

Faster Optimization Through Genetic Drift

Cella Florescu, Marc Kaufmann, Johannes Lengler, and Ulysse Schaller

Department of Computer Science, ETH Zürich, Zürich, Switzerland.
{cella.florescu, marc.kaufmann, johannes.lengler,
ulysse.schaller}@inf.ethz.ch

Abstract. The compact Genetic Algorithm (cGA), parameterized by its hypothetical population size K , offers a low-memory alternative to evolving a large offspring population of solutions. It evolves a probability distribution, biasing it towards promising samples. For the classical benchmark ONE-MAX, the cGA has two different modes of operation: a conservative one with small step sizes $\Theta(1/(\sqrt{n} \log n))$, which is slow but prevents genetic drift, and an aggressive one with large step sizes $\Theta(1/\log n)$, in which genetic drift leads to wrong decisions, but those are corrected efficiently. On ONE-MAX, an easy hill-climbing problem, both modes lead to optimization times of $\Theta(n \log n)$ and are thus equally efficient.

In this paper we study how both regimes change when we replace ONE-MAX by the harder hill-climbing problem DYNAMIC BIN-VAL. It turns out that the aggressive mode is not affected and still yields quasi-linear runtime $O(n \text{ polylog } n)$. However, the conservative mode becomes substantially slower, yielding a runtime of $\Omega(n^2)$, since genetic drift can only be avoided with smaller step sizes of $O(1/n)$. We complement our theoretical results with simulations.

Keywords: compact Genetic Algorithm · Genetic Drift · Estimation-of-Distribution Algorithm · Dynamic Binary Value

1 Introduction

Estimation-of-distribution algorithms (EDAs) are a family of randomized optimization heuristics in which the algorithm evolves a probability distribution over the search space. In each iteration, it samples solutions from this distribution, evaluates their quality (also called fitness), and updates the probability distribution accordingly. Examples in discrete domains include the cGA, UMDA, PBIL, ant colony systems like the MMAS, and multivariate systems like hBOA [26] or MIMIC [3], see [15] for a survey. EDAs turn out to be a powerful alternative to population-based heuristics like evolutionary and genetic algorithms. They have the advantage that they often sample from a wider region of the search space than their population-based alternatives, which makes them less susceptible to local deviations like (frozen or non-frozen) noise and local optima [6, 11, 12, 17, 30].

EDAs have been used for several decades, but theoretical investigations of EDAs have only started to gain momentum a few years ago. While for some aspects a clear picture has emerged, like that EDAs are able to cope with large amounts of noise [11], there is one aspect for which researchers have found a complex and ambiguous pattern: genetic drift. Genetic drift is the tendency of an algorithm to move through the fitness landscape even in absence of a clear signal-to-noise ratio.¹ While it is possible to avoid genetic drift by tracking the signal-to-noise ratio [8, 10], this conservative attempt of avoiding mistakes could potentially make the algorithm slow and inflexible. An alternative approach might be to embrace genetic drift, allow the algorithm to swiftly move through the search space, and let it correct mistakes as they appear.

Indeed, these two alternatives are exemplified by the behaviour of the compact Genetic Algorithm cGA on the pseudo-Boolean function ONEMAX [23]. The ONEMAX function assigns to a bit string $x \in \{0, 1\}^n$ the number of one-bits in x . It is one of the simplest and most classical hill-climbing benchmarks. The cGA maintains for each of the n coordinates a *frequency* p_i , which encodes the probability that the i -th bit is set to one in the distribution. In each iteration, it samples two solutions, and for each component i it shifts the frequency p_i by $1/K$ towards the value of the fitter of the two solutions, see Section 2 for full details. The step size $1/K$ determines how aggressively or conservatively the algorithm updates. It is well understood that the size of K determines whether genetic drift happens to a relevant extent on ONEMAX or not. If $K = \omega(\sqrt{n} \log n)$ then the frequencies move so slowly that the signal exceeds the noise, and all frequencies move slowly but steadily towards the upper boundary. This corresponds to the regime where genetic drift is avoided, and we refer to this as *conservative* regime. On the other hand, if $K = o(\sqrt{n} \log n)$ then the signal is weaker than the noise, and some bits move to the wrong boundary due to genetic drift. In the subsequent optimization process, these mistakes are then slowly corrected. We call this the *aggressive* regime.

It turns out that both regimes are equally efficient on ONEMAX. For suitable $K = C \log n$ with a large constant C , errors are corrected so quickly that the optimum is sampled in $O(n \log n)$ iterations.² On the other hand, if $K = C\sqrt{n} \log n$ with a large constant C , then the algorithm moves more slowly, but does not make any errors, which yields the same asymptotic runtime $O(n \log n)$. Both parameter settings are brittle with respect to smaller K : if either $K = C \log n$ or $K = C\sqrt{n} \log n$ are decreased only slightly, this results in a sudden loss of performance. On the other hand, if the parameter K is increased from either $K = C \log n$ or $K = C\sqrt{n} \log n$, then the performance deteriorates slowly but steadily. Hence, there are two optimal parameter settings for the cGA on ONE-

¹ The term *drift* is also used in the context of *drift analysis*, where it means the expected change, which is almost the opposite concept. The term “genetic drift” should not be confused with this other meaning of the term “drift”.

² This was only shown formally for the UMDA in [30] and [2], not for the cGA. However, [23] contains an informal argument why the results should also apply to the cGA.

MAX, a conservative one which avoids genetic drift and an aggressive one which embraces genetic drift.

Since the mentioned analysis was limited to ONEMAX, it remained open whether both modes of the algorithm also show comparable performance for other hill-climbing tasks. In this paper, we give a negative answer and show that for the harder hill-climbing problem DYNAMIC BINVAL, the aggressive mode still finds the optimum in quasi-linear time, while the conservative mode needs time $\Omega(n^2)$.

1.1 Our results

We investigate the cGA on the function DYNAMIC BINVAL, or DYNBV for short. This function, introduced in [20], builds on the classical linear test function BINARY VALUE that assigns to each binary string the integer that is represented by it in the binary number system. DYNBV is obtained by drawing at each iteration a random permutation of the weights and then evaluating all solutions with the permuted BINARY VALUE function, see Section 2 for a formal definition. BINARY VALUE is conjectured to be the hardest linear function³, and DYNBV is known to be the hardest dynamic linear function⁴ [20, 21]. This makes it the perfect benchmark for a hard hill-climbing task. For more discussion of the benchmark, see Section 1.2.

The conservative regime is slow. Our first main result is the following lower runtime bound, which holds for all $K = O(\text{poly}(n))$. In this range the runtime will at first increase quadratically in K , until K reaches the dimension of the search space - at which point the runtime dependency becomes $\Omega(K \cdot n)$.

Theorem 1. *Let $\bar{p} \in (0, \frac{1}{2})$ be arbitrary and consider the cGA with parameter $K = O(\text{poly}(n))$ and boundaries at \bar{p} and $1 - \bar{p}$ on DYNBV. Then with high probability, the optimum is not sampled during the first $\Omega(K \cdot \min\{K, n\})$ iterations.*

The reason for this is captured in Lemma 14, which states that in this period, there are always a linear number of bits which stay in some constant interval around their initialization value, so that it is exponentially unlikely to sample the optimum.

If we want to avoid genetic drift, we have to choose a rather large K (small step sizes) to overcome the small signal-to-noise ratio that is inherent in DYNBV. The following theorem states that any $K = O(n)$ will lead to substantial genetic drift and hence belongs to the aggressive regime. This agrees with the guidelines from [6] on how to avoid genetic drift.

³ For example for the $(1+1)$ -EA. It is a famous open problem to prove this formally.

⁴ It is not formally a dynamic linear function in the sense of [22], but can be obtained as a limit of such functions [20].

Theorem 2. *For every $\rho > 0$ and $\beta \in (\bar{p}, \frac{1}{2})$ there is $\delta > 0$ such that the following holds. Consider the cGA with parameter $K \leq \rho n$ on DYNBV. Then with high probability as $K \rightarrow \infty$ at least δn frequencies drop below β during optimization.*

Theorem 2 shows that indeed the only possibility to avoid substantial genetic drift is to set $K = \omega(n)$, which leads to a runtime of $\omega(n^2)$ by Theorem 1. Hence, DYNBV can not be optimized in quadratic time by any parameter setting of the cGA that avoids genetic drift. As we will see below, aggressive parameter settings that allow genetic drift are much more efficient. Before we come to this other regime, we complement Theorem 1 with a matching upper bound that holds when we are safely in the conservative regime with $K = \Omega(n \log n)$. To simplify the proof, we require a slight adjustment of the boundary values.⁵

Theorem 3. *Consider the cGA with parameter $K = \text{poly}(n)$ and boundaries at $\frac{1}{cn}$ and $1 - \frac{1}{cn}$ on DYNBV. If $K \geq c' \cdot n \log n$, and the constants $c, c' > 0$ are large enough, then the expected optimization time is $O(Kn)$.*

The main ingredient for this proof is to show that in the first polynomially many rounds, all frequencies will stay bounded away from the lower boundary (Proposition 18). Hence, the proof of Theorem 3 is similar to other proofs of upper runtime bounds in conservative regimes [10, 29].

Together, Theorems 1 and 3 give tight runtime bounds of $\Theta(Kn)$ in the conservative regime. This implies in particular that the runtime in this regime is much larger for DYNBV than for ONEMAX, where the runtime is $O(K\sqrt{n})$ [28].

We remark that the different transition points between conservative and aggressive regime ($K = \omega(n)$ in Theorem 3 and $K = \Omega(n \log n)$ in Theorem 3) are natural because there are different possible definitions of the conservative regime: that no frequency drops below $1/3$ (or any other fixed constant below $1/2$), that no frequency reaches the lower boundary, or that the number of frequencies hitting the lower boundary is sublinear. All these variants lead to different transition points between the conservative and aggressive regime.

The aggressive regime is fast. Our second result shows that in contrast, the optimization time of the cGA remains quasi-linear for small K , i.e. linear up to a poly-logarithmic factor. This corresponds to the aggressive regime where many frequencies reach the wrong boundary, but those errors are corrected efficiently. To make the analysis simpler, similar as for Theorem 3, we do not set the two boundary values at their standard values $1/n$ and $1 - 1/n$, but this time we even set them to $1/(n \text{ polylog } n)$ and $1 - 1/(n \text{ polylog } n)$. Moreover, we do not use the smallest possible (most aggressive) parameter choice $K = C \log n$ for the aggressive regime, but rather choose the slightly more conservative $K = \Theta(\log^2 n)$. Then we prove the following result.

⁵ More precisely, we set them to $\frac{1}{cn}, 1 - \frac{1}{cn}$ for a large enough constant $c > 0$. As our simulations, which are all conducted with boundaries $\frac{1}{n}, 1 - \frac{1}{n}$, show, this choice does not affect the asymptotic behaviour.

Theorem 4. *Consider the CGA with parameter $K = \Theta(\log^2 n)$ and boundaries $1/(n \log^7 n)$ and $1 - 1/(n \log^7 n)$ on DYNBV. Then the optimum is sampled in $O(n \cdot \text{polylog}(n))^6$ iterations with probability $1 - o(1)$.*

We note that we made no effort to optimize the exponent 7 of the poly-logarithmic factor in the boundaries. We conjecture that the true runtime for optimal parameters is $O(n \log n)$, and that this is achieved with the standard boundaries $1/n$ and $1 - 1/n$. Notably, this would mean that there is no substantial runtime difference in the aggressive regime with optimal parameters between ONEMAX and DYNBV, in stark contrast to the conservative regime. We do not quite show this statement, but we show it up to poly-logarithmic factors.

Compared to this conjecture, our analysis is likely not tight in several ways. Firstly, even for the given parameters we believe that the poly-logarithmic exponent of our runtime bound could be reduced at the cost of a more technical analysis. Secondly, both the conservative choice of K and the non-standard choice of the threshold likely bring us away from the optimal parameter. This simplifies the proof, but costs us performance, even though only logarithmic factors. We suspect that the optimal parameter setup is indeed the standard setup of $K = C \log n$ for a large constant C and boundaries $1/n$ and $1 - 1/n$. This is supported by the experiments presented in Section 7.

1.2 Discussion of the setup and related work

Signal steps and DYNBV. In order to understand genetic drift, a key question is how often each frequency receives a signal step. We call an iteration a *signal step* for frequency p_i if both solutions differ in this position i and the two values of this position are necessary to decide which of the two solution is fitter. When both solutions differ, but their values are irrelevant for identifying the fitter solution, then we call the iteration a *random walk step* for frequency p_i . In the initial phase of the CGA on ONEMAX, the probability of a signal step is $\Theta(1/\sqrt{n})$. For DYNBV, the signal probability is considerably weaker, namely of order $\Theta(1/n)$. In fact, in each iteration exactly one frequency receives a signal step, except when the two offspring sampled by the algorithm agree in every single bit.

As mentioned, DYNBV is the hardest dynamic linear function. This also holds in terms of the signal strength: when comparing two non-equal solutions, then exactly one frequency gets a signal, while all other frequencies perform a random walk step. This is the weakest signal strength among all dynamic linear functions, and even the hardest among all dynamic monotone functions [13], since every monotone function, static or dynamic, will always provide a signal step to at least one frequency when comparing two solutions.

Although it is a dynamic function, DYNBV provides a hill-climbing task in the sense that in each iteration and at any position, a one-bit gives a higher fitness than a zero-bit. Thus, pure hill-climbing heuristics such as Random Local

⁶ Our derived bound yields $O(n \log^{16} n)$ but this is not tight with regard to log-factors in various places, so we did not optimize for the exponent of the logarithm.

Search (RLS) can be highly efficient on this function. Moreover, DYNBV is more symmetric than the classical BINARY VALUE function, which makes the analysis simpler. All these properties make DYNBV the perfect benchmark for a theoretical runtime analysis of a hard hill-climbing task.

