_R^2 + \mathrm{E}(\Delta_t \mid M)2p_R(1-p_R). \tag{2}$$

Let us consider the generation of one offspring more closely, assuming a fixed P_t. A crucial insight, implied by the normal approximation, is that $p_R = \Pr(X > 2n/3)$ is monotone increasing in P_t (up to multiplicative errors of $1 - o(1)$) and approaches $1/2$. Even more, already if $P_t = 2n/3 - (V_t)^{1/2-\varepsilon}$ for some constant $\varepsilon > 0$, the probability $\Pr(X > 2n/3)$ becomes at least $1/2 - o(1)$ using the normal approximation. This follows since the density is at most $\frac{e^{-1/2}}{\sqrt{V_t}\sqrt{2\pi}} = O(1/\sqrt{V_t})$ so that $\Pr\left(2n/3 - (V_t)^{1/2-\varepsilon} \le X \le 2n/3\right) \le V_t^{1/2-\varepsilon} \cdot O(1/\sqrt{V_t}) = o(1)$.

Let us now fix $c > 0$ such that $P_t \in [2n/3 - c(V_t)^{1/2-\varepsilon}, 2n/3]$ and $p_R = \Pr(X > 2n/3) \ge 1/2 - 1/(V_t)^{1/2-\varepsilon} = 1/2 - o(1)$. Since $P_t \le 2n/3$, we also have $p_R \le 1/2$ and therefore

- $\Pr(R) = p_R^2 \le 1/4$
- $\Pr(M) = 2p_R(1-p_R) \ge 2(1/2 - o(1))(1/2) = 1/2 - o(1)$
- $\Pr(L) = (1-p_R)^2 \le (1/2 + o(1))^2 = 1/4 + o(1)$.

We next estimate the drift under the three events. To this end, we need bounds on the two conditional expectations $\mathrm{E}(X \mid X \le 2n/3)$ and $\mathrm{E}(X \mid X \ge 2n/3 + 1)$ since the conditions specify that an offspring is on the first and second slope, respectively. Using Lemma 3.2 with $\mu = P_t$, $\sigma = V_t$ and $t = 2n/3$, we have

$$\mathrm{E}(X \mid X \le 2n/3) = P_t - \sqrt{V_t} \cdot \frac{\phi((2n/3 - P_t)/\sqrt{V_t})}{\Phi((2n/3 - P_t)/\sqrt{V_t})}.$$

We note that $(2n/3 - P_t)/\sqrt{V_t} = O(1/V_t^{\varepsilon}) = o(1)$ by our choice of V_t. Hence, we have

$$\frac{\phi((2n/3 - P_t)/\sqrt{V_t})}{\Phi((2n/3 - P_t)/\sqrt{V_t})} = \frac{\phi(o(1))}{\Phi(o(1))} = \frac{(1 \pm o(1))\phi(0)}{(1 \pm o(1))\Phi(0)} = (1 \pm o(1))\sqrt{\frac{2}{\pi}}, \tag{3}$$

using the continuity of the density and distribution functions in the second step and the well-known equality $\frac{\phi(0)}{\Phi(0)} = \sqrt{2/\pi}$ stemming from the half-normal distribution in the third step. Together,

$$\mathrm{E}(X \mid X \le 2n/3) = P_t - (1 + o(1))\sqrt{2/\pi}\sqrt{V_t}.$$

In the very same way, we derive

$$\mathrm{E}(X \mid X > 2n/3) = P_t - (1 - o(1))\sqrt{2/\pi}\sqrt{V_t}.$$

Under the event M defined above, we have one offspring with at most $2n/3$ one-bits and another one with strictly more one bits. The update will reinforce the individual on the left slope and change the potential by the difference in the number of one-bits, divided

by K, assuming no frequencies at the border. To correct this for the boundary effects, we apply Lemma 8 in [4] and obtain an error term of at most $2/K$ in the expected change of potential. (Roughly speaking, this accounts for the fact that every frequency at the border flips with probability at most $2(1/n)(1-1/n)$ and that capping reduces its change by at most $2/K$.) Hence, we obtain

$$\begin{aligned}\mathrm{E}(\Delta_t \mid M) &\leq -\frac{1}{K}\big(\mathrm{E}(X \mid X > 2n/3) - \mathrm{E}(X \mid X \leq 2n/3)\big) + \frac{2}{K}\\ &= -\frac{1}{K}\left(\left(P_t + (1-o(1))\sqrt{\frac{2}{\pi}V_t}\right) - \left(P_t - (1+o(1))\sqrt{\frac{2}{\pi}V_t}\right)\right) + \frac{2}{K}\\ &= -(2-o(1))\frac{1}{K}\sqrt{\frac{2}{\pi}V_t},\end{aligned}$$

where we have used that $V_t = \omega(1)$.

