Article:
Sudholt, D. orcid.org/0000-0001-6020-1646 and Witt, C. (2019) On the choice of the update strength in estimation-of-distribution algorithms and ant colony optimization. Algorithmica, 81 (4). pp. 1450-1489. ISSN 0178-4617

https://doi.org/10.1007/s00453-018-0480-z

Algorithmica
https://doi.org/10.1007/s00453-018-0480-z

On the Choice of the Update Strength in Estimation-of-Distribution Algorithms and Ant Colony Optimization

Dirk Sudholt[1] · Carsten Witt[2]

Received: 5 June 2017 / Accepted: 9 July 2018

Abstract
Probabilistic model-building Genetic Algorithms (PMBGAs) are a class of metaheuristics that evolve probability distributions favoring optimal solutions in the underlying search space by repeatedly sampling from the distribution and updating it according to promising samples. We provide a rigorous runtime analysis concerning the update strength, a vital parameter in PMBGAs such as the step size $1/K$ in the so-called compact Genetic Algorithm (cGA) and the evaporation factor ρ in ant colony optimizers (ACO). While a large update strength is desirable for exploitation, there is a general trade-off: too strong updates can lead to unstable behavior and possibly poor performance. We demonstrate this trade-off for the cGA and a simple ACO algorithm on the well-known OneMax function. More precisely, we obtain lower bounds on the expected runtime of $\Omega(K\sqrt{n} + n \log n)$ and $\Omega(\sqrt{n}/\rho + n \log n)$, respectively, suggesting that the update strength should be limited to $1/K, \rho = O(1/(\sqrt{n} \log n))$. In fact, choosing $1/K, \rho \sim 1/(\sqrt{n} \log n)$ both algorithms efficiently optimize OneMax in expected time $\Theta(n \log n)$. Our analyses provide new insights into the stochastic behavior of PMBGAs and propose new guidelines for setting the update strength in global optimization.

Keywords Ant colony optimization · Estimation-of-distribution algorithms · Genetic Algorithms · Probabilistic model-building Genetic Algorithms · Runtime analysis · Theory of randomized search heuristics

An extended abstract of this article with parts of the results was presented at GECCO'16 [31].

✉ Dirk Sudholt
d.sudholt@sheffield.ac.uk

[1] University of Sheffield, Sheffield S1 4DP, UK

[2] DTU Compute,Technical University of Denmark, Kongens Lyngby, Denmark

Published online: 14 August 2018

1 Introduction

The term *probabilistic model-building Genetic Algorithms* describes a class of algorithms that construct a probabilistic model which is used to generate new search points. The model is adapted using information about previous search points. Both estimation-of-distribution algorithms (EDAs) and swarm intelligence algorithms including ant colony optimizers (ACO) and particle swarm optimizers (PSO) fall into this class. These algorithms generally behave differently from evolutionary algorithms where a population of search points fully describes the current state of the algorithm.

EDAs like the compact Genetic Algorithm (cGA) and many ACO algorithms update their probabilistic models by sampling new solutions and then updating the model according to information about good solutions found. In this work we focus on pseudo-Boolean optimization (finding global optima in $\{0, 1\}^n$, n the number of bits) and simple univariate probabilistic models, that is, for each bit there is a value p_i that determines the probability of setting the ith bit to 1 in a newly created solution.

Recently, the runtime analysis of such univariate EDAs has received increasing interest. Research has focused on the expected optimization time of not only cGA but also the univariate marginal distribution algorithm (UMDA), for which upper bounds [3,20,33] and lower bounds [18] on its expected runtime were obtained with respect to the problem $\text{OneMax}(x) := \sum_{i=1}^{n} x_i$, a simple hill-climbing task. Friedrich et al. [12,13] showed that the cGA is efficient on a noisy OneMax, even under extreme Gaussian noise. Moreover, Friedrich et al. [11] describe general properties of EDAs and how they are related to runtime analysis. In this paper, we follow up on work by Droste [7] on the cGA and by Neumann, Sudholt and Witt [26] on 2-MMAS$_{\text{ib}}$, an ACO algorithm that is closely related.