Related work. It has been shown that DYNAMIC BINVAL is harder to optimize by evolutionary algorithms than static monotone functions in various ways. The $(1, \lambda)$ -EA with self-adapting offspring population size fails on DYNAMIC BINVAL while succeeding on ONEMAX if the hyperparameters are not set correctly [14]. Furthermore, a “switching” variant of DYNAMIC BINVAL minimizes drift in the number of zeros at every search point for the $(1 + 1)$ -EA for any mutation rate at every search point, making it harder to optimize than any *static* monotone function [13].

We do not claim that the aggressive mode of the cGA is generally superior to the conservative mode. However, our results show that the other extreme position of avoiding genetic drift at all costs, does cost performance for DYNAMIC BINVAL. On the other hand, the conservative mode was shown to be superior on the function DECEPTIVELEADINGBLOCKS for some parameter settings [9, 16], though a discussion at a Dagstuhl seminar shows that opinions are split about the implications of these results [7]. We hope that further research will give a clearer and more nuanced picture on the benefits and drawbacks of genetic drift.

1.3 Overview of the paper

In Section 2 we describe the cGA and our benchmark DYNAMIC BINVAL, followed by important terminology and technical tools which we will use throughout the paper. Section 3 characterizes the dynamics of the marginal probabilities which are used in the generation of offspring. In Section 4, we prove the lower runtime bound, Theorem 1, and Theorem 2, thus establishing that any $K = O(n)$ will lead to genetic drift. In the subsequent Section 5, we prove Theorem 3, the upper runtime bound for the conservative regime. Section 6 establishes an upper runtime bound for the aggressive regime, culminating in Theorem 4. The paper concludes with the simulations in Section 7.

2 Setting

Our search space is always $\{0, 1\}^n$. We say that an event $\mathcal{E} = \mathcal{E}(n)$ holds *with high probability* or *whp* if $\Pr[\mathcal{E}] \rightarrow 1$ as $n \rightarrow \infty$. We may for simplicity omit the parameter t indicating the iteration when it is clear from context.

2.1 The Algorithm: the cGA with hypothetical population size K

We begin with an intuitive description of the cGA. Before the start of the algorithm, we fix a capping probability $\bar{p} < \frac{1}{2}$ and a hypothetical population size K , which we may think of as an inverse update strength. At every iteration,

the algorithm generates two offspring x and y independently of each other by the same sampling procedure. At iteration t , the i -th bit of the offspring to be sampled is set (independently of all other bits and of all previous iterations) to 1 with probability $p_{i,t}$ and set to 0 otherwise. The probabilities are initialized to $\frac{1}{2}$ for all bits and evolve according to the following procedure: At each iteration, the fitness of x and y are compared according to the fitness function f - in our case this will be DYNAMIC BINVAL. If the fitter offspring contains a 1 at position i , we increase the probability of sampling a 1 bit, p_i , by $\frac{1}{K}$ for the next iteration, otherwise we decrease it by the same amount. If there is no strictly fitter offspring, i.e. $f(x) = f(y)$, all probabilities $p_{i,t}$ remain unchanged in the next iteration. To ensure the algorithm does not get stuck by fixing one of the bits, i.e. sampling a 1 with probability 1 or sampling a 0 with probability 1, we restrict the possible values for probabilities $p_{i,t}$ to the interval $[\bar{p}, 1 - \bar{p}]$. If an update step would make a $p_{i,t}$ exceed these bounds, we set it to the boundary value instead. The algorithm stops when the optimum has been sampled (as one of the two offspring in a given iteration). The full pseudocode is provided in Algorithm 1.

Algorithm 1 cGA(f, K, \bar{p})

```

 $t \leftarrow 0$ 
 $p_{1,t} \leftarrow p_{2,t} \leftarrow \dots \leftarrow p_{n,t} \leftarrow \frac{1}{2}$ 
while optimum has not been sampled do
  for  $i \in \{1, 2, \dots, n\}$  do
     $x_i \leftarrow 1$  with probability  $p_{i,t}$  and 0 otherwise
     $y_i \leftarrow 1$  with probability  $p_{i,t}$  and 0 otherwise
  end for
  if  $f(x) = f(y)$  then
     $t \leftarrow t + 1$ 
    continue
  else if  $f(x) < f(y)$  then
    swap  $x$  and  $y$ 
  end if
  for  $i \in \{1, 2, \dots, n\}$  do
    if  $x_i > y_i$  then
       $p'_{i,t+1} \leftarrow p_{i,t} + \frac{1}{K}$ 
    else if  $x_i < y_i$  then
       $p'_{i,t+1} \leftarrow p_{i,t} - \frac{1}{K}$ 
    else
       $p'_{i,t+1} \leftarrow p_{i,t}$ 
    end if
     $p_{i,t+1} \leftarrow \min \{ \max \{ \bar{p}, p'_{i,t+1} \}, 1 - \bar{p} \}$ 
  end for
   $t \leftarrow t + 1$ 
end while

```

2.2 The Benchmark: DYNAMIC BINVAL

The benchmark we consider is the DYNAMIC BINVAL function.

In DYNAMIC BINVAL, at each iteration t , we draw uniformly at random (and independently of everything else) from the set of bijections from $\{1, \dots, n\}$ onto itself an element $\pi_t: \{1, 2, \dots, n\} \rightarrow \{1, 2, \dots, n\}$. Note that this can be seen as a permutation of the bits of the search point. The fitness function for iteration t is then given by

$$f_t(x) = \sum_{i=1}^n 2^{n-i} \cdot x_{\pi_t(i)}.$$

Intuitively, the offspring which has a 1 bit at the most significant position (given by the permutation π_t) at which the two offspring differ is considered fitter.

2.3 Terminology

Below, we introduce some useful concepts and terminology that will be used throughout the paper.

Signal step. A signal step of a bit $i \in \{1, 2, \dots, n\}$ is the increase in the marginal probability of bit i during an iteration t where the value of this bit in the two offspring was decisive. In other words, bit i performs a signal step at iteration t if and only if the two offspring differ at bit i and are equal at any other bit i' for which $\pi_t(i') < \pi_t(i)$.

Random step. A random step of a bit $i \in \{1, 2, \dots, n\}$ is the change in the marginal probability of bit i during an iteration where the value of this bit in the two offspring was not decisive. Thus, all changes in a bit's marginal probability that are not signal steps are random steps.

Sampling variance. The sampling variance at time t is the variance of the binomial distribution induced by the probabilities $p_{i,t}, i = 1, \dots, n$. We denote it by $V_t := \sum_{i=1}^n p_{i,t}(1 - p_{i,t})$. Intuitively, it is the sum at time t of the variances contributed by the probability of each frequency in the generating distribution.

Lower/Upper boundary. Recall from subsection 2.1 that the possible values for the probabilities $p_{i,t}$ are restricted to an interval $[\bar{p}, 1 - \bar{p}] \subsetneq [0, 1]$ for some value $\bar{p} = O(\frac{1}{n})$ to ensure that the algorithm does not get stuck. The values \bar{p} and $1 - \bar{p}$ are going to be referred to as lower and upper boundary respectively. For example, we say a frequency $p_{i,t}$ is at its upper boundary if $p_{i,t} = 1 - \bar{p}$. Note that the distances of the lower and upper boundaries from 0 and 1 is always the same, respectively. \bar{p} is fixed for the entire execution of the algorithm. It will always either be $\frac{1}{cn}$ or $\frac{1}{n \log^c n}$ for constant $c > 0$, and it will usually be clear from context which value of \bar{p} is assumed.

2.4 Drift analysis and concentration inequalities

The analysis of evolutionary algorithms relies heavily on drift theorems. These allow to transform statements about the expected one-iteration change of a potential function, a proxy for the function to be optimized, into runtime bounds. For an overview see [19]. We will need in particular the following negative drift theorem.

Theorem 5 (Theorem 2 in [25]). *Let $(X_t)_{t \in \mathbb{N}}$ be a stochastic process over some space $S \subseteq \mathbb{R}_0^+$, adapted to a filtration $(\mathcal{F}_t)_{t \in \mathbb{N}}$. Suppose there exist an interval $[a, b] \subseteq \mathbb{R}$ and, possibly depending on $l := b - a$, a drift bound $\varepsilon := \varepsilon(l) > 0$ as well as a scaling factor $r := r(l)$ such that for all $t \geq 0$ the following three conditions hold:*

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

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

We will further use the variable drift theorem.

Theorem 6 (Theorem 15 in [18]). *Let $(X_t)_{t \in \mathbb{N}}$ be a stochastic process over some state space $S \subseteq \{0\} \cup [x_{\min}, x_{\max}]$, adapted to a filtration $(\mathcal{F}_t)_{t \in \mathbb{N}}$, where $x_{\min} > 0$. Let $h(x) : [x_{\min}, x_{\max}] \rightarrow \mathbb{R}^+$ be a monotone increasing function such that $1/h$ is integrable on $[x_{\min}, x_{\max}]$ and $\mathbb{E}[X_t - X_{t+1} \mid \mathcal{F}_t] \geq h(X_t)$ if $X_t \geq x_{\min}$. Then it holds for the first hitting time $T := \min\{t : X_t = 0\}$ that*

$$\mathbb{E}[T \mid X_0] \leq \frac{x_{\min}}{h(x_{\min})} + \int_{x_{\min}}^{X_0} \frac{1}{h(x)} dx.$$

We will also need the next theorem in our proof of Lemma 19.

Theorem 7 (Theorem 1 in [24]). *Consider a Markov process $\{X_t\}_{t \geq 0}$ with state space S and a function $g : S \rightarrow \mathbb{R}_0^+$. Let $T := \inf\{t \geq 0 : g(X_t) = 0\}$. If there exists $\delta > 0$ such that for every time $t \geq 0$ and every state X_t with $g(X_t) > 0$ the condition $\mathbb{E}[g(X_t) - g(X_{t+1}) \mid X_t] \geq \delta$ holds, then*

$$\mathbb{E}[T \mid X_0] \geq \frac{g(X_0)}{\delta} \quad \text{and} \quad \mathbb{E}[T] \geq \frac{\mathbb{E}[g(X_0)]}{\delta}. \quad (1)$$

The same statements hold if T is redefined to $T := \inf\{t \geq 0 : g(X_t) \leq \kappa\}$ for some $\kappa > 0$ and the condition is relaxed to $\mathbb{E}[g(X_t) - g(X_{t+1}) \mid X_t \geq \kappa] \geq \delta$ for all $t \geq 0$ and all X_t .

In addition, we require the following tail bounds for hypergeometric random variables. More explicitly, we consider a random variable X with cumulative distribution function $h_X(M, N, n, i) = \Pr[X = i \mid M, N, n] = \frac{\binom{M}{i} \binom{N-M}{n-i}}{\binom{N}{n}}$. Here we

denote the population size by N , the number of success states in the population by M , the number of draws by n , and the number of observed successes by i . Then the following tail estimates hold.

Lemma 8 ([27, Section 5]). *Let X be a hypergeometric random variable with cumulative distribution function as described above. Let further $t > 0$. Then*

$$\Pr[X \geq \mathbb{E}[X] + tn] \leq e^{-2t^2n} \quad \text{and} \quad \Pr[X \leq \mathbb{E}[X] - tn] \leq e^{-2t^2n}.$$

Finally, we make use of the classical Chernoff bound.

Theorem 9 (Chernoff Bound [5, Section 1.10]). *Let X_1, \dots, X_n be independent random variables taking values in $[0, 1]$. Let $X = \sum_{i=1}^n X_i$ and let $\delta \in [0, 1]$. Then*

$$\Pr[X \geq (1 + \delta)\mathbb{E}[X]] \leq \exp\left(-\frac{\delta^2\mathbb{E}[X]}{3}\right).$$

3 Dynamics of the marginal probabilities

We start by analyzing how the cGA behaves on the DYNAMIC BINVAL at the level of a single iteration. The first proposition computes the probability that a bit i at which the two offspring differ gets a signal step (in some iteration t). This proposition captures the main difference between DYNBV and ONEMAX. For ONEMAX, a position which differs in the two offspring has probability $1/\max\{\sqrt{V_t}, 1\}$ to perform a signal step [29], and performs a random step otherwise. For DYNBV, the probability is $1/\max\{V_t, 1\}$, so the term $\sqrt{V_t}$ is replaced by V_t . Hence, signal steps are more likely for ONEMAX, and the signal-to-noise ratio of ONEMAX is larger than for DYNBV. This corresponds to the fact that ONEMAX is a particularly easy function to optimize.

Proposition 10. *Let $K \geq 1$ and $\bar{p} \in (0, \frac{1}{2})$ be arbitrary, and consider the algorithm $\text{cGA}(K, \bar{p})$ on DYNBV and some bit $i \in [n]$ and iteration t . For the permutation π_t drawn at iteration t , we denote by $S_{i,t}$ the event that all bits $i' \in [n]$ that appear before i in the permutation, i.e. such that $\pi_t(i') < \pi_t(i)$, are equal in the two offspring. Then it holds that*

$$\Pr[S_{i,t}] = \Theta\left(\frac{1}{\max\{V_t, 1\}}\right).$$

Proof. To compute this probability, consider the random variable $D_{i,t}$, which denotes the number of bits $i' \neq i$ that are different in the two offspring. If we denote by $V_t^{\neq i} = \sum_{j \neq i} p_{j,t}(1 - p_{j,t})$ the variance of the bits $i' \neq i$, we can observe that $\mathbb{E}[D_{i,t}] = \sum_{j \neq i} 2p_{j,t}(1 - p_{j,t}) = 2V_t^{\neq i}$.

The main observation that will now allow us to find asymptotic bounds for $\Pr[S_{i,t}]$ is the following: Once we have fixed the number $D_{i,t} = k \in \{0, 1, \dots, n-1\}$ of bits $i' \neq i$ that differ in the two offspring, the probability of drawing a

permutation π_t such that bit i receives the signal is $\frac{1}{k+1}$. This is because bit i only receives the signal if $\pi_t(i) < \pi_t(i')$ for all bits $i' \neq i$ that differ in the offspring. Since the permutation π_t is drawn uniformly at random, the probability that $\pi_t(i)$ assumes the smallest value from a set of $k+1$ numbers is $\frac{1}{k+1}$.