We are left with the drift under L and R; we only analyze L since both cases are analogous. Here we sample two offspring conditional on both having at most $2n/3$ one-bits. Let X_1 and X_2 denote the random number of one-bits of two offspring and assume w. l. o. g. that $X_1 \leq X_2$. Similarly as for $\mathrm{E}(\Delta_t \mid M)$, the potential drift is then

$$\mathrm{E}(\Delta_t \mid L) \leq \frac{1}{K} \cdot \mathrm{E}(X_2 - X_1 \mid X_1 \leq X_2 \leq 2n/3) + \frac{2}{K}.$$

Compared to the case M analyzed above, the difference $X_2 - X_1$ in the number of one-bits tends to be smaller since the two offspring are sampled on the same slope, whereas under M the offspring are on different slopes and $X_2 - X_1$ is typically larger. Using that X_1 is normally distributed with variance V_t, Lemma 3.2 implies

$$\mathrm{E}(X_1 \mid X_1 \leq s) = P_t - \sqrt{V_t} \cdot \frac{\phi((s-P_t)/\sqrt{V_t})}{\Phi((s-P_t)/\sqrt{V_t})},$$

where we identify $s = X_2$, assuming $X_2 \leq 2n/3$. Hence,

$$\mathrm{E}(s - X_1 \mid s; X_1 \leq s) = s - P_t + \sqrt{V_t} \cdot \frac{\phi((s-P_t)/\sqrt{V_t})}{\Phi((s-P_t)/\sqrt{V_t})}$$

As shown in Lemma 3.2, the right-hand side of the last equation is monotone increasing in s. Hence,

$$\begin{aligned}&\mathrm{E}(X_2 - X_1 \mid X_1 \leq X_2 \leq 2n/3) \leq \mathrm{E}(2n/3 - X_1 \mid X_1 \leq 2n/3)\\ &\quad = 2n/3 - P_t + \sqrt{V_t} \cdot \frac{\phi((2n/3-P_t)/\sqrt{V_t})}{\Phi((2n/3-P_t)/\sqrt{V_t})}\end{aligned}$$

(simplifying the fraction using (3))

$$\leq 2n/3 - P_t + (1+o(1))\sqrt{(2/\pi)V_t} = (1+o(1))\sqrt{(2/\pi)V_t}$$

so, since $V_t = \omega(1)$, we have both

$$\mathrm{E}(\Delta_t \mid L) \leq \frac{(1+o(1))}{K}\sqrt{\frac{2}{\pi}V_t} \text{ and } \mathrm{E}(\Delta_t \mid R) \leq \frac{(1+o(1))}{K}\sqrt{\frac{2}{\pi}V_t}.$$

Plugging the above bounds in (2), we obtain

$$\begin{aligned}\mathrm{E}(\Delta_t \mid P_t) &\leq \frac{1}{K}\Big((p_R^2 + (1-p_R)^2)\cdot(1+o(1))\sqrt{(2/\pi)V_t}\\ &\qquad + 2p_R(1-p_R)(-(2-o(1))\sqrt{(2/\pi)V_t})\Big)\\ &= \frac{1}{K}\left((1/2+o(1))\sqrt{(2/\pi)V_t} - (1/2-o(1))\cdot 2\sqrt{(2/\pi)V_t}\right)\\ &= -(1/2-o(1))\frac{1}{K}\sqrt{(2/\pi)V_t}.\end{aligned}$$

□

4 JUSTIFYING THE ASSUMPTION OF SUPER-CONSTANT SAMPLING VARIANCE

The approximation with (truncated) normal distributions used in the proof of the drift estimate from Theorem 3.1 hinges on the sampling variance being $\omega(1)$. We now try to convince the reader why we believe that the sampling variance is $\omega(1)$, for interesting K.

4.1 Rigorous Variance Bounds for OneMax

To this end, we first discuss the variance on OneMax, for which rigorous bounds of $\omega(1)$ have been shown by Lengler et al. [20] in a medium parameter regime for K (as will be defined below). The following statement was implicitly shown in [20] and can be deduced from [20, Lemma 18] and its proof. A discussion will follow.

Theorem 4.1. *Consider the cGA on OneMax with $K = \Omega(\log^3 n)$ and $K = O(n^{1/2}/(\log(n)\log\log(n)))$. With probability $1-e^{-\Omega(K^{1/4})}$, there exist times $t_1 = O(K^2\log^2 n)$, $t_2 = O(K^2\log^2(n)\log\log K)$ and $t_3 > t_2$ such that the following statements hold.*

(1) For all $t \in [t_1, t_2]$, $V_t \geq \Omega(K^{1/2})$.
(2) The number of frequencies at the lower border at time t_2 is $\Omega(n)$. For all $t_3 > t_2$ such that there are $\Omega(n)$ frequencies at the lower border at all times in (t_2, t_3), with probability $1 - O(t_3/(K^2\log^2 n))\cdot\exp(-\Omega(K^{1/3}))$,

$$V_t \in \Omega(K^{2/3}) \cap O(K^{4/3}).$$

There is an initial phase of the first $O(K^2\log^2 n) = o(n)$ iterations for which no lower bound on the variance is shown in [20]. After this phase, we have a lower bound of $\Omega(K^{1/2})$ on the variance that quickly improves to a lower bound of $\Omega(K^{2/3})$. The latter bound applies with good probability as long as there are still $\Omega(n)$ frequencies at the lower border.