The cGA was introduced by Harik et al. [15]. In brief, it simulates the behavior of a Genetic Algorithm with population size K in a more compact fashion. In each iteration two solutions are generated, and if they differ in fitness, p_i is updated by $\pm 1/K$ in the direction of the fitter individual. Here $1/K$ reflects the strength of the update of the probabilistic model. Simple ACO algorithms based on the Max–Min ant system (MMAS) [29], using the iteration-best update rule, behave similarly: they generate a number λ of solutions and reinforce the best solution amongst these by increasing values p_i, here called *pheromones*, according to $(1 - \rho)p_i + \rho$ if the best solution had bit i set to 1, and $(1 - \rho)p_i$ otherwise. Here the parameter $0 < \rho < 1$ is called *evaporation factor*; it plays a similar role to the update strength $1/K$ for cGA.

Neumann et al. [26] showed that $\lambda = 2$ ants suffice to optimize the function OneMax, in expected time $O(\sqrt{n}/\rho)$ if the update strength is chosen small enough, $\rho \leq 1/(c\sqrt{n} \log n)$ for a suitably large constant $c > 0$. This is $O(n \log n)$ for $\rho = 1/(c\sqrt{n} \log n)$. If ρ is chosen unreasonably large, $\rho \geq c'/(\log n)$ for some $c' > 0$, the algorithm shows a chaotic behavior and needs exponential time even on this very simple function. In a more general sense, this result suggests that for global optimization such high update strengths should be avoided for any problem, unless the problem contains many global optima.

However, these results leave open a wide gap of parameter values between $\sim 1/(\log n)$ and $\sim 1/(\sqrt{n} \log n)$, for which no results are available. This leaves open the question of which update strengths are optimal, and for which values performance degrades. Understanding the working principles of the underlying probabilistic model remains an important open problem for both cGA and ACO algorithms. This is evident from the lack of reasonable lower bounds. The previous best known direct lower bound for MMAS algorithms for reasonable parameters was $\Omega((\log n)/\rho - \log n)$ [25, Theorem 5]; this bound holds for all functions with a unique global optimum. The best known lower bound for cGA on OneMax is $\Omega(K\sqrt{n})$ [7]. There are more general bounds from black-box complexity theory [6,8], showing that the expected runtime of comparison-based algorithms such as MMAS must be $\Omega(n)$ on OneMax. However, these black-box bounds do not yield direct insight into the stochastic behavior of the algorithms and do not shed light on the dependency of the algorithms' performance on the update strength.

In this paper, we study 2-MMAS$_{\text{ib}}$ and cGA with a much more detailed analysis that provides such insights through rigorous runtime analysis. We prove lower bounds of $\Omega(K\sqrt{n} + n\log n)$ and $\Omega(\sqrt{n}/\rho + n\log n)$ on OneMax. The terms $K\sqrt{n}$ and $\sqrt{n}/\rho$ indicate that the runtime decreases when the update strength $1/K$ or ρ is increased. However, the added terms $+\ n\log n$ set a limit: there is no asymptotic decrease and hence no benefit for choosing update strengths $1/K$ or ρ growing faster than $1/(\sqrt{n}\log n)$. The reason is that in this regime both algorithms suffer from a phenomenon well known in evolutionary biology and evolutionary computation as *genetic drift*: the probabilistic model attains extreme values simply due to the randomness of the sampling process, ignoring or overruling information about the quality of solutions. In our context, genetic drift leads to incorrect decisions being made. Correcting these incorrect decisions requires time $\Omega(n\log n)$. These lower bounds hold in expectation and with high probability; hence, they accurately reflect the algorithms' typical performance.

We further show that these bounds are tight for $1/K, \rho \leq 1/(c\sqrt{n}\log n)$. In this parameter regime the impact of genetic drift is bounded and hence these parameter choices provably lead to the best asymptotic performance on OneMax for arbitrary problem sizes n.

The lower bounds formally apply to OneMax, but we believe that they also apply more generally to functions with few optima. Among all functions with a unique global optimum, the function OneMax is provably the easiest function for certain evolutionary algorithms (see [5] for a proof for the (1+1) EA and [30,32] for extensions to populations), and similar results were shown for the cGA on linear functions by Droste [7]. We believe that the lower bounds give general performance limits for all functions with a unique global optimum. However, new arguments will be required to prove (or disprove) this formally.