To reach the desired conclusion, we will first show that $\Pr[S_{i,t}] = \Theta\left(\frac{1}{\max\{V_t^{\neq i}, 1\}}\right)$. Intuitively, this holds because the number $D_{i,t}$ of bits $i' \neq i$ that differ in the offspring is concentrated around its expected value (since it is a sum of independent Bernoulli-distributed random variables), and $\mathbb{E}[D_{i,t}] = 2V_t^{\neq i}$. Formally, this can be shown via Chernoff bounds after conditioning on the value of $D_{i,t}$.

For the lower bound we obtain:

$$\begin{aligned} \Pr[S_{i,t}] &= \sum_{k=0}^{n-1} \Pr[S_{i,t} \mid D_{i,t} = k] \cdot \Pr[D_{i,t} = k] \\ &= \sum_{k=0}^{n-1} \frac{1}{k+1} \cdot \Pr[D_{i,t} = k] \\ &\geq \sum_{k=0}^{\lfloor \frac{3}{2}\mathbb{E}[D_{i,t}] \rfloor - 1} \frac{1}{k+1} \cdot \Pr[D_{i,t} = k]. \end{aligned}$$

Following this truncation of the sum, we can use the fact that for the first factors in the sum it holds that $\frac{1}{k+1} \geq \frac{1}{\lfloor \frac{3}{2}\mathbb{E}[D_{i,t}] \rfloor}$:

$$\begin{aligned} \Pr[S_{i,t}] &\geq \frac{1}{\lfloor \frac{3}{2}\mathbb{E}[D_{i,t}] \rfloor} \cdot \sum_{k=0}^{\lfloor \frac{3}{2}\mathbb{E}[D_{i,t}] \rfloor - 1} \Pr[D_{i,t} = k] \\ &\geq \frac{1}{\frac{3}{2}\mathbb{E}[D_{i,t}]} \cdot \left(1 - \Pr\left[D_{i,t} > \left\lfloor \frac{3}{2}\mathbb{E}[D_{i,t}] \right\rfloor\right]\right) \\ &= \frac{1}{\frac{3}{2}\mathbb{E}[D_{i,t}]} \cdot \left(1 - \Pr\left[D_{i,t} \geq \left\lfloor \frac{3}{2}\mathbb{E}[D_{i,t}] \right\rfloor + 1\right]\right) \\ &\geq \frac{1}{\frac{3}{2}\mathbb{E}[D_{i,t}]} \cdot \left(1 - \Pr\left[D_{i,t} \geq \frac{3}{2}\mathbb{E}[D_{i,t}]\right]\right). \end{aligned}$$

The last equality is due to the fact that $D_{i,t}$ only assumes values in \mathbb{N} .

Since $D_{i,t}$ is a random variable that is the sum of independent Bernoulli-distributed random variables with $\mathbb{E}[D_{i,t}] = 2V_t^{\neq i}$, we can apply the Chernoff bounds to further get:

$$\begin{aligned} \Pr[S_{i,t}] &\geq \frac{1}{\frac{3}{2}\mathbb{E}[D_{i,t}]} \cdot \left(1 - e^{-\frac{1}{12}\mathbb{E}[D_{i,t}]}\right) \\ &= \frac{1}{3V_t^{\neq i}} \cdot \left(1 - e^{-\frac{1}{6}V_t^{\neq i}}\right) \\ &= \Omega\left(\frac{1}{\max\{V_t^{\neq i}, 1\}}\right), \end{aligned} \tag{2}$$

where the last step holds for $V_t^{\neq i} \geq 1$ because then $1 - e^{-\frac{1}{6}V_t^{\neq i}} \geq 1 - e^{-1/6} = \Omega(1)$, and it holds for $V_t^{\neq i} < 1$ because then $1 - e^{-\frac{1}{6}V_t^{\neq i}} = \Theta(V_t^{\neq i})$.

The upper bound can be computed in a very similar manner, again by conditioning on the value of $D_{i,t}$.

$$\begin{aligned}
\Pr[S_{i,t}] &= \sum_{k=0}^{n-1} \Pr[S_{i,t} \mid D_{i,t} = k] \cdot \Pr[D_{i,t} = k] \\
&= \sum_{k=0}^{n-1} \frac{1}{k+1} \cdot \Pr[D_{i,t} = k] \\
&= \sum_{k=\lceil \frac{1}{2}\mathbb{E}[D_{i,t}] \rceil - 1}^{n-1} \frac{1}{k+1} \cdot \Pr[D_{i,t} = k] + \sum_{k=0}^{\lceil \frac{1}{2}\mathbb{E}[D_{i,t}] \rceil - 2} \frac{1}{k+1} \cdot \Pr[D_{i,t} = k] \\
&\leq \frac{1}{\lceil \frac{1}{2}\mathbb{E}[D_{i,t}] \rceil} \cdot \sum_{k=\lceil \frac{1}{2}\mathbb{E}[D_{i,t}] \rceil - 1}^{n-1} \Pr[D_{i,t} = k] + \sum_{k=0}^{\lceil \frac{1}{2}\mathbb{E}[D_{i,t}] \rceil - 2} 1 \cdot \Pr[D_{i,t} = k].
\end{aligned}$$

Since, we are dealing with probabilities of disjoint events, we know that

$$\sum_{k=\lceil \frac{1}{2}\mathbb{E}[D_{i,t}] \rceil - 1}^{n-1} \Pr[D_{i,t} = k] \leq 1.$$

We can employ Chernoff bounds again to obtain:

$$\begin{aligned}
\Pr[S_{i,t}] &\leq \frac{1}{\frac{1}{2}\mathbb{E}[D_{i,t}]} + \Pr\left[D_{i,t} \leq \left\lceil \frac{1}{2}\mathbb{E}[D_{i,t}] \right\rceil - 2\right] \\
&\leq \frac{1}{\frac{1}{2}\mathbb{E}[D_{i,t}]} + \Pr\left[D_{i,t} \leq \frac{1}{2}\mathbb{E}[D_{i,t}]\right] \\
&\quad \text{(since } D_{i,t} \text{ only assumes values in } \mathbb{N}\text{)} \\
&\leq \frac{1}{\frac{1}{2}\mathbb{E}[D_{i,t}]} + e^{-\frac{1}{8}\mathbb{E}[D_{i,t}]} \quad \text{(Chernoff bounds)} \\
&= \frac{1}{V_t^{\neq i}} + e^{-\frac{1}{4}V_t^{\neq i}} = O\left(\frac{1}{V_t^{\neq i}}\right).
\end{aligned}$$

Since $\Pr[S_{i,t}]$ is a probability, we clearly also have $\Pr[S_{i,t}] = O(1)$, and hence $\Pr[S_{i,t}] = O\left(\frac{1}{\max\{V_t^{\neq i}, 1\}}\right)$.

We have shown that $\Pr[S_{i,t}] = \Theta\left(\frac{1}{\max\{V_t^{\neq i}, 1\}}\right)$, so in order to conclude the proof it remains to show that $\max\{V_t^{\neq i}, 1\} = \Theta(\max\{V_t, 1\})$. Clearly $\max\{V_t^{\neq i}, 1\} \leq \max\{V_t, 1\}$ since $V_t^{\neq i} \leq V_t$. On the other hand, $V_t - V_t^{\neq i} = p_{i,t}(1 - p_{i,t}) \leq \frac{1}{4}$, and hence

$$V_t = V_t^{\neq i} + (V_t - V_t^{\neq i}) \leq V_t^{\neq i} + \frac{1}{4} = O(\max\{V_t^{\neq i}, 1\}),$$

which concludes the proof. \square

Using Proposition 10, one can describe the transition matrix of the marginal probabilities, which then allows to compute the drift of these marginal probabilities. As for Proposition 10, the same formulas would hold for ONEMAX with V_t replaced by $\sqrt{V_t}$.

Proposition 11. *Let $K \geq 1$ and $\bar{p} \in (0, \frac{1}{2})$ be arbitrary, and consider the algorithm $cGA(K, \bar{p})$ on DYNBV. Then for all $i \in [n]$ and $t \in \mathbb{N}$ we have $p_{i,t+1} = \min \{ \max \{ \bar{p}, p'_{i,t+1} \}, 1 - \bar{p} \}$ where*

$$p'_{i,t+1} = \begin{cases} p_{i,t}, & \text{with probability } 1 - 2p_{i,t}(1 - p_{i,t}) \\ p_{i,t} + \frac{1}{K}, & \text{with probability } \left(\frac{1}{2} + \Theta \left(\frac{1}{\max\{V_t, 1\}} \right) \right) 2p_{i,t}(1 - p_{i,t}) . \\ p_{i,t} - \frac{1}{K}, & \text{with probability } \left(\frac{1}{2} - \Theta \left(\frac{1}{\max\{V_t, 1\}} \right) \right) 2p_{i,t}(1 - p_{i,t}) \end{cases}$$

This implies $\mathbb{E}[p_{i,t+1} - p_{i,t} \mid p_{i,t}] = \Theta \left(\frac{p_{i,t}(1-p_{i,t})}{K \cdot \max\{V_t, 1\}} \right)$, where the lower bound requires $p_{i,t} < 1 - \bar{p}$ and the upper bound requires $p_{i,t} > \bar{p}$.

Proof. This proof follows the idea of Lemma 2 in [23]. The main idea is simply to use that the two offspring differ at bit i in iteration t with probability $2p_{i,t}(1 - p_{i,t})$, and that this probability does not change if we condition on $S_{i,t}$ since this event only influences the offspring value at some bits i' different from i .

First note that $p'_{i,t+1} \neq p_{i,t}$ if and only if the two bits at position i in the offspring are sampled differently. This implies that the event $p'_{i,t+1} = p_{i,t}$ happens with probability $1 - p_{i,t}(1 - p_{i,t}) - (1 - p_{i,t})p_{i,t} = 1 - 2p_{i,t}(1 - p_{i,t})$. We will now derive asymptotically tight bounds for the probability that $p'_{i,t+1} = p_{i,t} + \frac{1}{K}$, as the bounds for $p'_{i,t+1} = p_{i,t} - \frac{1}{K}$ follow analogously by symmetry.

Note that during one iteration, at most one bit in the offspring takes a signal step, and all other bits take random steps. To be more precise, there exists a signal step if and only if the two offspring have a bit in which they differ.

Consider therefore an arbitrary iteration t . Let $S_{i,t}$ be the event defined in Proposition 10, and let $\bar{S}_{i,t}$ be its complement, namely that bit i did not receive the signal for sure (which implies that bit i will do a random step). Conditioning on these events, we can rewrite the probability of $p'_{i,t} = p_{i,t} + \frac{1}{K}$ as:

$$\begin{aligned} \Pr \left[p'_{i,t+1} = p_{i,t} + \frac{1}{K} \right] &= \Pr \left[p'_{i,t+1} = p_{i,t} + \frac{1}{K} \mid S_{i,t} \right] \cdot \Pr[S_{i,t}] \\ &\quad + \Pr \left[p'_{i,t+1} = p_{i,t} + \frac{1}{K} \mid \bar{S}_{i,t} \right] \cdot \Pr[\bar{S}_{i,t}]. \end{aligned} \tag{3}$$

Let us now analyze the factors in the formula above. If bit i has the chance of receiving the signal, then it has the possibility to determine whether the update with respect to the first or second offspring. Therefore, in order to observe an increase in its frequency, it only needs to happen that the two offspring differ at bit i . Thus, the first conditional probability is

$$\Pr \left[p'_{i,t+1} = p_{i,t} + \frac{1}{K} \mid S_{i,t} \right] = 2p_{i,t}(1 - p_{i,t}). \tag{4}$$

On the other hand, if bit $\bar{S}_{i,t}$ takes place, then bit i does not contribute to the decision whether to update with respect to the first or second offspring. Thus, we observe an increase in its marginal frequency only if bit i is set to 1 in the fitter individual and to 0 in the other. This leads to the conditional probability

$$\Pr\left[p'_{i,t+1} = p_{i,t} + \frac{1}{K} \mid \bar{S}_{i,t}\right] = p_{i,t}(1 - p_{i,t}). \quad (5)$$

As already mentioned in Proposition 10, the probability of the event $S_{i,t}$ taking place is nothing but the probability that all the bits $i' \neq i$ for which $\pi_t(i') < \pi_t(i)$ are equal in the two offspring, and it holds that

$$\Pr[S_{i,t}] = \Theta\left(\frac{1}{\max\{V_t, 1\}}\right).$$

Combining this result with equations (3), (4) and (5) yields:

$$\begin{aligned} \Pr\left[p'_{i,t+1} = p_{i,t} + \frac{1}{K}\right] &= 2p_{i,t}(1 - p_{i,t}) \cdot \Pr[S_{i,t}] + p_{i,t}(1 - p_{i,t}) \cdot (1 - \Pr[S_{i,t}]) \\ &= p_{i,t}(1 - p_{i,t}) \cdot (1 + \Pr[S_{i,t}]) \\ &= p_{i,t}(1 - p_{i,t}) \cdot \left(1 + \Theta\left(\frac{1}{\max\{V_t, 1\}}\right)\right) \end{aligned} \quad (6)$$

The final statement of the lemma on the expectation follows easily from the aforementioned probability bounds and the way the cGA caps the probabilities at the boundaries. \square

4 Lower bound on the runtime

In this section, we prove lower bounds for the runtime of the compact Genetic Algorithm on DYNAMIC BINVAL when K is polynomial in the number of bits. The idea is to bound the number of signal steps that a given bit makes over a certain number of iterations, and use this bound to show that a linear number of frequencies stay a constant distance away from the boundaries for $\Omega(K \cdot \min\{K, n\})$ iterations. As long as this is the case, the probability to sample the optimum in any given iteration is exponentially small, and hence a union bound over the iterations gives us a high probability lower bound on the runtime. We further prove Theorem 2, which states that genetic drift occurs whenever $K = O(n)$.

We start by upper bounding the probability that a fixed frequency gets a signal step in a given iteration, under the condition that enough bits have marginal probabilities far away from the boundaries.