We describe the main idea behind the analysis in [20], and the proof of Theorem 4.1. First we observe that a frequency at a border contributes only $1/n \cdot (1-1/n)$ to the variance, while frequencies that are off their borders contribute a much larger amount. Hence bounding the variance is achieved by studying the number and position of off-border frequencies.

Lemma 18 in [20] considers the situation after the first $O(K^2)$ iterations, when a linear number of frequencies has reached the lower border, with probability $1 - e^{-\Omega(K^{1/2})}$. Then the authors consider periods of $\Theta(K^2\log^2 n)$ iterations and show that in a period, frequencies tend to leave their borders to perform a random walk. This random walk ends when a border is reached. (The frequency may then start another random walk during the period.) Frequencies that perform a random walk contribute a term of $p_{i,t}(1-p_{i,t})$ to the sampling variance. Hence the variance in future iterations can be bounded by analyzing these random walks. The dynamics are intricate since the random walks show a positive drift that depends on the current sampling variance. The drift has a potentially significant impact on the random walks; for instance, it can decide whether a random walk started at the lower border crosses the whole range $[1/n, 1-1/n]$ and ends up at the upper border, or whether it returns to the lower border. Lengler et al. [20] argue that the cGA experiences a feedback loop since the current sampling variance influences future sampling variances. This feedback loop

has a considerable lag as the effects of a small or large sampling variance are felt at later stages of the frequencies' random walks.

One idea from [20] is to assume that we have lower and upper bounds on the sampling variance during a period as this can then be used to bound the drift for the frequencies' random walks from above and below, and to establish bounds for the sampling variance in the next period. This is formalized in the so-called stabilization lemma, Lemma 7 in [20], in which lower and upper bounds on the variance in one period are used to show tighter lower and upper bounds in the next period. Part (a) of Lemma 7 in [20] assumes trivial bounds on the sampling variance and applies after the short, initial phase of $O(K^2)$ steps, when $\Theta(n)$ frequencies have reached the lower border. Part (a) of the stabilization lemma then yields that after a further period of $CK^2 \log^2 n$ iterations, $C > 0$ a sufficiently large constant, the variance is guaranteed to be at least $\Omega(K^{1/2})$ for the next at least period of $CK^2 \log^2 n$ steps, with probability at least $1 - e^{-\Omega(K^{1/2})}$. Lemma 18 in [20] then applies Part (b) of the stabilization lemma iteratively to obtain tighter lower and upper bounds. More specifically, after $O(\log \log K)$ periods, the variance is guaranteed to be in $\Omega(K^{2/3})$ and $O(n^{4/3})$. Since the number of frequencies at the lower border only changes very slowly (in comparison to the length of a period), we still have $\Theta(n)$ frequencies at the lower border at this point in time. While this is the case, the stabilization lemma can still be applied to show that these variance bounds are maintained. Each application of the stabilization lemma has a failure probability of $\exp(-\Omega(K^{1/3}))$, thus a union bound over $t_3/(CK^2 \log^2 n)$ applications of the lemma yields the probability bound stated in Theorem 4.1.

4.2 Trying to Translate Results to Cliff

We conjecture that Theorem 4.1 also holds when replacing OneMax with Cliff. We do not have a proof for this statement, but we will argue why this conjecture seems plausible, and what the challenges are in translating results from [20] on OneMax to Cliff.

Both Cliff and OneMax are functions of unitation, hence on both functions the dynamics can be analyzed by considering individual frequencies. On both functions, frequencies are likely to reach borders within $O(K^2)$ iterations and then frequencies may detach from their borders to perform a random walk. Thus, the approach from [20] that leads to the stabilization lemma can also be applied to Cliff.

These random walks are similar for both functions. If we pick an arbitrary but fixed frequency i then, for both functions, there are steps in which that frequency has no effect on the selection of the fitter offspring and then increasing the frequency has the same probability as decreasing it. These steps were called *random walk steps* in [26]. There are other steps, called *biased steps* in [26], in which a frequency can only increase on OneMax. This happens, for instance, when all other bits have the same number of ones in both samples x and y and then the i-th frequency determines which search point is reinforced. If exactly one of x_i and y_i is 1, that solution is reinforced and the i-th frequency increases. The probability of a biased step is $\Theta(1/\sqrt{V_t})$ and the expected drift of the i-th frequency is $\Theta\left(\frac{p_{i,t}(1-p_{i,t})}{K\sqrt{V_t}}\right)$ if no border is hit.

On Cliff, the situation is similar. If the i-th frequency has no impact on selection, a random walk step occurs. Biased steps may occur when all other bits have the same number of ones on both samples or when the number of ones on all other bits is precisely $2n/3$. In the latter case, the i-th frequency may decide whether a sample has $2n/3$ ones (i. e. is on top of the cliff) or $2n/3 + 1$ ones (i. e. is at the bottom of the second slope). It is not difficult to show that the probability of a biased step is $\Theta(1/\sqrt{V_t})$ as for OneMax.