From a technical point of view, our work uses a novel approach: using a second-order potential function to approximate the distribution of hitting times for a random walk that underlies changes in the probabilistic model. This approach has been recently picked up in [19] to analyze a different type of EDAs and we are confident that it will find further applications.

Finally, by pointing out similarities between cGA and 2-MMAS$_{\text{ib}}$, using the same analytical framework to understand changes in the probabilistic model, we make a step towards a unified theory of probabilistic model-building Genetic Algorithms.

This paper is structured as follows. Section 2 introduces the algorithms and Sect. 3 presents important analytical concepts. Section 4 proves efficient upper bounds for small update strengths, whereas Sect. 5 deals with the lower bounds for large update strengths. We finish with some conclusions.

2 Preliminaries

In the remainder, $p_t = (p_{t,1}, \ldots, p_{t,n})$ denotes a vector of probabilities and $x_t = (x_{t,1}, \ldots, x_{t,n})$, $y_t = (y_{t,1}, \ldots, y_{t,n})$ denote search points from $\{0, 1\}^n$. Hence $p_{t,i}$ refers to the i-th entry of p_t and $x_{t,i}$ refers to the ith bit in x_t.

Algorithm 1: Compact Genetic Algorithm (cGA)

```
$t \leftarrow 0$
$p_{t,1} \leftarrow p_{t,2} \leftarrow \cdots \leftarrow p_{t,n} \leftarrow 1/2$
while termination criterion not met do
    for $i \in \{1, \ldots, n\}$ do
        $x_{t,i} \leftarrow 1$ with prob. $p_{t,i}$, $x_{t,i} \leftarrow 0$ with prob. $1 - p_{t,i}$
    for $i \in \{1, \ldots, n\}$ do
        $y_{t,i} \leftarrow 1$ with prob. $p_{t,i}$, $y_{t,i} \leftarrow 0$ with prob. $1 - p_{t,i}$
    if $f(x_t) < f(y_t)$ then swap $x_t$ and $y_t$ for $i \in \{1, \ldots, n\}$ do
        if $x_{t,i} > y_{t,i}$ then $p_{t+1,i} \leftarrow p_{t,i} + 1/K$ if $x_{t,i} < y_{t,i}$ then $p_{t+1,i} \leftarrow p_{t,i} - 1/K$ if
        $x_{t,i} = y_{t,i}$ then $p_{t+1,i} \leftarrow p_{t,i}$ Restrict $p_{t+1,i}$ to be within $[1/n, 1 - 1/n]$
    $t \leftarrow t + 1$
```

Our presentation of cGA follows Droste [7]; see also Friedrich et al. [12]. The parameter $1/K$ is called update strength (classically, K is called population size) and the $p_{t,i}$ are called marginal probabilities. Pseudocode of cGA is shown in Algorithm 1. The cGA in each iteration generates two search points according to the probabilistic model. Then the better solution is reinforced: if the two solutions differ on some bit i, the probabilistic model $p_{t,i}$ is adjusted in the direction of the better solution, using a step size of $1/K$. If the two solutions have equal values on bit i then $p_{t,i}$ remains unchanged.

The simple MMAS algorithm 2-MMAS$_{\text{ib}}$, analyzed before in [26],[1] is shown in Algorithm 2. Note that the two algorithms only differ in the update mechanism. In contrast to cGA, 2-MMAS$_{\text{ib}}$ always changes the probabilistic model by either decreasing values $p_{t,i}$ to $(1 - \rho)p_{t,i}$ or increasing it to $(1 - \rho)p_{t,i} + \rho$. Here ρ determines the strength of the update. In the context of ACO, $p_{t,i}$ are usually called pheromone values, however we also refer to them as marginal probabilities to unify our approach to both algorithms.

[1] The 2-MMAS$_{\text{ib}}$ in [26] used a randomized tie-breaking rule that swaps x and y with probability $1/2$ if $f(x) = f(x)$. We omit this swap to ease presentation without changing the stochastic behavior; namely, conditioning on creating two specific samples x and y, where $x \neq y$, in one of the two possible orders, the probability of sampling x first is $1/2$ due to the independence of the trials.