Corollary 12. *Let $K \geq 1$ and $\bar{p} \in (0, \frac{1}{2})$ be arbitrary, and consider the algorithm cGA(K, \bar{p}) on DYNBV. Assume that at iteration t there are at least γn bits whose marginal probabilities are within $[\frac{1}{6}, \frac{5}{6}]$, for some constant $\gamma > 0$. Then the probability of having a signal step on any fixed bit is $O(\frac{1}{n})$.*

Proof. Because there are at least γn bits whose marginal probabilities are within the interval $[\frac{1}{6}, \frac{5}{6}]$, and for each of these we can lower bound the product $p_{i,t}(1 - p_{i,t}) \geq \frac{1}{6} \cdot \frac{5}{6} = \frac{5}{36}$, we have

$$V_t = \sum_{i=1}^n p_{i,t}(1 - p_{i,t}) \geq \frac{5\gamma}{36}n = \Omega(n).$$

Now fix a bit i at iteration t , and recall the event $S_{i,t}$ from Proposition 10. This event is a necessary condition for bit i to have a signal step in iteration t , and hence (using Proposition 10 and the bound on the sampling variance) the probability that this happens is at most

$$\Pr[S_{i,t}] = \Theta\left(\frac{1}{\max\{V_t, 1\}}\right) = O\left(\frac{1}{n}\right).$$

□

Next, we prove a lemma that guarantees that the displacement of the marginal probabilities caused by $O(K^2)$ random steps is bounded in absolute value by $\frac{1}{6}$ for a linear number of bits. The approach is similar to the proof of Lemma 13 in [28].

Lemma 13. *Let $K \geq 1$ and $\bar{p} \in (0, \frac{1}{2})$ be arbitrary, and consider the algorithm $cGA(K, \bar{p})$ on DYNBV. Consider a fixed frequency $i \in [n]$ and let $t \leq \alpha K^2$, where $\alpha > 0$ is a small enough constant. With probability $\Theta(1)$, the first t random steps of frequency i lead to a total change of the bit's marginal probability that is within $[-\frac{1}{6}, \frac{1}{6}]$.*

Moreover, for a small enough constant $\gamma > 0$, the probability that the above holds for less than γn bits among the first $\frac{n}{2}$ bits is $2^{-\Omega(n)}$, regardless of the decisions made on the last $\frac{n}{2}$ bits.

Proof. Fix an arbitrary bit i . Define the random variables $Z_k = K \cdot (p_{i,k+1} - p_{i,k})$ to be the scaled change of the bit's marginal probability at an iteration k where bit i does a random step. Note that these variables take values in $\{-1, 0, 1\}$. The proof consists in bounding the maximum value that the (partial) sums of the Z_k 's can take over the first t random steps, using a "maximal" version of Chernoff's bound.

We thus define $Y_j := \sum_{k=1}^j Z_k$ as the total progress in the first $j \in \{1, 2, \dots, t\}$ random steps. Due to the symmetry and the fact that we start with the marginal probability equal to $\frac{1}{2}$, we can conclude $\mathbb{E}[Y_j] = 0$. Since the change in the probabilities is bounded by $\frac{1}{K}$, we need $Y_j \geq \frac{K}{6}$ in order to have a total change that exceeds $\frac{1}{6}$. Using the Chernoff-Hoeffding bounds from Theorems 1.11 and 1.13 in [4], we can bound the probability of this event happening during the first t random steps:

$$\Pr\left[\max_{j \in \{1, \dots, t\}} Y_j \geq \frac{K}{6}\right] \leq \exp\left(-2 \cdot \frac{(\frac{K}{6})^2}{\sum_{j=1}^t 2^j}\right) = \exp\left(-\frac{K^2}{72t}\right) \leq \exp\left(-\frac{1}{72\alpha}\right).$$

Leveraging the symmetry of this process, we can obtain the same bound for the probability of the total change going below $-\frac{1}{6}$

$$\Pr \left[\min_{j \in \{1, \dots, t\}} Y_j \leq -\frac{K}{6} \right] = \Pr \left[\max_{j \in \{1, \dots, t\}} -Y_j \geq \frac{K}{6} \right] \leq \exp \left(-\frac{1}{72\alpha} \right).$$

Now using the union bound, we can bound the probability that the total change stays within $[-\frac{1}{6}, \frac{1}{6}]$ the first t random steps, denoted by the $P^{\leq t}$:

$$\begin{aligned} P^{\leq t} &= 1 - \Pr \left[\left(\max_{j \in \{1, \dots, t\}} Y_j > \frac{K}{6} \right) \cup \left(\min_{j \in \{1, \dots, t\}} Y_j < -\frac{K}{6} \right) \right] \\ &\geq 1 - \Pr \left[\max_{j \in \{1, \dots, t\}} Y_j > \frac{K}{6} \right] - \Pr \left[\min_{j \in \{1, \dots, t\}} Y_j < -\frac{K}{6} \right] \\ &\geq 1 - \Pr \left[\max_{j \in \{1, \dots, t\}} Y_j \geq \frac{K}{6} \right] - \Pr \left[\min_{j \in \{1, \dots, t\}} Y_j \leq -\frac{K}{6} \right] \\ &\geq 1 - 2 \exp \left(-\frac{1}{72\alpha} \right). \end{aligned}$$

Choosing $\alpha < \frac{1}{72 \ln 4}$ we get $P^{\leq t} \geq \frac{1}{2} = \Omega(1)$, which proves the first part of the lemma.

Since by definition the random steps of different frequencies behave independently of each other, the second part of the lemma follows directly using Chernoff bounds. \square

In the next lemma, we put things together and show that there is a constant fraction of bits whose marginal probabilities stay bounded away from the boundaries (and hence receive few signal steps). The proof structure follows that of Lemma 15 in [28], and the idea is to bound the accumulated effect of signal steps using Chernoff bounds.

Lemma 14. *Let $K \geq 1$ and $\bar{p} \in (0, \frac{1}{2})$ be arbitrary, and consider the algorithm $CGA(K, \bar{p})$ on DYNBV. There exist constants $\alpha, \gamma > 0$, such that the following holds with high probability, regardless of the last $\frac{n}{2}$ bits (i.e., an adversary may choose the value of those bits in the offspring). There is a subset S of γn bits among the first $\frac{n}{2}$ bits such that during the first $t := \alpha K^2$ iterations:*

- i) the marginal probabilities of all bits in S always lie in the interval $[\frac{1}{6}, \frac{5}{6}]$;*
- ii) the total number of signal steps for each bit in S is bounded by $\frac{K}{6}$, leading to a displacement of at most $\frac{1}{6}$.*

Proof. Note that the first part holds at initialization. In order to prove the first part for all iterations, we will look at the probability that the property we want to prove gets violated. We will show that in each iteration, the property gets violated only if an event of probability exponentially small in K occurs. Taking the union bound over t iterations then shows that the property must hold throughout all of the t iterations.

If $\gamma > 0$ is a sufficiently small constant, it follows from the second part of Lemma 13 that with probability $1 - 2^{-\Omega(n)}$ at least $2\gamma n$ bits among the first $\frac{n}{2}$ ones experience a displacement from the random steps that lies within $[-\frac{1}{6}, \frac{1}{6}]$ during the first $t \leq \alpha K^2$ iterations. For the remainder of the proof, we will assume that this event takes place, and we will consider S' to be a set of $2\gamma n$ such bits.

To complete the proof, we distinguish two cases. Either $K \leq \log^2 n$. Then $t = o(n)$. Since at most one bit per iteration takes a signal step, the total number of bits in S' which ever take a signal step is at most $t \leq \gamma n$, for n sufficiently large. Hence, at least γn bits in S' never take a signal step, which completes this case.

For $K \geq \log^2 n$ we will show that for all bits in the set S' , with high probability, the total displacement caused by the signal steps is at most $\frac{1}{6}$, so we may take $S := S'$. According to Corollary 12, for every bit in S' , the probability of it taking a signal step in an iteration is bounded by $\frac{c}{n}$, where $c > 0$ is the constant from the lemma's O -notation. As long as the conditions for Corollary 12 hold, the expected number of signal steps of a fixed bit in t iterations will be at most $t \cdot \frac{c}{n} \leq \alpha K \cdot \min\{K, n\} \cdot \frac{c}{n} \leq \alpha c K$.

We know that each signal step changes the marginal probability of a bit by $\frac{1}{K}$, so a necessary condition for increasing a bit's marginal probability by an additional $\frac{1}{6}$ would be to take at least $\frac{K}{6}$ signal steps during the t iterations. By choosing the constant $\alpha > 0$ so small enough that $\alpha c K \leq \frac{1}{2} \cdot \frac{K}{6}$, Chernoff's bound ensures that the probability of getting at least $\frac{K}{6}$ signal steps in the first t iterations is at most $e^{-\Omega(K)}$. Thus, to violate the first property, an event of probability $e^{-\Omega(K)}$ has to take place for any bit in S' and any length of time, otherwise all properties hold.

Finally, we can take a union bound over all $2\gamma n$ bits in the set S' , as well as all t iterations, and obtain that the desired property gets violated with probability $t \cdot 2\gamma n \cdot e^{-\Omega(K)} \leq \alpha K^2 \cdot 2\gamma n \cdot e^{-\Omega(K)}$. Since $K \geq \log^2 n$, this failure probability is $o(1)$, and since the other property is a direct implication of Lemma 13, the proof is complete. \square

The main result of this section is then a direct consequence of the above lemma.

Theorem 15. *Let $K = O(\text{poly}(n))$ and $\bar{p} \in (0, \frac{1}{2})$ be arbitrary, and consider the algorithm $\text{cGA}(K, \bar{p})$ on DYNBV . Then with high probability the optimum is not sampled during the first $\Omega(K \cdot \min\{K, n\})$ iterations.*

Proof. Let $\alpha > 0$ be the constant from Lemma 14. By this lemma, the probability of sampling the optimum in one given iteration among the first $\alpha K \cdot \min\{K, n\}$ iterations is at most $2 \cdot \left(\frac{5}{6}\right)^{\gamma n} = e^{-\Omega(n)}$, since all γn bits with marginal probability in the interval $[\frac{1}{6}, \frac{5}{6}]$ have to be set to 1 (and there are two offspring in each iteration). Taking a union bound over these $\Omega(K \cdot \min\{K, n\})$ iterations and adding the error probability from Lemma 14 concludes the proof. \square

We are also in a position to show the existence of genetic drift. For the convenience of the reader, we restate Theorem 2 below before finishing its proof.

Theorem 16. *Let $K = \omega(1)$ and $\bar{p} \in (0, \frac{1}{2})$ be arbitrary, and consider the algorithm $\text{CGA}(K, \bar{p})$ on DYNBV . For every $\rho > 0$ and $\beta \in (\bar{p}, \frac{1}{2})$ there is $\delta > 0$ such that the following holds. Assuming that $K \leq \rho n$, then with high probability at least δn frequencies drop below β during optimization.*

Proof. We begin by establishing that the sample variance V_t remains of linear order for our period of interest. This is again due to a linear number of bits which stay in middle interval around their initialization values. Concretely, by Lemma 14, for some small constants $\alpha, \gamma > 0$, with high probability there exists a subset of γn bits from the first $\frac{n}{2}$ such that the marginal probabilities of these bits remain in the interval $[\frac{1}{6}, \frac{5}{6}]$ for the first αK^2 iterations. In particular, this implies that $V_t = \Omega(n)$ for all $t \leq \alpha K^2$.

Consider now the other $\frac{n}{2}$ bits, denoting them by $D_0 \subseteq [n]$. By Corollary 12, each of them has probability of $O(\frac{1}{n})$ to take a signal step during these αK^2 iterations. This implies that the expected number of signal steps taken by one of these bits during this timespan is $\alpha K^2 \cdot O(\frac{1}{n}) \leq O(K)$, where the inequality holds since $K \leq \rho n$. Hence the number of signal steps taken by a given bit in D_0 is stochastically dominated by a binomial random variable $\text{Bin}(\alpha K^2, O(\frac{1}{n}))$.

Chernoff bounds therefore imply the existence of a constant $\eta > 0$ such that a bit in D_0 takes at most ηK signal steps during the first αK^2 iterations with constant probability.

Let us now focus on the subset $D_1 \subseteq D_0$ of bits that have taken at most ηK signal steps throughout the first αK^2 iterations. Since the probability that a bit in D_0 is also in D_1 is constant, we get by a similar argument (involving Chernoff bounds and stochastic dominance), that $|D_1| = \Omega(n)$ with high probability. Note that all the bits in D_1 make at most ηK signal steps during the αK^2 iterations, and as long as their marginal probabilities are in the interval $[\beta, 1 - \beta]$, the probability that a given iteration is a random step is lower bounded by a constant. Therefore, since $K = \omega(1)$ (and hence $K = o(K^2)$), we can again use Chernoff bounds to deduce that the number of random steps they make is at least εK^2 with constant probability for some small constant $\varepsilon > 0$. Now let $D_2 \subseteq D_1$ denote the set of bits from D_1 that have taken at least εK^2 random steps throughout the first αK^2 iterations. By Chernoff, we know that $|D_2| = \Omega(n)$ with high probability.

Consider now the effect these random steps will have on the marginal probability of a bit in D_2 . To this effect, we will inspect some properties of unbiased random walks with step size 1. Let X_N denote the endpoint of such an unbiased random walk with N steps. Then, by the Central Limit Theorem, for every $c \in \mathbb{R}$ it holds that

$$\Pr[X_N \leq c\sqrt{N}] \rightarrow \Phi(c), \text{ as } N \rightarrow \infty,$$

where Φ is the cumulative distribution function of the standard normal distribution. Let us instantiate $N := \varepsilon K^2$ and $c := \frac{-1/2 + \beta - \eta}{\sqrt{\varepsilon}}$. Observe that $N \rightarrow \infty$

is equivalent to $K \rightarrow \infty$, hence for K large enough

$$\Pr[X_N \leq (-\frac{1}{2} + \beta - \eta)K] = \Pr[X_N \leq c\sqrt{N}] \geq \frac{\Phi(c)}{2}.$$

We return now to our setting of εK^2 random steps of size $\frac{1}{K}$. Until this random walk reaches distance $\frac{1}{2} - \beta + \eta$ from its starting point, it behaves exactly like a rescaled version of the unbiased random walk.