One key difference is that on Cliff the drift can be either positive or negative. It is positive when the cGA focuses its search on one particular slope. However, the drift can be negative when sampling close to the cliff and the two offspring lie on different slopes.

A central argument from [20] is that the variance can be accurately described by studying the so-called *lifetime contribution* of one frequency, which is the total contribution that the frequency makes to the variance while the frequency does not reach any border. The lifetime contribution is then bounded from above and below by using a worst-case perspective for the drift: in each iteration, the drift may be chosen arbitrarily from a range between 0 and a maximum value that depends on V_t (cf. Lemma 11 and 12 in [20]). This worst-case view was necessary to deal with dependencies between frequencies and the intricate feedback loops. For these bounds it is crucial that the drift is always non-negative. To translate this approach to Cliff, one would have to allow negative drift values in addition to positive ones. This means that sudden changes between positive and negative drift values are possible and, ultimately, the worst-case bounds on the lifetime contribution become too weak to prove that the variance stabilizes to super-constant values. We conjecture that the worst-case view is too pessimistic here as the real dynamics are unlikely to rapidly switch between regimes where the drift is noticeably positive and noticeably negative. Providing rigorous arguments remains a challenge for future work.

4.3 Empirical Evidence

We provide empirical evidence to support our belief that the variance is super-constant for interesting ranges of K, including the ones from Theorem 4.1. We recorded the variance after $100\sqrt{n}K + 100K^2$ iterations and report averages taken over 100 runs, for increasing n and K chosen as different functions of n. The time bound is motivated by the upper bound of $O(\sqrt{n}K)$ for the expected optimization time of the cGA on OneMax for $K = \Omega(\sqrt{n} \log n)$ [26] and the upper bound of $O(K^2)$ for the expected time for random walks reaching a border (see [26] and Lemma 4 in [20]). Both results do not specify constants, hence we put a generous constant of 100 to enable the cGA to reach a state where the variance has stabilized.

The values of K are chosen from $\{\log n, n^{0.45}, \sqrt{n}, n^{0.75}, n\}$. The value $K = n^{0.45}$ is captured by Theorem 4.1 whereas $K = \log n$ and $K = \sqrt{n}$ are just outside the medium parameter regime.

On the left-hand side of Figure 2 we can see the variance scaling with n. For $K = \log n$, the variance seems to remain constant. Runs for $K = n$ were only performed up to $n = 340$ and the variance does not seem to increase, apart from a spike for small values of n. (This spike persisted when increasing the time limit to $100K^2 \log^2 n$.) All other values of K yield curves that have a clear upward trend. The right-hand side shows the same data, normalized by dividing by $K^{1/2}$ and for all K except $K = n$, the normalized values are strikingly close to 1. All curves but $K = \log n$ appear to be stable, suggesting that a variance lower bound of $\Omega(K^{1/2})$ might apply for medium K.

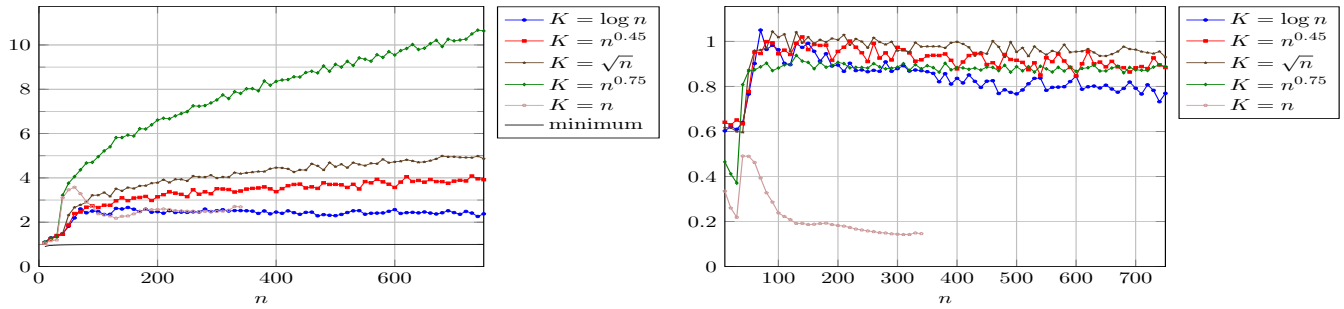

Figure 2: Variance after $100\sqrt{n}K + 100K^2$ iterations, averaged over 100 runs, for increasing n and different values of K. The black line indicates the minimum variance of $1 - 1/n$. The plot on the right-hand side shows the same curves divided by $K^{1/2}$.

5 A SEMI-RIGOROUS EXPONENTIAL LOWER BOUND

As explained in Section 4, we believe that the sampling variance of the cGA with potential P_t around the cliff is $\omega(1)$ and tends to lie stable for a long time. Under these assumptions, we can formally prove that the potential cannot efficiently cross the interval of negative drift around $2n/3$, which was established above in Theorem 3.1.