Algorithm 2: 2-MMAS$_{\text{ib}}$

```
$t \leftarrow 0$
$p_{t,1} \leftarrow p_{t,2} \leftarrow \cdots \leftarrow p_{t,n} \leftarrow 1/2$
while termination criterion not met do
    for $i \in \{1, \ldots, n\}$ do
        $x_{t,i} \leftarrow 1$ with prob. $p_{t,i}$, $x_{t,i} \leftarrow 0$ with prob. $1 - p_{t,i}$
    for $i \in \{1, \ldots, n\}$ do
        $y_{t,i} \leftarrow 1$ with prob. $p_{t,i}$, $y_{t,i} \leftarrow 0$ with prob. $1 - p_{t,i}$
    if $f(x) < f(y)$ then swap $x$ and $y$ for $i \in \{1, \ldots, n\}$ do
        if $x_{t,i} \geq y_{t,i}$ then $p_{t+1,i} \leftarrow (1-\rho)p_{t,i} + \rho$ if $x_{t,i} < y_{t,i}$ then $p_{t+1,i} \leftarrow (1-\rho)p_{t,i}$
        Restrict $p_{t+1,i}$ to be within $[1/n, 1 - 1/n]$
    $t \leftarrow t + 1$
```

We note that the marginal probabilities for both algorithms are restricted to the interval $[1/n, 1 - 1/n]$. These bounds are used such that the algorithms always show a finite expected optimization time, as otherwise certain bits can be irreversibly fixed to 0 or 1. Our results also apply to algorithms without these borders: our analysis can be easily adapted to show that when the optimum is found efficiently in the presence of borders, it is found with high probability when borders are removed, and when the algorithm is inefficient, many bits are fixed opposite to the optimum.

There are intriguing similarities in the definition of cGA and 2-MMAS$_{\text{ib}}$, despite these two algorithms coming from quite different strands from the natural computation community. As mentioned earlier, they only differ in the update mechanism: cGA uses a symmetrical update rule with $1/K$ as the amount of change and changes a marginal probability if and only if both offspring differ in the corresponding bit value. 2-MMAS$_{\text{ib}}$ will always change a marginal probability in either positive or negative direction by a value dependent on its current state; however, the maximum absolute change will always be at most ρ. We are not the first to point out these similarities (e. g., see the survey by Hauschild and Pelikan [16], who embrace both algorithms under the umbrella of EDAs). However, our analyses will reveal the surprising insight that both cGA and 2-MMAS$_{\text{ib}}$ have the same runtime behavior as well as the same optimal parameter set on OneMax and can be analyzed with almost the same techniques.

Several parts of our analysis will consider random variables X that follow the so-called *Poisson-binomial* distribution with probability vector $(p_1, \ldots, p_n)$. Then X is the sum of n Bernoulli trials with possibly different success probabilities p_i, $1 \leq i \leq n$, i. e., $X = X_1 + \cdots + X_n$, where $X_i = 1$ with probability p_i and $X_i = 0$ with probability $1 - p_i$, independently for all trials. Note that the number of ones in the search points x_t and y_t sampled at time t by the cGA and 2-MMAS$_{\text{ib}}$ follows the Poisson-binomial distribution with probability vector $(p_{t,1}, \ldots, p_{t,n})$, which is why this distribution appears naturally in the analysis of OneMax. Section A.3 in the Appendix describes powerful bounds for such Poisson-binomially distributed random variables.

In the remainder of the paper, “poly(n)” is used as a shorthand for “$n^{O(1)}$.”

3 On the Dynamics of the Probabilistic Model

We first elaborate on the stochastic processes underlying the probabilistic model in both algorithms. These insights will then be used to prove upper runtime bounds for small update strengths in Sect. 4 and lower runtime bounds for large update strengths in Sect. 5.

We fix an arbitrary bit i and $p_{t,i}$, its marginal probability at time t. Note that $p_{t,i}$ is a random variable, and so is its random change $\Delta_t := p_{t+1,i} - p_{t,i}$ in one step. This change depends on whether the value of bit i matters for the decision whether to update with respect to the first bit string x sampled in iteration t (using p_t as sampling distribution) or the second one y (cf. also [26]). More precisely, we inspect $D_t := |x_t| - |x_{t,i}| - (|y_t| - |y_{t,i}|)$, which is the change of OneMax-value at bits other than i.