Moreover, by symmetry both endpoints (namely $\frac{1}{2} - \beta + \eta$ and $-\frac{1}{2} + \beta - \eta$) have the same probability of having been reached first, i.e. the probability to end at $-\frac{1}{2} + \beta - \eta$ is $\frac{1}{2}$.

Hence, the probability that these εK^2 random steps will cause a total (negative) displacement of at least $\frac{1}{2} - \beta + \eta$ from the starting point is at least $\frac{\Phi(c)}{4}$.

Thus, for any bit in D_2 , we have that the cumulative effect of the random and signal steps results in a marginal probability of at most $\frac{1}{2} + \eta - (\frac{1}{2} - \beta + \eta) = \beta$ with constant probability. Consider $D_3 \subseteq D_2$ to be the set of bits for which this holds true. We deduce that $|D_3| \geq \delta n$ with high probability using Chernoff bounds, for some constant $\delta > 0$ chosen small enough, which concludes the proof. \square

5 Upper bound on the runtime for the conservative regime

This section is dedicated to showing an upper bound on the runtime for population sizes $K = \Omega(n \log n)$ that matches Theorem 15. To facilitate the proof, we slightly adjust the boundaries by setting $\bar{p} = \frac{1}{cn}$ for c large enough. Precisely, our goal will be to show the following result.

Theorem 17. *Let $K = \text{poly}(n)$ and let $c, c' > 0$ be sufficiently large constants. If $K \geq c' \cdot n \log n$, then the expected optimization time of the algorithm $\text{cGA}(K, \frac{1}{cn})$ on DYNBV is $O(Kn)$.*

We begin with a proposition that is the analogue of Lemma 4 in [28]. Intuitively, it states that for $K = \Omega(n \log n)$, the marginal probability of any bit above at least some constant threshold β will not drop below a fixed constant $0 < \alpha < \beta$ in the next polynomially many iterations. Even more succinctly, even moderately high marginal probabilities do not drop by much for a while.

Proposition 18. *Let $\bar{p} \in (0, \frac{1}{2})$ be arbitrary and let $K \geq 1$, and consider the algorithm $\text{cGA}(K, \bar{p})$ on DYNBV . Let $\bar{p} < \alpha < \beta < 1 - \bar{p}$ and $\gamma > 0$ be constants. There exists a constant $c' > 0$ (possibly depending on α, β , and γ) such that for a specific bit the following holds: If the bit has marginal probability at least β and $K \geq c' \cdot n \log n$, then the probability that during the following n^γ iterations the marginal probability decreases below α is at most $O(n^{-\gamma})$.*

Proof. The goal of this proof is to apply the negative drift theorem, Theorem 5, so it will be rather technical, with the bulk of the task consisting of setting the parameters and verifying the three conditions of the negative drift theorem. To this extent, we first fix an arbitrary frequency i and consider the scaled stochastic process $X_t := Kp_{i,t}$.

To apply Theorem 5, we choose the interval bounds $a := K\alpha$ and $b := K\beta$, which yields $l := b - a = K(\beta - \alpha)$. To establish the first condition of the theorem, we scale the result from Proposition 11 by K and obtain the following bound on the drift (thus, also fixing $\varepsilon := 4\xi\alpha(1 - \beta)/n$ in the meantime

$$\begin{aligned} \mathbb{E}[X_{t+1} - X_t \mid \mathcal{F}_t; a < X_t < b] &\geq K \cdot \xi \cdot \frac{p_{i,t}(1 - p_{i,t})}{KV_t} \\ &\geq \xi \frac{4\alpha(1 - \beta)}{n}, \end{aligned}$$

where $\xi > 0$ is the constant from the Θ -notation in Proposition 11, and the final bound was obtained by observing that $V_t = \sum_{j=1}^n p_{j,t}(1 - p_{j,t}) \leq \sum_{j=1}^n \frac{1}{4} = \frac{n}{4}$.

The second condition of the theorem can be trivially fulfilled by choosing $r := 2$. We note that for $j = 0$, the condition holds because $e^{-0} = 1$ by definition and the left-hand side is a probability, hence again by definition upper-bounded by 1 - and for $j \geq 1$ we have that $\Pr[|X_{t+1} - X_t| \geq 2j] = 0 \leq e^{-j}$, because $|X_{t+1} - X_t| \leq 1$ in the scaled process, as the marginal frequencies change by at most $\frac{1}{K}$.

We verify the last condition of the theorem using the fact that $K \geq c \cdot n \log n$:

$$\begin{aligned} \frac{\varepsilon l}{132 \log(r/\varepsilon)} &= \frac{4\xi\alpha(1 - \beta)}{n} \cdot K(\beta - \alpha) \cdot \frac{1}{132} (\log n - \log(2\xi\alpha(1 - \beta)))^{-1} \\ &\geq \frac{\xi\alpha(1 - \beta)(\beta - \alpha)}{33} \cdot c \cdot \frac{\log n}{\log n - \log(2\xi\alpha(1 - \beta))}. \end{aligned}$$

As all $\xi, \alpha, \beta > 0$ are constants and $(1 - \beta)(\beta - \alpha)$ is positive since $1 > \beta > \alpha$, the term above is positive for large enough n . From the limiting value $\lim_{n \rightarrow \infty} \frac{\log n}{\log n - \log(2\xi\alpha(1 - \beta))} = 1$, we can conclude that there exists a constant $c > 0$ for which the term above is greater or equal to $4 = r^2$.

Furthermore, note that the constant $c > 0$ in $K \geq c \cdot n \log n$ can further be increased to ensure that the following inequality holds

$$\begin{aligned} \frac{\varepsilon l}{132r^2} &= \frac{4\xi\alpha(1 - \beta) \cdot K(\beta - \alpha)}{132 \cdot 4 \cdot n} \\ &\geq \frac{\xi\alpha(1 - \beta)(\beta - \alpha)}{132} \cdot c \cdot \log n \\ &\geq \gamma \ln n, \end{aligned}$$

for an arbitrarily large constant γ . This is possible because all other quantities, that is, ξ, β, α do not depend on n and have been fixed beforehand.

Now we are ready to conclude. Since, by assumption, we started with $X_0 \geq b$, we can establish via the negative drift theorem that $\Pr[T \leq n^\gamma] = O(n^{-\gamma})$, where T is the first time when the fixed marginal frequency drops below α . \square

The next lemma, which is modeled after Lemma 2 in [24], gives an upper bound on the time until the marginal probability reaches any threshold τ , irrespective of where the marginal probability starts. In the proof of the main theorem of this section, this lemma will allow us to upper-bound the "recovery time" needed for bits whose marginal probabilities travel to the lower boundary.

Lemma 19. *Let $c > 0$ be a constant and let $K \geq 1$. Let $\tau \in [\frac{1}{cn}, 1 - \frac{1}{cn}]$ and consider some fixed bit i in the cGA($K, \frac{1}{cn}$) on DYNBV with an arbitrary value for the initial marginal probability. Then the expected time until the marginal probability p_i of this bit reaches at least $p_i = \tau$ is $O(Kn^2)$.*

Proof. Let us inspect the expected change in probability for the marginal frequency of bit i at an arbitrary iteration t for which $p_{i,t} < \tau$. From Proposition 11 we can fix $\xi > 0$ as the constant in the Θ -notation and write:

$$\begin{aligned} \mathbb{E}[p_{i,t+1} \mid p_{i,t}] &\geq p_{i,t} + \xi \cdot \frac{p_{i,t}(1-p_{i,t})}{KV_t} \\ &\geq p_{i,t} + \xi \cdot \frac{1}{cn} \left(1 - \frac{1}{cn}\right) \cdot \frac{1}{K} \cdot \frac{4}{n} \\ &= p_{i,t} + \Omega\left(\frac{1}{Kn^2}\right), \end{aligned}$$

where the second inequality stems from the fact that the marginal frequencies are capped at $\frac{1}{cn}$ and $1 - \frac{1}{cn}$ at the boundaries, and $V_t = \sum_{j=1}^n p_{j,t}(1-p_{j,t}) \leq \sum_{j=1}^n \frac{1}{4} = \frac{n}{4}$.

The claim now follows by applying Theorem 7 with drift bound $\delta = \Omega\left(\frac{1}{Kn^2}\right)$, threshold $\kappa = 1 - \tau$, and distance function g with $g(x) = 1 - x$. \square

We are now ready to prove the main theorem of this section. Its proof follows the structure of the proof of Theorem 2 in [28].

Proof (of Theorem 17). Recall that we assumed without loss of generality that $\frac{1}{K}$ "divides" $\frac{1}{2} - \frac{1}{cn}$, which implies that the marginal frequencies are restricted to the set

$$\left\{ \frac{1}{cn}, \frac{1}{cn} + \frac{1}{K}, \dots, \frac{1}{2}, \dots, 1 - \frac{1}{cn} - \frac{1}{K}, 1 - \frac{1}{cn} \right\}.$$

The structure of this proof will follow the one of Theorem 2 in [28]. We will show that starting with a state where all frequencies are at least $\frac{1}{2}$ at the same time, with probability $\Omega(1)$, after $O(Kn)$ iterations either the global optimum has been found, which we call *success*, or at least one of the frequencies has dropped below some arbitrary constant $\eta < \frac{1}{2}$, which we call *failure*. We do not instantiate the constant η . Concretely, if a failure occurred, then we will prove that one can recover from it - without too big of an overhead in expectation. Recovering means that we return to a state where all marginal frequencies are at least $\frac{1}{2}$. The expected time until either a success or a failure takes place is $O(Kn)$. Then, if no failure occurs during $O(Kn)$ iterations, we show that with probability $\Omega(1)$ the optimum has been found.

Firstly, let us prove that recovering from a failure is not too costly. Choose a constant $\gamma > 0$ such that $n^\gamma \geq K \cdot n^5$ (such a constant exists since $K = \text{poly}(n)$). By Proposition 18, instantiated with $\alpha = \eta < \frac{1}{2}$ and $\beta = \frac{1}{2}$, the probability of a failure in n^γ iterations is at most $O(n^{-\gamma})$, provided the constant c in $K \geq c \cdot n \log n$ is large enough. From Lemma 19 we can conclude that the expected time for the marginal probability of a specific bit to reach the upper boundary is always bounded by $O(Kn^2)$, regardless of the initial state. We note that the expected time until all probabilities have reached the upper boundary at least once is in $O(Kn^3 \log n)$, as it is stochastically dominated by a simple modification of the coupon collector problem: We split the time into phases of length $\kappa \cdot Kn^2$ for some constant $\kappa > 0$ that is large enough. In each of these phases, by Markov's inequality and Lemma 19, each marginal probability travels to the upper boundary with probability at least $p > 0$, where p is a constant. Therefore, by the coupon collector problem, we need in expectation $O(n \log n)$ such phases for all bits to reach the upper boundary at least once. Once a bit reaches the upper boundary, we apply Proposition 18 with $\alpha = \frac{1}{2}$ and $\beta = \frac{2}{3}$ to get that the probability of a marginal frequency decreasing below $\frac{1}{2}$ in time n^γ is at most $O(n^{-\gamma})$. By the union bound, the probability that there is a bit for which this happens is at most $O(n^{-\gamma+1})$. If this does not happen, all bits assume values of at least $\frac{1}{2}$ simultaneously, and we can say that we successfully recovered from a failure.

As the probability of a bit falling below $\frac{1}{2}$ during the recovery phase is at most $n^{-\gamma+1}$, the expected number of restarts of this recovery phase is $O(n^{-\gamma+1})$. Therefore, considering the expected time until all bits recover to values that are at least $\frac{1}{2}$ only adds an additional term of $O(n^{-\gamma+1} \cdot Kn^3 \log n) = o(1)$ to the expectation (recall that $n^\gamma \geq Kn^5$).

All that remains to show now is that after $O(Kn)$ iterations without failure the probability of having found the optimum is in $\Omega(1)$. To this end we will consider a suitable potential function, and apply the variable drift theorem (see [19]). We will apply the variable drift theorem as long as the sampling variance stays above an arbitrary constant threshold. Since the variance will be a lower bound on the potential, this will imply that the potential also assumes values above a certain threshold. In the subsequent, we will show that once the variance drops below a certain constant threshold, the potential will as well (as we are in the case where all marginal probabilities are at least η), and we will sample the optimum with at least constant probability.

The potential function we consider is $\varphi_t := n - \frac{1}{c} - \sum_{i=1}^n p_{i,t}$, which measures the distance to the ideal setting where all frequencies lie at the upper boundary. Recall here that we are inspecting the modified version of the cGA where the upper boundary is at $1 - \frac{1}{cn}$.

Assume for now that $V_t \geq 1$ holds for the sampling variance V_t . This assumption is helpful, as it will allow us to get asymptotic bounds on the drift whose constants do not depend on c (see Proposition 11).

We will now inspect the drift of a fixed bit $i \in \{1, \dots, n\}$ by distinguishing two cases based on whether the frequency is at the upper boundary or not:

- I. $p_{i,t} > 1 - \frac{1}{cn} - \frac{1}{K}$: By our assumption that $\frac{1}{K}$ “divides” $\frac{1}{2} - \frac{1}{cn}$, we conclude that $p_{i,t} = 1 - \frac{1}{cn}$, therefore the frequency of bit i can only decrease. A decrease by $\frac{1}{K}$ happens with probability at most $\frac{1}{cn}$: We have to sample a 0 on bit i in the “losing” offspring, and a 1 in the “winning” one.⁷ This happens with probability $p_{i,t}(1 - p_{i,t}) \leq \frac{1}{cn}$. Thus:

$$\begin{aligned} \mathbb{E}\left[1 - \frac{1}{cn} - p_{i,t+1} \mid p_{i,t}\right] &\leq \left(1 - \frac{1}{cn} - p_{i,t}\right) + \frac{1}{Kcn} \\ &= \frac{1}{Kcn}. \end{aligned}$$

- II. $p_{i,t} \leq 1 - \frac{1}{cn} - \frac{1}{K}$: In this case, Proposition 11 becomes applicable and we obtain:

$$\mathbb{E}\left[1 - \frac{1}{cn} - p_{i,t+1} \mid p_{i,t}\right] \leq \left(1 - \frac{1}{cn} - p_{i,t}\right) - \xi \cdot \frac{p_{i,t}(1 - p_{i,t})}{KV_t},$$

where $\xi > 0$ is the constant from Proposition 11 hidden by the Θ -notation and does not depend on c' .