The following lemma assumes a variance in an interval $[v_\ell, v_u]$ while the potential is in the drift interval $[2n/3 - (v_\ell)^{1-\varepsilon}, 2n/3]$. Under conditions on K, v_ℓ and v_u, discussed after the proof, the time to pass the drift interval is exponential, with high probability.

Lemma 5.1. *Assume that there are a constant $\varepsilon' > 0$, functions $v_\ell = v_\ell(n) = \omega(1)$, $v_u = v_u(n) \le n$ and a constant $c > 1$ such that the property $V_t \in [v_\ell, v_u]$ holds for all points in time t where $P_t \in [a', b'] := [2n/3 - v_\ell^{1/2-\varepsilon'}, 2n/3]$. Assume $K \ge n^{\varepsilon'} v_u / v_\ell^{1-\varepsilon'}$ and $K = O(\sqrt{n})$ and define the hitting time $T := \min\{t \ge 0 \mid P_t \ge b'\}$. Then there is a constant $c' > 0$ such that given $P_0 \le a'$, it holds that $\Pr\left(T \le e^{c' K v_\ell^{1-\varepsilon'}/v_u}\right) = 2^{-\Omega(K v_\ell^{1-\varepsilon'}/v_u)}$.*

Proof. We verify the three conditions of the negative drift theorem with scaling (Theorem 2.1). Its parameters are chosen as $X_t = -KP_t$, $a = -Kb'$ and $b = -Ka'$.

For the first item, which deals with a lower bound on the drift, we use Theorem 3.1 to obtain (for $X_t = -KP_t$) the drift bound $\varepsilon = \varepsilon(n) = c_1\sqrt{v_\ell}$ for a constant $c_1 > 0$ within the interval $[a, b]$ of length $\ell = b - a = K v_\ell^{1/2-\varepsilon'}$.

To verify the second condition, we use Corollary 2.3 dealing with the concentration of the one-step change $|P_{t+1} - P_t|$ depending on the variance. Since we apply it to the scaled process $X_t = -KP_t$, the parameter λ is implicitly multiplied by K. Hence, there is an $r = c_2\sqrt{v_u}$ for some sufficiently large constant $c_2 > 0$ such that

$$\Pr(|X_{t+1} - X_t| \ge jr \mid \mathcal{F}; a < X_t) \le 2e^{-\min\{j^2 c_2^2, jc_2\sqrt{v_u}\}/3} \le e^{-j}$$

for $j \in \mathbb{N}_0$, using that $\sqrt{v_u} \ge \sqrt{v_\ell} \ge 1$. Note that the condition $a < X_t$ is equivalent to $P_t < b'$ and $|X_{t+1} - X_t| = |K(P_{t+1} - P_t)|$.

We now analyze the exponent in the final bound on T. We have $r^2 = c_2^2 v_u$ and therefore

$$\frac{\varepsilon\ell}{132r^2} \ge \frac{c_1\sqrt{v_\ell} K v_\ell^{1/2-\varepsilon'}}{132 c_2^2 v_u} = \frac{c_1 K v_\ell^{1-\varepsilon'}}{132 c_2^2 v_u},$$

which immediately leads to the exponent claimed in the statement of this lemma by setting $c' = c_1/(132c_2^2)$. Note that the exponent becomes $\Omega(n^{\varepsilon'})$ if $K \ge n^{\varepsilon'} v_u / v_\ell^{1-\varepsilon'}$. However, we still have to verify the third condition of the drift theorem.

First of all, we have $r = c_2\sqrt{v_u} \ge 1$ by choosing c_2 large enough and thus satisfy the lower bound on r^2 in the third condition. Next, we note that $r/\varepsilon \le \frac{c_2\sqrt{v_u}}{c_1\sqrt{v_\ell}} = O(\sqrt{n})$ since $v_u \le n$ and therefore $\log(r/\varepsilon) = O(\log n)$. Hence, we can use the bound $\frac{\varepsilon\ell}{132r^2} = \Omega(n^{\varepsilon'})$, assuming $K \ge n^{\varepsilon'} v_u / v_\ell^{1-\varepsilon'}$ and $K = O(\sqrt{n})$, from the previous paragraph to show that $\frac{\varepsilon\ell}{r^2 132 \log(r/\varepsilon)} \ge 1$, which is equivalent to the upper bound on r^2 in the third condition of the drift theorem. □