We assume $p_{t,i}$ to be bounded away from the borders such that Δ_t is not affected by the borders. Then cGA experiences two different kinds of steps:

Random-walk steps If $|D_t| \geq 2$, then bit i does not affect the decision whether to update with respect to x_t or y_t. For $\Delta_t > 0$ it is necessary that bit i is sampled differently. Hence, the $p_{t,i}$-value increases and decreases by $1/K$ with equal probability $p_{t,i}(1 - p_{t,i})$; with the remaining probability $p_{t+1,i} = p_{t,i}$. In this case, Δ_t can be described by a variable F_t where

$$F_t := \begin{cases} +1/K & \text{with probability } p_{t,i}(1-p_{t,i}), \\ -1/K & \text{with probability } p_{t,i}(1-p_{t,i}), \\ 0 & \text{with the remaining probability.} \end{cases}$$

We call a step where $|D_t| \geq 2$ a *random-walk step (rw-step)* since the process in such a step is a fair random walk (with self-loops) as $\mathrm{E}\big(\Delta_t \mid p_{t,i}, |D_t| \geq 2\big) = \mathrm{E}\big(F_t \mid p_{t,i}\big) = 0$.

If $D_t = 1$ then $|x_{t+1}| \geq |y_{t+1}|$ such that x_{t+1} and y_{t+1} are never swapped in line 8 of cGA. Hence, the same argumentation as in the previous case applies and the process performs an rw-step as well.

Biased steps If $D_t = -1$ then x_{t+1} and y_{t+1} are swapped unless bit i is sampled to 1 in x_{t+1} and to 0 in y_{t+1}. Hence, both events of sampling bit i differently increase the $p_{t,i}$-value. We have $\Delta_t = 1/K$ with probability $2p_{t,i}(1-p_{t,i})$ and $\Delta_t = 0$ otherwise.

If $D_t = 0$ then as in the case $D_t = -1$ both events of sampling bit i differently increase the $p_{t,i}$-value. Hence, we again have $\Delta_t = 1/K$ with probability $2p_{t,i}(1 - p_{t,i})$ and $\Delta_t = 0$ otherwise. Let B_t be a random variable such that

$$B_t := \begin{cases} +1/K & \text{with probability } 2p_{t,i}(1-p_{t,i}), \\ 0 & \text{with the remaining probability.} \end{cases}$$

Hence, in the cases $D_t = -1$ and $D_t = 0$ we get that Δ_t has the same distribution as B_t. We call such a step a *biased step (b-step)* since $\mathrm{E}\big(\Delta_t \mid p_{t,i}, D_t \in \{-1, 0\}\big) = \mathrm{E}\big(B_t \mid p_{t,i}\big) = 2p_{t,i}(1 - p_{t,i})/K > 0$ here.

Whether a step is an rw-step or b-step for bit i depends only on circumstances being external to the bit (and independent of it). Let R_t be the event that $D_t = 1$ or $|D_t| \geq 2$. We get the equality

$$\Delta_t = F_t \cdot \mathrm{P}[R_t] + B_t \cdot (1 - \mathrm{P}[R_t]), \tag{1}$$

which we denote as *superposition.* Informally, the change of $p_{t,i}$-value is a superposition of a fair (unbiased) random walk and biased steps. The fair random walk reflects the *genetic drift* underlying the process, i. e. the variance in the process may lead the algorithm to move in a random direction. In contrast, the biased steps reflect steps where the algorithm *learns* about which bit value leads to a better fitness at the considered bit position. We remark that the superposition of two different behaviors as formulated here is related to the approach taken in [2], where an EDA called UMDA was decomposed into a derandomized, deterministic EDA and a stochastic component modeling genetic drift.

For 2-MMAS$_{\text{ib}}$, structurally this kind of superposition holds as well, however, the underlying random variables look somewhat different.

Random-walk steps If $|D_t| \geq 2$ or $D_t = 1$, then the considered bit does not affect the choice whether to update with respect to x_t or y_t. Hence, the marginal probability of the considered bit increases with probability $p_{t,i}$ and decreases with probability $1 - p_{t,i}$.

We get that $\Delta_t = p_{t+1,i} - p_{t,i}$ is distributed as F_t in this case, where F_t is a random variable such that

$$F_t := \begin{cases} \rho \cdot (1 - p_{t,i}) & \text{with probability } p_{t,i}, \\ -\rho \cdot p_{t,i} & \text{with probability } 1 - p_{t,i}. \end{cases}$$

We call such a step an rw-step in analogy to cGA as in expectation the current state does not change: $\mathrm{E}\big(\Delta_t \mid p_{t,i}, |D_t| \geq 2 \vee D_t = 1\big) = \mathrm{E}\big(F_t \mid p_{t,i}\big) = 0$.