The sampling variance V_t can be bounded from above by:

$$V_t = \sum_{j=1}^n p_{j,t}(1 - p_{j,t}) \leq \sum_{j=1}^n (1 - p_{j,t}) = \varphi_t + \frac{1}{c}.$$

Given that we assumed to be in the situation where no failure occurred, we may assume that $p_{i,t} \geq \eta$ and $1 - p_{i,t} \geq 1 - \frac{1}{cn} - p_{i,t}$. This yields the following bound on the drift:

$$\begin{aligned} \mathbb{E}\left[1 - \frac{1}{cn} - p_{i,t+1} \mid p_{i,t}\right] &\leq \left(1 - \frac{1}{cn} - p_{i,t}\right) - \xi \cdot \frac{\eta(1 - p_{i,t})}{K(\varphi_t + 1/c)} \\ &\leq \left(1 - \frac{1}{cn} - p_{i,t}\right) \cdot \left(1 - \frac{\xi\eta}{K(\varphi_t + 1/c)}\right). \end{aligned}$$

This case distinction now allows us to write the expression for the total drift. Let therefore $B_t \subseteq \{1, 2, \dots, n\}$ be the set of the indices of the bits that are at

⁷ I.e. the less fit respectively fitter offspring.

the upper boundary at iteration t . The expected change in potential is:

$$\begin{aligned}
\mathbb{E}[\varphi_{t+1} \mid \varphi_t] &= \sum_{i \in B_t} \mathbb{E}\left[1 - \frac{1}{cn} - p_{i,t+1} \mid p_{i,t}\right] + \sum_{i \notin B_t} \mathbb{E}\left[1 - \frac{1}{cn} - p_{i,t+1} \mid p_{i,t}\right] \\
&\leq \sum_{i \in B_t} \frac{1}{Kcn} + \sum_{i \notin B_t} \left(1 - \frac{1}{cn} - p_{i,t}\right) \cdot \left(1 - \frac{\xi\eta}{K(\varphi_t + 1/c)}\right) \\
&= \frac{|B_t|}{Kcn} + \left(1 - \frac{\xi\eta}{K(\varphi_t + 1/c)}\right) \cdot \sum_{i \notin B_t} \left(1 - \frac{1}{cn} - p_{i,t}\right) \\
&\leq \frac{1}{Kc} + \left(1 - \frac{\xi\eta}{K(\varphi_t + 1/c)}\right) \cdot \sum_{i \notin B_t} \left(1 - \frac{1}{cn} - p_{i,t}\right) \\
&\leq \frac{1}{Kc} + \left(1 - \frac{\xi\eta}{K(\varphi_t + 1/c)}\right) \cdot \sum_{i=1}^n \left(1 - \frac{1}{cn} - p_{i,t}\right) \\
&= \frac{1}{Kc} + \left(1 - \frac{\xi\eta}{K(\varphi_t + 1/c)}\right) \cdot \varphi_t \\
&= \varphi_t + \frac{1}{K} \cdot \left(\frac{1}{c} - \frac{\xi\eta \cdot \varphi_t}{\varphi_t + 1/c}\right),
\end{aligned} \tag{7}$$

where the second inequality holds because $|B_t| \leq n$, and the third one holds due to the fact that the frequencies are capped at $1 - \frac{1}{cn}$, and, thus, we have $1 - \frac{1}{cn} - p_{i,t} = 0$ for $i \in B_t$.

We will now prove that for any given $\varepsilon > 0$, we can fix the constant $c > 0$ in such a way that $\frac{\xi\eta\varphi_t}{\varphi_t + 1/c} \geq (1 + \varepsilon) \cdot \frac{1}{c}$ holds for all large enough φ_t . Intuitively, this should hold, because once $\varepsilon > 0$ is fixed, $\lim_{\varphi_t \rightarrow \infty} \frac{\varphi_t}{\varphi_t + 1/c} = 1$, both ξ and η are constants that are already fixed, and then c' can be adapted appropriately.

In particular, we may choose $c = \max\left\{1, \frac{2(1+\varepsilon)}{\xi\eta}\right\}$. Then, for all $\varphi_t \geq V_t \geq 1$ we obtain

$$\begin{aligned}
(1 + \varepsilon) \cdot \frac{1}{c} &= \min\left\{\frac{(1 + \varepsilon)\xi\eta}{2(1 + \varepsilon)}, 1 + \varepsilon\right\} \\
&\leq \xi\eta \cdot \frac{1}{2} \\
&\leq \xi\eta \cdot \frac{\varphi_t}{\varphi_t + 1} \\
&\leq \xi\eta \cdot \frac{\varphi_t}{\varphi_t + 1/c},
\end{aligned}$$

where in the second inequality we used that $f(x) = \frac{x}{x+1}$ is increasing on $[1, \infty)$ and lower-bounded by $f(1) = \frac{1}{2}$ and in the last inequality our choice of c .

Let us now fix an arbitrary $\varepsilon > 0$, and take $c = \max\left\{1, \frac{2(1+\varepsilon)}{\xi\eta}\right\}$. We know from our previous derivation that for $\varphi_t \geq 1$ it holds that $\frac{\xi\eta\varphi_t}{\varphi_t + 1/c} \geq (1 + \varepsilon) \cdot \frac{1}{c}$.

Thus, we can continue bounding Equation (7) for $\varphi_t \geq 1$ by:

$$\begin{aligned} \mathbb{E}[\varphi_{t+1} \mid \varphi_t] &\leq \varphi_t + \frac{1}{K} \cdot \left(\frac{1}{c} - \frac{\xi\eta \cdot \varphi_t}{\varphi_t + 1/c} \right) \\ &\leq \varphi_t + \frac{1}{K} \cdot \left(\frac{1}{c} - (1 + \varepsilon) \cdot \frac{1}{c} \right) \\ &= \varphi_t - \frac{1}{K} \cdot \frac{\varepsilon}{c}. \end{aligned}$$

In other words, we have obtained that $\mathbb{E}[\varphi_t - \varphi_{t+1} \mid \varphi_t] \geq \frac{\varepsilon}{c} \cdot \frac{1}{K}$, if $\varphi_t \geq 1$.

Recall that for the previous bounds on the drift we have assumed $V_t \geq 1$ (which also implied $\varphi_t \geq 1$). If, however, we are in the case that $V_t < 1$, we can leverage the fact that we are in a series of iterations without failure. In particular, this means that all the frequencies are at least equal to the constant η , which also implies that there is a constant upper bound $C \geq 1$ on the potential. To see the latter, recall that $V_t = \sum_{j=1}^n p_{j,t}(1 - p_{j,t})$. Using $p_{i,t} \geq \eta$ and $V_t < 1$, we get

$$1 > V_t = \sum_{j=1}^n p_{j,t}(1 - p_{j,t}) \geq \sum_{j=1}^n \eta(1 - p_{j,t}) = \eta(\varphi_t + \frac{1}{c}).$$

Therefore any constant $C > \frac{1}{\eta} - \frac{1}{c}$ works.

We now apply the variable drift theorem, Theorem 6, to bound the expected time for the potential φ to decrease from any initial value $\varphi \leq n$ to a value $\varphi \leq C$. For this, we will use the (constant) drift function $h(\varphi) = \frac{\varepsilon}{Kc}$, since we have established in the paragraphs above that it represents a lower bound on the expected change if the potential is at least $C \geq 1$.

While the variable drift theorem only considers the hitting time of 0, we note that it is still acceptable to use this theorem in our case: The process we consider instead is one where all the states with potential $\varphi \in [0, C)$ are merged into a single “state 0”. In this modified process, the smallest state larger than 0 is $x_{\min} = C$. This modification can only increase the drift (all iterations that previously reduced a potential above C to one below C now reduce the potential directly to 0 instead), so the drift of this process is still bounded below by $h(\varphi)$ for all states with $\varphi \geq C$.

Thus, by the variable drift theorem, the expected time until we hit a state with potential 0 in our modified process is:

$$\frac{C}{h(C)} + \int_C^n \frac{1}{h(x)} dx = \frac{Kc}{\varepsilon} + \int_C^n \frac{Kc}{\varepsilon} dx = O(Kn).$$

In order to end our proof, consider an iteration where $\varphi \leq C$. Given that we assumed we are not in a failure scenario, all the marginal frequencies of the bits attain values that are at least η , (recall that $\eta < \frac{1}{2}$ was a fixed constant). The probability of sampling a value of 1 on all bits simultaneously in this case is minimal in the extreme setting where a maximal number of bits have marginal probabilities equal to η , and all other bits (except at most one) have marginal

probabilities at their upper boundaries. Then, the probability of sampling the optimum in one iteration is at least $(1 - \frac{1}{cn})^n \cdot \eta^{\lceil \frac{c}{1-c} \rceil} = \Omega(1)$. Therefore, in a successful phase the optimum is found with at least constant probability. \square

6 Upper bound on the runtime for the aggressive regime

In this section, we analyze the cGA when $K = \Theta(\log^2 n)$ and prove Theorem 4. Throughout the whole section we will consider a capping probability of $\bar{p} := \frac{1}{n \log^c n}$ for some constant $c > 0$, but our simulations indicate that the result should also hold for the classical $\frac{1}{n}$ capping probability. We prove our main theorem in this regime for $c = 7$, and the proof would go through for any $c \geq 7$. It would be possible to reduce this constant, but we aimed for simpler proofs and have not tried to optimize it.

As in the previous section, we assume that $K \cdot (\frac{1}{2} - \frac{1}{n \log^c n})$ is an integer, so that the marginal frequencies are always in the set

$$\left\{ \frac{1}{n \log^c n}, \frac{1}{n \log^c n} + \frac{1}{K}, \dots, \frac{1}{2}, \dots, 1 - \frac{1}{n \log^c n} - \frac{1}{K}, 1 - \frac{1}{n \log^c n} \right\}.$$

The proof of Theorem 4 proceeds in four main steps.

First, we will show that due to the high genetic drift, the frequencies essentially start by executing random walks until they reach one of the boundaries. As a consequence, the sampling variance V_t drops from $\Theta(n)$ to $O(\log n)$ during the first $O(\text{polylog}(n))$ iterations, and then stays below $O(\log n)$ for the remainder of the optimization time with high probability. We call this initial phase the *burn-in phase*.

Then, given this bound on the sampling variance $V_t = O(\log n)$, we prove that it is unlikely to have frequencies at the upper boundary drop below a constant. This basically ensures that, while frequencies from the lower boundary may reach the upper boundary, the converse does not happen.

Using standard arguments on random walks and geometric distributions, we then argue that indeed all the frequencies starting from the lower boundary after the burn-in phase reach the upper boundary (at least once) within $O(n \cdot \text{polylog}(n))$ iterations.

This puts us in a situation where $n - O(\text{polylog}(n))$ frequencies are at the upper boundary, and the remaining $O(\text{polylog}(n))$ frequencies are lower bounded by a constant. We refine the analysis for this situation to show that with high probability all $O(\text{polylog}(n))$ frequencies reach the upper boundary while no frequency detaches from the upper boundary, and this process only takes $O(\text{polylog}(n))$ iterations. Hence we finally reach a state where $p_{i,t} = 1 - \frac{1}{n \log^c n}$ for all positions i , from which the optimum is then sampled with high probability in a single iteration, and that terminates the algorithm.

The first proposition covers the initial burn-in phase, and shows that afterwards the sampling variance stays low for at least a quadratic number of

iterations (which suffices for our purposes since we will prove that whp the algorithm terminates in quasi-linear time). We first require a lemma that bounds the time until a given frequency reaches one of the boundaries.

Lemma 20 (adapted from Lemma 6 in [23]). *Let $c > 0$ be a constant and $K = \omega(1)$, and consider the frequency $p_{i,t}$ of a bit i of the algorithm $\text{cGA}(K, \frac{1}{n \log^c n})$ on DYNBV . Let T denote the first time that $p_{i,t}$ reaches one of the boundaries. Then for every initial value $p_{i,0}$ and all $r \geq 8$, $\mathbb{E}[T \mid p_{i,0}] \leq 4K^2 \ln K$ and $\Pr[T \geq rK^2 \ln K \mid p_{i,0}] \leq 2^{-\lfloor r/8 \rfloor}$.*

Proof. This was proven in [23, Lemma 6]. Note that even though the lemma is stated for the case of the ONEMAX function, the proof does not employ any properties of the function to optimize, but rather gives a general statement for any random walk that has a certain probability of being stationary in each state. This result thus also applies to the DYNAMIC BINVAL function, since by Proposition 11 the first display equation in the proof of Lemma 6 in [23] is satisfied in our case as well. \square

Proposition 21. *For $K = \Theta(\log^2 n)$ consider the algorithm $\text{cGA}(K, \frac{1}{n \log^7 n})$ on DYNBV . After the first $O(K^3 \log n)$ iterations, with high probability the sampling variance V_t will stay below $O(\log n)$ for at least n^2 consecutive iterations.*

Proof. Let us divide the optimization time into phases of length $\kappa \cdot K^3 \log n$, where $\kappa > 0$ is a constant that will be fixed later. The idea of the proof is to show that with high probability, during a single such phase, all the frequencies that were not at the boundaries at the start of the phase will return to one of the boundaries, and the number of frequencies that detach from the boundaries during that same phase is in $O(\log n)$. Then, this means that with the exception of the first phase, all other phases will have at most $O(\log n)$ frequencies that are not at any boundary, which implies that the variance will stay within $O(\log n)$ for the rest of the optimization time. This follows from the fact that the variance of any frequency at the boundary is $\frac{1}{n \log^7 n} \left(1 - \frac{1}{n \log^7 n}\right) < \frac{1}{n \log^7 n}$, and the variance of a frequency not at a boundary is at most $\frac{1}{2} \cdot \left(1 - \frac{1}{2}\right) = \frac{1}{4}$. Thus, an upper bound on the variance is $n \cdot \frac{1}{n \log^7 n} + O(\log n) \cdot \frac{1}{4} = O(\log n)$.