To apply Lemma 5.1, we need bounds on the variance while P_t is in the drift interval. From the discussions above, we conjecture that the variance stabilizes around a value $v^* = \Omega(K^{1/2})$, which would allow us to have $v_u/v_\ell = \Theta(1)$, hence $K v_\ell^{1-\varepsilon'}/v_u = K(v_\ell/v_u)^{1-\varepsilon'}/v_u^{\varepsilon'} = \Omega(Kn^{-\varepsilon'})$ (since $v_u \le n$). In this case, we obtain an exponential bound already for $K = \Omega(n^{2\varepsilon'})$. If the variance is allowed to oscillate between v_ℓ and v_u that are not of the same asymptotic order, we can still apply the lemma under reasonable assumptions. If the variance is allowed to oscillate between a lower bound v_ℓ and an upper bound v_u such that $v_u/v_\ell \le K^{1-\delta}$ for a constant $\delta > 0$ (like, e. g., with the bounds $\Omega(K^{2/3})$ and $O(K^{4/3})$ appearing in Theorem 4.1), then $K v_\ell^{1-\varepsilon'}/v_u \ge (K/K^{1-\delta}) v_u^{-\varepsilon'} = K^\delta / v_u^{\varepsilon'} \ge K^\delta / n^{\varepsilon'}$, which is still polynomially growing in K if, e. g., $K \ge n^{2\varepsilon'/\delta}$. Here $\varepsilon' > 0$ can be chosen arbitrarily small.

The potential does not jump over the negative drift interval. To rule out that the cGA optimizes Cliff efficiently, we also have to prove it unlikely that the potential changes drastically in one step and "jumps over the drift interval" $[2n/3 - V_t^{1/2-\varepsilon}, 2n/3]$. However, it is not difficult to prove that the following event is unlikely: there is a point of time t where $P_t < 2n/3 - (V_t)^{1/2-\varepsilon}$ but $P_{t+1} > 2n/3$. Since $V_{t+1} \ge (1 - 1/K)V_t \ge V_t/2$, the length of the drift interval stemming from Theorem 3.1 is at most halved in the transition from time t to $t+1$. Hence, if $P_{t+1} \le P_t + (V_t/2)^{1/2-\varepsilon}/2$, then $P_{t+1} \le 2n/3 - (V_{t+1})^{1/2-\varepsilon}/2$, i. e., the process has not jumped over the interval. For simplicity, we work with the lower bound $(V_t)^{1/2-\varepsilon}/4$ on the potential difference in the following. We can then easily apply Corollary 2.3 to show that the variance does not change by at least $(V_t)^{1/2-\varepsilon}/4$ in a step with overwhelming probability.

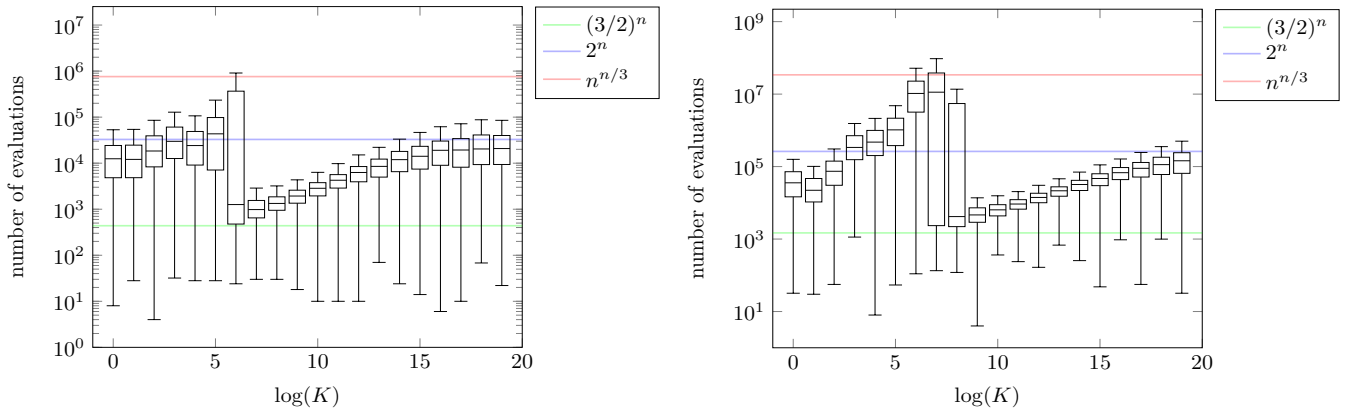

Figure 3: Box plots for the number of evaluations in 1000 runs on Cliff with $n = 15$ (left) and $n = 18$ (right) and exponentially increasing values of K. The plots also show values of $(3/2)^n$, 2^n and $n^{n/3}$ for comparison.

We choose $\lambda = K(V_t)^{1/2-\varepsilon}/4$ and obtain a failure probability of $e^{-\Omega(\min\{\lambda^2/V_t,\lambda\})} = e^{-\Omega(\min\{K^2/V_t^{2\varepsilon},KV_t^{1/2-\varepsilon}\})}$. If we have $K \geq n^{2\varepsilon}$ then the probability of increasing the potential by $V_t^{1/2-\varepsilon}/4 \leq (V_t/2)^{1/2-\varepsilon}/2$ in one step has probability $e^{-\Omega(K)}$, using $V_t \leq n$.