Biased steps If $D_t = 0$ or $D_t = -1$ then the marginal probability can only decrease if both offspring sample a 0 at bit i; otherwise it will increase. The difference Δ_t is a random variable

$$B_t := \begin{cases} \rho \cdot (1 - p_{t,i}) & \text{with probability } 1 - (1 - p_{t,i})^2, \\ -\rho \cdot p_{t,i} & \text{with probability } (1 - p_{t,i})^2. \end{cases}$$

This is called a biased step (b-step) as $\mathrm{E}\big(\Delta_t \mid p_{t,i}, D_t \in \{-1, 0\}\big) = \mathrm{E}\big(B_t \mid p_{t,i}\big) = \rho \cdot (1 - p_{t,i}) \cdot (1 - (1 - p_{t,i})^2) - \rho \cdot p_{t,i} \cdot (1 - p_{t,i})^2 = \rho(1 - p_{t,i})(1 - (1 - p_{t,i})^2 - p_{t,i}(1 - p_{t,i})) = \rho p_{t,i}(1 - p_{t,i}) > 0$.

Altogether, the superposition for 2-MMAS$_{\text{ib}}$ is also given by (1), with the modified meaning of B_t and F_t.

The strength of the update plays a key role here: if the update is too strong, large steps are made during updates, and genetic drift through rw-steps may overwhelm the probabilistic model, leading to “wrong” decisions being made in individual bits. On the other hand, small updates imply that rw-steps have a bounded impact, and the algorithm receives more time to learn optimal bit values in b-steps. We will formalize these insights in the following sections en route to proving rigorous upper and lower runtime bounds. Informally, one main challenge is to understand the stochastic process induced by the mixture of b- and rw-steps.

4 Small Update Strengths are Efficient

We first show that small update strengths are efficient for OneMax. This has been shown for 2-MMAS$_{\text{ib}}$ in [26].

Theorem 1 ([26]) *If $\rho \le 1/(c\sqrt{n}\log n)$ for a sufficiently large constant $c > 0$ and $\rho \ge 1/\text{poly}(n)$ then 2-MMAS$_{ib}$ optimizes* OneMax *in expected time $O(\sqrt{n}/\rho)$.*

For $\rho = 1/(c\sqrt{n}\log n)$ the runtime bound is $O(n\log n)$.

Here we exploit the similarities between both algorithms to prove an analogous result for cGA.

Theorem 2 *The expected optimization time of cGA on* OneMax *with $K \ge c\sqrt{n}\log n$ for a sufficiently large $c > 0$ and $K = \text{poly}(n)$ is $O(\sqrt{n}K)$. This is $O(n\log n)$ for $K = c\sqrt{n}\log n$.*

The analysis follows the approach for 2-MMAS$_{\text{ib}}$ in [26], adapted to the different update rule, and using modern tools like *variable drift analysis*[2] [17] and drift analysis with tail bounds [21]. We also extend previous work by showing in Sect. 4.1 that the upper bound for cGA holds with high probability (see Theorem 5 in Sect. 4.1). The main idea is that marginal probabilities are likely to increase from their initial values of 1/2. If the update strength is chosen small enough, the effect of genetic drift (as present in rw-steps) is bounded such that with high probability all bits never reach marginal probabilities below 1/3. Under this condition, we show that the marginal probabilities have a tendency (stochastic drift) to move to their upper borders, such that then the optimum is found with good probability.

The following lemma uses considerations and notation from Sect. 3 to establish a *stochastic drift*, i.e. a positive trend towards optimal bit values, for cGA. We use the same notation as in Sect. 3.