Consider an arbitrary phase. Our goal is to show that all the bits that are not at a boundary at the start of the phase will eventually reach one of the boundaries by the end of the $\kappa \cdot K^3 \log n$ iterations with high probability. Then, we can conclude the proof by showing that the number of frequencies that detach from a boundary during this same phase is within $C \cdot \log n$, where $C > 0$ is a constant that will be fixed later.

To prove the first part, let us fix an arbitrary $i \in \{1, 2, \dots, n\}$ and assume that the current phase we are interested in has started at some iteration $t > 0$. Then, it follows from Lemma 20 that with probability at least $\frac{1}{2}$, there will be an iteration $t' \in [t, t + 8 \cdot K^3]$ for which $p_{i,t'} \in \left\{ \frac{1}{n \log^7 n}, 1 - \frac{1}{n \log^7 n} \right\}$. By segmenting our phase of $\kappa \cdot K^3 \log n$ iterations into intervals of length $8 \cdot K^3$, we obtain that

the probability frequency i never reaching one of the boundaries within the phase is at most $(1 - \frac{1}{2})^{\lfloor \kappa \cdot K^3 \log n / (8 \cdot K^3) \rfloor} = (\frac{1}{2})^{\lfloor \kappa \log n / 8 \rfloor}$. By choosing $\kappa := 40$ this number is at most $n^{-3.4}$ for sufficiently large n ,

and via a union bound over the first n^2 phases and n frequencies, with high probability all bits that start any phase off-boundary will return to one of the boundaries within $\kappa \cdot K^3 \log n$ iterations.

What is now left to show is that there exists a constant $C > 0$ such that at most $C \cdot \log n$ frequencies detach from the boundaries during any phase of $\kappa \cdot K^3 \log n$ iterations with high probability.

Let the random variable N denote the number of bits that detach from any of the two boundaries during the $\kappa K^3 \log n$ iterations of the phase. N is then stochastically dominated by a sum of $n \cdot \kappa K^3 \log n$ independent Bernoulli-distributed random variables with parameter $\frac{2}{n \log^\gamma n}$. This is because in each of the iterations, at most n bits can detach, and the probability of detaching from a boundary is bounded by $\frac{2}{n \log^\gamma n}$: A frequency can only detach if the two bits in the offspring differ, and for a frequency at a boundary, this happens with probability at most $2 \cdot \frac{1}{n \log^\gamma n} \cdot \left(1 - \frac{1}{n \log^\gamma n}\right) < \frac{2}{n \log^\gamma n}$.

Hence, in expectation at most $\mathbb{E}[N] \leq 2 \cdot \frac{n \cdot \kappa K^3 \log n}{n \log^\gamma n} = 2\kappa \cdot \frac{K^3}{\log^6 n} = O(1)$ frequencies detach during the phase, where we used $K = \Theta(\log^2 n)$. Since N is stochastically dominated by a binomial distribution $\text{Bin}\left(n \cdot \kappa K^3 \log n, \frac{2}{n \log^\gamma n}\right)$, we can employ Chernoff bounds to get $\Pr[N \geq C \log n] = e^{-C \cdot \Omega(\log n)}$. As in the first part of this proof, we can now choose a suitable C that will allow us to perform a union bound over all phases. Hence, with high probability, no phase has more than $C \cdot \log n = O(\log n)$ bits detaching from any boundary during the entire optimization time. \square

In the next lemma, we show that frequencies that reach the upper boundary stay above a constant for the following n^γ iterations with high probability, where γ is a constant that we are free to choose (we will later pick $\gamma = 2$). The proof is a fairly straightforward application of the negative drift theorem (Theorem 5).

Lemma 22. *Let $\bar{p} \in (0, \frac{1}{2})$ be arbitrary and let $K \geq 1$, and consider the algorithm $\text{cGA}(K, \bar{p})$ on DYNBV . Let $\bar{p} < \alpha < \beta < 1 - \bar{p}$ and $\gamma > 0$ be constants. Assume that $V_t = O(\log n)$ holds for the variance all throughout the optimization time. Then there exists a constant $c' > 0$ (possibly depending on α , β , and γ) such that for a specific bit the following holds: If the bit has marginal probability at least β and $K \geq c' \cdot \log^2 n$, then the probability that during the following n^γ iterations the marginal probability decreases below α is at most $O(n^{-\gamma})$.*

Proof. The proof of this lemma follows very closely that of Proposition 18. The only difference is in the lower bound on the drift, where we here use our stronger bound $V_t = O(\log n)$, while the proof of Proposition 18 uses the general bound $V_t \leq \frac{n}{4}$.

We first fix an arbitrary frequency i and look at the scaled stochastic process $X_t := K p_{i,t}$.

In order to apply Theorem 5, we choose the interval bounds $a := K\alpha$ and $b := K\beta$, and therefore also get $l := b - a = K(\beta - \alpha)$. To establish the first condition of the theorem, we scale the result from Proposition 11 by K and obtain the following bound on the drift (thus, also fixing $\varepsilon := \xi\alpha(1 - \beta)$ in the meantime):

$$\begin{aligned} \mathbb{E}[X_{t+1} - X_t \mid \mathcal{F}_t; a < X_t < b] &\geq K \cdot \xi \cdot \frac{p_{i,t}(1 - p_{i,t})}{K \cdot \max\{1, V_t\}} \\ &\geq \frac{\xi\alpha(1 - \beta)}{C \log n}, \end{aligned}$$

where $\xi > 0$ is the constant from the Θ -notation in Proposition 11, and the last inequality uses the fact that by assumption $V_t \leq C \cdot \log n$ for some constant $C > 0$.

The second condition of the theorem can be trivially fulfilled by choosing $r := 2$. We note that for $j = 0$ the condition obviously holds, and for $j \geq 1$ we have that $\Pr[|X_{t+1} - X_t| \geq 2j] = 0 \leq e^{-j}$, because $|X_{t+1} - X_t| \leq 1$ in the scaled process, as the marginal frequencies change by at most $\frac{1}{K}$.

We now move on to verifying the last condition of the theorem using the fact that $K \geq c' \cdot \log^2 n$:

$$\begin{aligned} \frac{\varepsilon l}{132 \log(r/\varepsilon)} &= \frac{\xi\alpha(1 - \beta)}{C \log n} \cdot K(\beta - \alpha) \cdot \frac{1}{132 \cdot \log((2C \log n)/(\xi\alpha(1 - \beta)))} \\ &\geq \frac{\xi\alpha(1 - \beta)(\beta - \alpha)}{132C \cdot \log((2C \log n)/(\xi\alpha(1 - \beta)))} \cdot c' \cdot \log n. \end{aligned}$$

Note that $\xi > 0$ can be chosen small enough such that the term above is positive for any choice of α and β . As all $\xi, \alpha, \beta > 0$ are constants, the term above is greater or equal to $4 = r^2$ for large enough n , since $\frac{\log n}{\log \log n} = \omega(1)$.

Furthermore, note that the constant $c' > 0$ in $K \geq c' \cdot \log^2 n$ can further be adapted to fulfill the following inequality:

$$\begin{aligned} \frac{\varepsilon l}{132r^2} &= \frac{\xi\alpha(1 - \beta) \cdot K(\beta - \alpha)}{132 \cdot 4 \cdot C \log n} \\ &\geq \frac{\xi\alpha(1 - \beta)(\beta - \alpha)}{528C} \cdot c' \cdot \log n \\ &\geq \gamma \ln n, \end{aligned}$$

because all other quantities which do not depend on n are fixed constants.

Since, by assumption, we started with $X_0 \geq b$, we can establish via the negative drift theorem that $\Pr[T \leq n^\gamma] = O(n^{-\gamma})$, where T is first time the fixed marginal frequency goes below α . \square

The next proposition bounds the expected time until a given frequency, starting from the lower boundary, reaches the upper boundary.

Proposition 23. *Let $c > 0$ be a constant and $K \geq 1$, and consider the algorithm $cGA(K, \frac{1}{n \log^c n})$ on DYNBV. Let $i \in \{1, \dots, n\}$ be an arbitrary bit at the lower boundary, and assume that $V_i = O(\log n)$ for the rest of the optimization. Then, the expected number of iterations until the frequency of bit i reaches the upper boundary is in $O(K^4 n \log^c n)$.*

Proof. Fix a frequency i and assume that it is at the lower boundary. The proof proceeds in three steps. In the first step, we lower bound the probability that frequency i leaves the lower boundary in a given iteration. In the second step, we couple the random walk $p_{i,t}$ of this frequency to the fair coin gambler ruin's random walk with self-loops, which allows us to leverage the classical bounds on this problem. Finally, we use Lemma 20 to deal with the self-loops and bound the number of iterations needed for the gambler's ruin random walk to terminate.

Let $D_{i,t}$ be the event that frequency i detaches from the lower boundary at iteration t , and $U_{i,t}$ be the event that, assuming $p_{i,t} = \frac{1}{K}$, the next boundary it reaches (after an arbitrary number of iterations) is the upper boundary. As a first step of the proof, we will be interested in showing that for any t :

$$\Pr[D_{i,t} \cap U_{i,t+1}] = \Omega\left(\frac{1}{Kn \log^c n}\right).$$

This can be shown by writing $\Pr[D_{i,t} \cap U_{i,t+1}] = \Pr[D_{i,t}] \cdot \Pr[U_{i,t+1} \mid D_{i,t}]$ and inspecting the two factors separately.

A simple lower bound for $\Pr[D_{i,t}]$ can be obtained as follows: Since frequency i is at a lower boundary, its value can only increase. This happens exactly when different values for bit i are sampled in the two offspring, and the winning offspring is the one that sampled bit i valued at 1. Moreover, the probability that the winning offspring is the one which has sampled the 1 is at least $\frac{1}{2}$. This yields the lower bound

$$\Pr[D_{i,t}] \geq 2 \cdot \frac{1}{n \log^c n} \left(1 - \frac{1}{n \log^c n}\right) \cdot \frac{1}{2} = \Omega\left(\frac{1}{n \log^c n}\right).$$

For the factor $\Pr[U_{i,t+1} \mid D_{i,t}]$ we will inspect the following process that dominates ours stochastically. Let us consider the gambler's ruin problem, where the probability of transitioning to the adjacent states is $\frac{1}{2}$ each. This is a pessimistic scenario, as we have already shown in Proposition 11 that there is a (possibly small) positive drift towards higher values of the marginal frequencies: by that proposition, $\Pr[p_{i,t'+1} = p + 1/K \mid p_{i,t'} = p] \geq \Pr[p_{i,t'+1} = p - 1/K \mid p_{i,t'} = p]$ for all $p < 1 - 1/(n \log^c n)$ and all $t' \geq t$. But in this gambler's ruin process, we reach the upper boundary with probability $\frac{1}{K}$ [1], therefore we can conclude that $\Pr[U_{i,t+1} \mid D_{i,t}] = \Omega\left(\frac{1}{K}\right)$. Note that since we are only interested in the probability of reaching the upper boundary (and not in the number of iterations within this happens), ignoring self-loops in our random walk is admissible.

We will now inspect the frequency of bit i , that is at the lower boundary, and split the iterations until it reaches the upper boundary into phases. A phase starts at the iteration t' when $p_{i,t'}$ is at the lower boundary, and ends at the first

iteration $t'' > t'$ for which $p_{i,t''} \in \left\{ \frac{1}{n \log^c n}, 1 - \frac{1}{n \log^c n} \right\}$. Thus, a phase contains all iterations $t \in (t', t'')$ in which the frequency of bit i is at neither of the two boundaries. We call a phase successful if at its end the frequency of bit i is at the upper boundary, otherwise we deem it unsuccessful.

Above, we have shown that the probability of a phase being successful is in $\Omega\left(\frac{1}{Kn \log^c n}\right)$. We therefore expect to go through $O(Kn \log^c n)$ phases until a success is encountered.

All that is left to show for us to reach our desired conclusion, is that the expected duration of a phase is $O(K^3)$ iterations. But this follows immediately from Lemma 20, since

$$4K^2 \ln K = O(K^3).$$

□

We are now ready to prove Theorem 4. The proof first consists of putting Proposition 21, Lemma 22 and Proposition 23 together, in order to reach a situation where almost all frequencies are at the upper boundary, and the remaining frequencies are lower bounded by a constant. From there, we just use a union bound over a series of high probability events that together imply the termination of the algorithm within an additional $\text{polylog}(n)$ iterations. For convenience, we restate the main theorem before starting the proof.

Theorem 24. *For $K = \Theta(\log^2 n)$ consider the algorithm $\text{cGA}(K, \frac{1}{n \log^7 n})$ on DYNBV. Then there exists a constant c' such that if $K \geq c' \log^2 n$, then the optimum is sampled in $O(n \cdot \text{polylog}(n))$ iterations with high probability.*

Proof. From Proposition 21 we know that for this chosen configuration of boundaries, with high probability we will have $V_t = O(\log n)$ for the next n^2 iterations after a burn-in phase that lasts $O(K^3 \log n) = O(\log^7 n)$ iterations.

We want to now show that, eventually, all frequencies will reach the upper boundary and not drop below a constant. By applying Lemma 22 with $\gamma = 2$ and taking a union bound over all n frequencies, we know that by choosing c' large enough then with high probability no frequency that ever reaches the upper boundary will drop below $\alpha = \frac{3}{4}$ (this value for α is just an arbitrary constant) during the following n^2 iterations.

Now, we can use Markov's inequality and Proposition 23 to claim that there is a constant probability for one specific bit i , which started at the lower boundary, to reach the upper boundary within $O(K^4 n \log^7 n)$ iterations (note that the proof of Proposition 21 implies that every bit will reach either of the boundaries after the burn-in phase, as all of them start out in the middle). Thus, by using arguments similar to the ones in the proof of Proposition 21, with high probability bit i will have reached the upper boundary after $O(K^4 n \log^7 n \cdot \log n) = O(n \log^{16} n)$. Moreover, the O -notation again allows us to pick suitable constants such that a union bound over all n frequencies will still yield that all of them will have reached the upper boundary after $O(n \log^{16} n)$ iterations with high probability.