Altogether, the analyses presented in this section show that the potential, under reasonable assumptions on the sampling variance, takes exponential time to exceed the value $2n/3$. If $P_t \leq 2n/3$, sampling the optimum of cliff, i. e., the all-ones string, has probability at most $(3/4)^{n/9} = 2^{-\Omega(n)}$ since at least $n/9$ frequencies have to be below $3/4$. Hence, the optimum is not sampled in this situation with overwhelming probability. We have proven the following theorem.

Theorem 5.2. *In the setting of Lemma 5.1 the optimization time of the cGA on Cliff is $2^{\Omega(n^{\varepsilon'})}$ with probability $1 - 2^{-\Omega(n^{\varepsilon'})}$.*

6 EMPIRICAL RUNTIMES & OPEN PROBLEMS

Figure 3 shows boxplots highlighting the median runtimes and their distributions for $n = 15$ and $n = 18$ and K being set to a power of 2 from 1 to 2^{19}. Note that the y axis uses a logarithmic scale.

As can be seen, the parameter K has a significant impact on the runtime and its parameter landscape seems complex. For both problem sizes, the median runtime for small K is close to 2^n. We suspect that, for extreme updates, the cGA shows a chaotic behavior resembling random search (cf. Theorem 17 in [20] for OneMax). This is caused by extreme genetic drift [6, 26].

As K grows, the median runtime increases considerably, with some runs exceeding $n^{n/3}$ evaluations. For medium K, the frequencies tend to reach their borders quickly due to genetic drift. If most frequencies remain at their borders and the potential is close to $2n/3$, roughly $n/3$ frequencies must be at the lower border and the probability of sampling the optimum from there is at most $n^{-n/3}$.

This regime is followed by a sudden and steep drop at $K = 2^6 = 64$ for $n = 15$ and $K = 2^8 = 256$ for $n = 18$, respectively. When doubling K one more time, the distribution is highly concentrated around the median, and runtimes are close to $(3/2)^n$. We suspect that in this parameter range, where K is exponential in n, the frequencies increase slowly and evenly. When the potential reaches $2n/3$ and all frequencies are similar to $2/3$, the cGA would have a probability of roughly $(2/3)^n$ to sample the optimal solution. This would explain the phase transition where we suspect the expected optimization time to drop from $n^{\Omega(n)}$ to $2^{O(n)}$ and possibly even to at most $c^n \cdot \text{poly}(n)$ for a constant $3/2 \leq c < 2$.

When increasing K even further, the frequencies remain so close to their initial values of $1/2$ that the cGA behaves like random search again. Indeed, the median runtime seems to approach 2^n for the largest values of K examined.

So, the best choice of K seems to be in the sweet spot where the frequencies are able to rise equally towards a potential of $2n/3$ and stay there for long enough to sample the optimum.

We finish with some open problems related to these observations.

Open Problem 1. *Sharpen the variance bounds from [20] on OneMax, Theorem 4.1, and add variance bounds for times in $[0, t_1]$.*

Open Problem 2. *Prove rigorously that for the cGA on Cliff, with high probability, the variance is super-constant throughout an exponential period of time, for appropriate update strengths $K = n^{\Omega(1)}$.*

Open Problem 3. *Prove a lower bound of $n^{\Omega(n)}$ for the cGA on Cliff for appropriate values of K below the observed phase transition.*

Open Problem 4. *Prove an upper bound of $c^n \cdot \text{poly}(n)$ for a constant $c < 2$ for exponential K beyond the observed phase transition.*

In order to address the open problem 4, it might be necessary to prove that the frequencies tend to increase evenly.

Open Problem 5. *Prove that, for the cGA on OneMax or Cliff with large values of K, the frequencies tend to increase evenly from their initial value of $1/2$, and that they remain concentrated around the expectation for a period of time.*

Solving open problem 5 may not be as easy as it looks. Many standard concentration bounds do not apply since the frequencies are not independent and each step may have a large knock-on effect on future frequency dynamics, as discussed for the powerful method of bounded martingale differences in [3, Section 10.3].

ACKNOWLEDGMENTS

This work has been supported by the Australian Research Council (ARC) through grant FT200100536 and by the Independent Research Fund Denmark through grant DFF-FNU 8021-00260B.

REFERENCES

[1] D. Corus, P. S. Oliveto, and D. Yazdani. When hypermutations and ageing enable artificial immune systems to outperform evolutionary algorithms. *Theoretical Computer Science*, 832:166–185, 2020.

[2] D. Dang and P. K. Lehre. Simplified runtime analysis of estimation of distribution algorithms. In *Proceedings of the Genetic and Evolutionary Computation Conference (GECCO '15)*, pages 513–518, 2015.

[3] B. Doerr. Probabilistic tools for the analysis of randomized optimization heuristics. In B. Doerr and F. Neumann, editors, *Theory of Evolutionary Computation – Recent Developments in Discrete Optimization*, pages 1–87. Springer, 2020.

[4] B. Doerr. The runtime of the compact genetic algorithm on jump functions. *Algorithmica*, 83:3059–3107, 2021.