Once all the frequencies have reached the upper boundary at least at some point, then the variance bound $V_t = O(\log n)$ implies that all but $O(\log^3 n)$ of

the frequencies are at the upper boundary. Indeed, any frequency $p_{i,t}$ that is not at the upper boundary must be in the interval $[3/4, 1 - 1/K]$, and in particular contributes at least $p_{i,t}(1 - p_{i,t}) \geq \Omega(1/K)$ to the variance V_t . Hence, there can be at most $O(K \log n) = O(\log^3 n)$ such frequencies. Let $C \geq 0$ be such that the number of frequencies that are not at the upper boundary is $C \log^3 n$, and without loss of generality assume that these frequencies are indexed by $i = 1, 2, \dots, C \log^3 n$. Since we are in a situation where none of the frequencies will drop below $3/4$, we know that if the frequencies return to a boundary, they will return to the upper one.

We claim that from this point on, with high probability the optimum is sampled during the next $\log^6 n$ iterations. Consider the following events: for $i \in \{1, 2, \dots, C \log^3 n\}$, let R_i denote the event that frequency i reaches the upper boundary during this phase of $\log^6 n$ iterations. Additionally, consider the event D that no frequency that is currently at the upper boundary or reaches the upper boundary detaches from it during this phase.

By Lemma 20, we know that $\Pr[R_i] \geq 1 - \frac{1}{n}$, since $\log^6 n = \omega(\log n \cdot K^2 \ln K)$ and hence we can take $r_{20} = A \log n$ for an arbitrarily large constant A . Hence, by a union bound over the $C \log^3 n$ bits that are not at the upper boundary, we have

$$\Pr \left[\bigcap_{i=1}^{C \log^3 n} R_i \right] \geq 1 - O\left(\frac{\log^3 n}{n}\right).$$

On the other hand, by a union bound over all n frequencies and all $\log^6 n$ iterations, $\Pr[D] \geq 1 - n \log^6 n \cdot O\left(\frac{1}{n \log^7 n}\right) = 1 - O\left(\frac{1}{\log n}\right)$, since the probability that one given bit detaches from the upper boundary during one given iterations is $O\left(\frac{1}{n \log^7 n}\right)$.

Hence $\Pr\left[D \cap \bigcap_{i=1}^{C \log^3 n} R_i\right] \geq 1 - o(1)$, and on the event $D \cap \bigcap_{i=1}^{C \log^3 n} R_i$ we have at least one iteration where all the frequencies are at the upper boundary. When all frequencies are at the upper boundary, then the probability of sampling the optimum in that iteration is at least (since we are sampling two offspring)

$$\left(1 - \frac{1}{n \log^7 n}\right)^n = \left(\left(1 - \frac{1}{n \log^7 n}\right)^{-n \log^7 n}\right)^{-1/\log^7 n} = 1 - o(1).$$

The RHS tends to 1 as $n \rightarrow \infty$, because $\lim_{n \rightarrow \infty} \left(1 - \frac{1}{n \log^7 n}\right)^{-n \log^7 n} = e$ and $\lim_{n \rightarrow \infty} \frac{1}{\log^7 n} = 0$. \square

7 Simulations

In this section, we provide simulations that complement our theoretical analysis. All figures depict the optimization of DYNBV but for varying hypothetical

population size K .⁸ The dimension of the search space is always $n = 300$. The probabilities $p_i, i = 1, \dots, n$ are initialized with $\frac{1}{2}$ as in the pseudocode of the algorithm. The lower and upper boundary are set at $\frac{1}{n}$ and at $1 - \frac{1}{n}$ respectively. The algorithm stops when the optimum has been sampled, or after 200'000 iterations if the optimum has not been sampled at that point. The code for the simulations is provided on request.

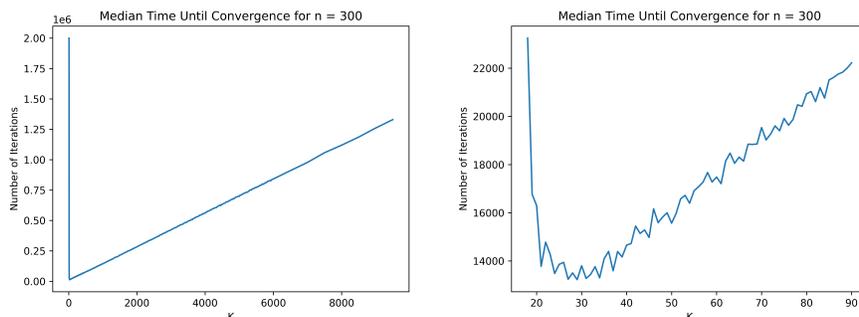


Fig. 1: Number of iterations for the optimization of DYNAMIC BINVAL with the cGA when $6 \leq K \leq 10000$. The right plot shows the subinterval $18 \leq K \leq 90$. The median over 50 runs is plotted.

The regime of small population sizes is shown in the right plot of Figure 1. Even at a small search space dimension of $n = 300$, the asymptotic speed-up of small K is clearly visible. For $K = 6, 7$, the optimum is not reached before the number of iterations are capped. This is in line with the observation that even for ONEMAX, when $K = o(\log(n))$, the runtime of the cGA becomes exponential. However, for $10 \leq K \leq 20$ we observe a phase transition, with the minimal runtime attained for hypothetical population sizes K around 30. Due to the small problem dimension, it is difficult to tell if the threshold is located at $K = \Theta(\log n)$, at $K = \Theta(\text{polylog}(n))$, or even $K = \Theta(n^c)$ for some small $c < 1$. But the data is consistent with the theoretical result that the optimum is obtained for the sublinear K regime of genetic drift.

For $K = \Omega(n \log n)$, Theorems 14 and 3 show an asymptotically tight runtime bound of $\Theta(Kn)$. Figure 1 covers a range of $K = 6$ up to $K = 10000$, exceeding the search space dimension of $n = 300$ by 2-3 orders of magnitude. We see indeed that the runtime increases proportionally to K , thus confirming our theoretical findings. In particular we see that, contrary to the optimization of ONEMAX, there are no local minima after the transition from the exponen-

⁸ For $6 \leq K \leq 420$ all integer values of K are simulated, for $421 \leq K \leq 1000$ all integer multiples of 5, for $1001 \leq K \leq 6000$ integer multiples of 20, for $6001 \leq K \leq 10000$ integer multiples of 500.

tial to the polynomial regime. Furthermore, the plot indicates that the runtime scales linearly with K in practice much earlier than our theoretical bound from Theorem 3.

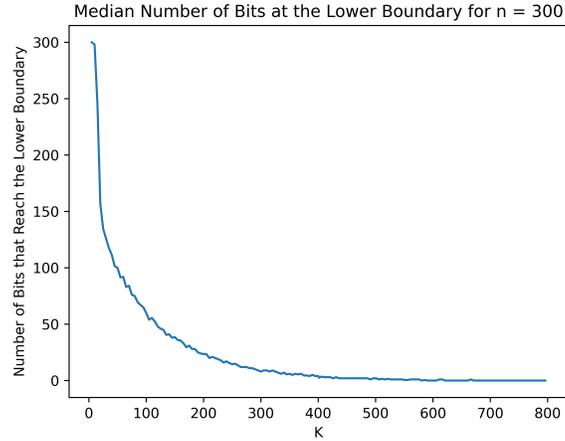


Fig. 2: Number of bits that reach the lower boundary $1 - \frac{1}{n}$ for the range $5 \leq K \leq 800$. The median over 20 runs is plotted.

In Figure 2, we see that after an initial exponential decrease, which is similar to the initial exponential runtime decrease in Figure 1, the number of frequencies ever reaching the lower boundary tapers off only slowly. In particular, for the empirically optimal value $K \approx 30$ from Figure 1 still many frequencies reach the lower boundary, confirming that this is in the aggressive regime of strong genetic drift. Until $K = n$, there is still a double-digit number of bits which reach the lower boundary. Only after approximately $K = 500 = \frac{5}{3}n$ the median drops to zero.

Acknowledgments. M.K. and U.S. were supported by the Swiss National Science Foundation [grant number 200021_192079]. The Dagstuhl seminar 22182 “Estimation-of-Distribution Algorithms: Theory and Applications” gave inspiration for this work.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Coolidge, J.L.: The gambler’s ruin. *The Annals of Mathematics* **10**(4), 181–192 (1909)

2. Dang, D.C., Lehre, P.K., Nguyen, P.T.H.: Level-based analysis of the univariate marginal distribution algorithm. *Algorithmica* **81** (02 2019)
3. De Bonet, J., Isbell, C., Viola, P.: Mimic: Finding optima by estimating probability densities. *Advances in neural information processing systems* **9** (1996)
4. Doerr, B.: Analyzing randomized search heuristics: Tools from probability theory. In: *Theory of Randomized Search Heuristics: Foundations and Recent Developments*, pp. 1–20. World Scientific (2011)
5. Doerr, B.: Probabilistic tools for the analysis of randomized optimization heuristics. In: *Theory of evolutionary computation*, pp. 1–87. Springer (2020)
6. Doerr, B.: The runtime of the compact genetic algorithm on jump functions. *Algorithmica* **83**, 3059–3107 (2021)
7. Doerr, B., Krejca, M., Lehre, P.K.: Estimation-of-distribution algorithms: Theory and applications. *Penal discussion* (2022). <https://doi.org/10.4230/DagRep.12.5.17>
8. Doerr, B., Krejca, M.S.: Significance-based estimation-of-distribution algorithms. In: *Proceedings of the Genetic and Evolutionary Computation Conference*. pp. 1483–1490 (2018)
9. Doerr, B., Krejca, M.S.: The univariate marginal distribution algorithm copes well with deception and epistasis. *Evolutionary Computation* **29**(4), 543–563 (2021)
10. Doerr, B., Zheng, W.: Sharp bounds for genetic drift in estimation of distribution algorithms. *IEEE Transactions on Evolutionary Computation* **24**(6), 1140–1149 (2020)
11. Friedrich, T., Kötzing, T., Krejca, M.S., Sutton, A.M.: The compact genetic algorithm is efficient under extreme gaussian noise. *IEEE Transactions on Evolutionary Computation* **21**(3), 477–490 (2016)
12. Friedrich, T., Kötzing, T., Neumann, F., Radhakrishnan, A.: Theoretical study of optimizing rugged landscapes with the cga. In: *International Conference on Parallel Problem Solving from Nature*. pp. 586–599. Springer (2022)
13. Kaufmann, M., Larcher, M., Lengler, J., Sieberling, O.: Hardest monotone functions for evolutionary algorithms (2023)
14. Kaufmann, M., Larcher, M., Lengler, J., Zou, X.: Onemax is not the easiest function for fitness improvements. In: *Evolutionary Computation in Combinatorial Optimization: 23rd European Conference, EvoCOP 2023, Held as Part of EvoStar 2023, Brno, Czech Republic, April 12–14, 2023, Proceedings*. p. 162–178. Springer-Verlag, Berlin, Heidelberg (2023)
15. Krejca, M., Witt, C.: Theory of Estimation-of-Distribution Algorithms, pp. 405–442 (01 2020). https://doi.org/10.1007/978-3-030-29414-4_9
16. Lehre, P.K., Nguyen, P.T.H.: On the limitations of the univariate marginal distribution algorithm to deception and where bivariate edas might help. In: *Proceedings of the 15th ACM/SIGEVO Conference on Foundations of Genetic Algorithms*. pp. 154–168 (2019)
17. Lehre, P.K., Nguyen, P.T.H.: Runtime analysis of the univariate marginal distribution algorithm under low selective pressure and prior noise. In: *Proceedings of the genetic and evolutionary computation conference*. pp. 1497–1505 (2019)
18. Lehre, P.K., Witt, C.: General drift analysis with tail bounds. *arXiv preprint arXiv:1307.2559* (2013)
19. Lengler, J.: Drift analysis. *Theory of evolutionary computation: Recent developments in discrete optimization* pp. 89–131 (2020)
20. Lengler, J., Meier, J.: Large population sizes and crossover help in dynamic environments. *Natural Computing* pp. 1–15 (2022)

21. Lengler, J., Riedi, S.: Runtime analysis of the $(\mu + 1)$ -ea on the dynamic binval function. *Evolutionary Computation in Combinatorial Optimization* **12692**, 84–99 (2021)
22. Lengler, J., Schaller, U.: The $(1 + 1)$ -ea on noisy linear functions with random positive weights. In: 2018 IEEE Symposium Series on Computational Intelligence (SSCI). pp. 712–719. IEEE (2018)
23. Lengler, J., Sudholt, D., Witt, C.: The complex parameter landscape of the compact genetic algorithm. *Algorithmica* **83**, 1096–1137 (2021)
24. Neumann, F., Sudholt, D., Witt, C.: A few ants are enough: Aco with iteration-best update. In: Proceedings of the 12th annual conference on Genetic and evolutionary computation. pp. 63–70 (2010)
25. Oliveto, P.S., Witt, C.: Improved time complexity analysis of the simple genetic algorithm. *Theoretical Computer Science* **605**, 21–41 (2015)
26. Pelikan, M., Lin, T.K.: Parameter-less hierarchical boa. In: Genetic and Evolutionary Computation Conference. pp. 24–35. Springer (2004)
27. Skala, M.: Hypergeometric tail inequalities: ending the insanity. arXiv preprint arXiv:1311.5939 (2013)
28. Sudholt, D., Witt, C.: On the choice of the update strength in estimation-of-distribution algorithms and ant colony optimization. *Algorithmica* **81**, 1450–1489 (2019)
29. Witt, C.: Upper bounds on the running time of the univariate marginal distribution algorithm on onemax. *Algorithmica* **81**, 632–667 (2019)
30. Witt, C.: How majority-vote crossover and estimation-of-distribution algorithms cope with fitness valleys. *Theoretical Computer Science* **940**, 18–42 (2023)