[5] B. Doerr and F. Neumann, editors. *Theory of Evolutionary Computation—Recent Developments in Discrete Optimization*. Springer, 2020. Also available at https://cs.adelaide.edu.au/~frank/papers/TheoryBook2019-selfarchived.pdf.

[6] B. Doerr and W. Zheng. Sharp bounds for genetic drift in estimation of distribution algorithms. *IEEE Transactions on Evolutionary Computation*, 24(6):1140–1149, 2020.

[7] S. Droste. A rigorous analysis of the compact genetic algorithm for linear functions. *Natural Computing*, 5(3):257–283, 2006.

[8] W. Feller. *An Introduction to Probability Theory and Its Applications*, volume 2. Wiley, 1971.

[9] T. Friedrich, T. Kötzing, M. S. Krejca, and A. M. Sutton. The compact genetic algorithm is efficient under extreme Gaussian noise. *IEEE Transactions on Evolutionary Computation*, 21(3):477–490, 2017.

[10] V. Hasenöhrl and A. M. Sutton. On the runtime dynamics of the compact genetic algorithm on jump functions. In *Proceedings of the Genetic and Evolutionary Computation Conference, (GECCO 2018)*, pages 967–974. ACM, 2018.

[11] M. Hauschild and M. Pelikan. An introduction and survey of estimation of distribution algorithms. *Swarm and Evolutionary Computation*, 1(3):111–128, 2011.

[12] M. A. Hevia Fajardo and D. Sudholt. Self-adjusting offspring population sizes outperform fixed parameters on the cliff function. In *Proceedings of the 16th ACM/SIGEVO Conference on Foundations of Genetic Algorithms (FOGA 2021)*, pages 1–15. ACM Press, 2021.

[13] W. Hoeffding. On the distribution of the number of successes in independent trials. *The Annals of Mathematical Statistics*, 27(3):713 – 721, 1956.

[14] J. Jägersküpper and T. Storch. When the plus strategy outperforms the comma strategy – and when not. In *Proceedings of the IEEE Symposium on Foundations of Computational Intelligence (FOCI 2007)*, pages 25–32, 2007.

[15] T. Jansen. *Analyzing Evolutionary Algorithms – The Computer Science Perspective*. Springer, 2013.

[16] N. L. Johnson, S. Kotz, and N. Balakrishnan. *Continuous Univariate Distributions*, volume 1. Wiley, 2nd edition, 1994.

[17] M. S. Krejca and C. Witt. Theory of estimation-of-distribution algorithms. In B. Doerr and F. Neumann, editors, *Theory of Evolutionary Computation: Recent Developments in Discrete Optimization*, pages 406–442. Springer, 2019.

[18] P. K. Lehre and P. T. H. Nguyen. Tight bounds on runtime of the univariate marginal distribution algorithm via anti-concentration. In *Proceedings of the Genetic and Evolutionary Computation Conference (GECCO '17)*, pages 1383–1390. ACM Press, 2017.

[19] J. Lengler, D. Sudholt, and C. Witt. Medium step sizes are harmful for the compact genetic algorithm. In *Proceedings of the Genetic and Evolutionary Computation Conference (GECCO '18)*, pages 1499–1506. ACM Press, 2018.

[20] J. Lengler, D. Sudholt, and C. Witt. The complex parameter landscape of the compact genetic algorithm. *Algorithmica*, 83(4):1096–1137, 2021.

[21] A. Lissovoi, P. S. Oliveto, and J. A. Warwicker. On the time complexity of algorithm selection hyper-heuristics for multimodal optimisation. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI 2019)*, volume 33, pages 2322–2329, 2019.

[22] F. Neumann and C. Witt. *Bioinspired Computation in Combinatorial Optimization – Algorithms and Their Computational Complexity*. Springer, 2010.

[23] P. S. Oliveto and C. Witt. Improved time complexity analysis of the simple genetic algorithm. *Theoretical Computer Science*, 605:21–41, 2015.

[24] M. Pelikan, M. Hauschild, and F. G. Lobo. Estimation of distribution algorithms. In J. Kacprzyk and W. Pedrycz, editors, *Handbook of Computational Intelligence*, pages 899–928. Springer, 2015.

[25] M. R. Sampford. Some inequalities on Mill's ratio and related functions. *The Annals of Mathematical Statistics*, 24(1):130–132, 1953.

[26] D. Sudholt and C. Witt. On the choice of the update strength in estimation-of-distribution algorithms and ant colony optimization. *Algorithmica*, 81(4): 1450–1489, 2019.

[27] C. Witt. Upper bounds on the runtime of the Univariate Marginal Distribution Algorithm on OneMax. In *Proceedings of the Genetic and Evolutionary Computation Conference (GECCO '17)*, pages 1415–1422. ACM Press, 2017.

[28] C. Witt. On crossing fitness valleys with majority-vote crossover and estimation-of-distribution algorithms. In *Proceedings of the 16th ACM/SIGEVO Conference on Foundations of Genetic Algorithms (FOGA 2021)*, pages 1–15. ACM Press, 2021.