Lemma 3 *If $1/n + 1/K \le p_{t,i} \le 1 - 1/n - 1/K$ then*

$$\mathrm{E}\big(\Delta_t \mid p_{t,i}\big) \ge \frac{2}{11}\,\frac{p_{t,i}(1-p_{t,i})}{K}\left(\sum_{j\ne i} p_{t,j}(1-p_{t,j})\right)^{-1/2}.$$

Proof The assumptions on $p_{t,i}$ assure that $p_{t+1,i}$ is not affected by the borders $1/n$ and $1-1/n$. Then the expected change is given by the expectation of the superposition (1):

$$\mathrm{E}\big(\Delta_t \mid p_{t,i}\big) = \mathrm{E}\big(F_t \mid p_{t,i}\big)\cdot \mathrm{P}[R_t] + \mathrm{E}\big(B_t \mid p_{t,i}\big)\cdot(1-\mathrm{P}[R_t]).$$

[2] The term “drift” is used in both “genetic drift” and in “drift analysis.” In the latter, “drift” is used to indicate the expected progress towards a target. We sometimes use the term “stochastic drift” to distinguish it from “genetic drift”. Drift theorems always refer to stochastic drift.

From Sect. 3 we know $\mathrm{E}\bigl(F_t \mid p_{t,i}\bigr) = 0$ and $\mathrm{E}\bigl(B_t \mid p_{t,i}\bigr) = 2p_{t,i}(1-p_{t,i})/K$. Further,

$$1 - \mathrm{P}[R_t] \geq \mathrm{P}[D_t = 0] \geq \frac{1}{11}\left(\sum_{j \neq i} p_{t,j}(1 - p_{t,j})\right)^{-1/2},$$

where the last inequality was shown in [26, proof of Lemma 1]. Here we exploit that cGA and 2-MMAS$_{\text{ib}}$ use the same construction procedure. Together this proves the claim. □

Note that the term $\left(\sum_{j \neq i} p_{t,j}(1 - p_{t,j})\right)^{1/2}$ reflects the standard deviation of the sampling distribution on all bits $j \neq i$.

Lemma 3 indicates that the drift increases with the update strength $1/K$. However, a too large value for $1/K$ also increases genetic drift. The following lemma shows that, if $1/K$ is not too large, this positive drift implies that the marginal probabilities will generally move to higher values and are unlikely to decrease by a constant.

Lemma 4 *Let $0 < \alpha < \beta < 1$ be two constants. For each constant $\gamma > 0$ there exists a constant $c_\gamma > 0$ (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 $K \geq c_\gamma \sqrt{n} \log n$ then the probability that during the following n^γ steps the marginal probability decreases below α is at most $O(n^{-\gamma})$.*

Proof The proof uses a similar approach as the proof of Lemma 3 in [26], using $1/K$ instead of ρ and drift bounds from Lemma 3.

The aim is to apply the negative drift theorem, Theorem 20 in the Appendix, with respect to the stochastic process $X_t := Kp_{t,i}$, obtained by scaling the process on the marginal probabilities of the considered bit i by a factor of K. Note that the X_t-process is on $\{K/n, 1, 2, \ldots, K-1, K-K/n, K\}$.

We use the interval $[a, b] := [\alpha K, \beta K]$ in the drift theorem. To establish the first condition of the drift theorem, we use Lemma 3. Hence, we obtain the following bound on the drift

$$\begin{aligned}\mathrm{E}(X_{t+1} - X_t \mid X_t; a < X_t < b) &\geq K \cdot \frac{2\alpha(1-\beta)}{11K} \cdot \left(\sum_{j \neq i} p_{t,j}(1 - p_{t,j})\right)^{-1/2} \\ &\geq \frac{\alpha(1-\beta)}{11\sqrt{n}} =: \varepsilon\end{aligned}$$

using that $a < X_t < b$ implies $\alpha < p_{t,i} < \beta$, and estimating $p_{t,j}(1 - p_{t,j}) \leq 1/4$ for all j and t.

For the second condition, we note that always $|X_t - X_{t+1}| \leq 1$ since the marginal probabilities change by at most $1/K$. Hence, the second condition is trivially satisfied by choosing $r := 2$.

To verify the third condition, we will use that $K \geq c_\gamma \sqrt{n} \log n$ for a constant c_γ that may depend on α, β and γ. We compute, using $\ell := (\beta - \alpha)K$ and r, ε defined above,

$$\frac{\varepsilon\ell}{132\log(r/\varepsilon)} \geq \frac{(\beta-\alpha)K\alpha(1-\beta)}{132\cdot 11\sqrt{n}\log(\Theta(\sqrt{n}))} \geq \frac{\alpha(\beta-\alpha)(1-\beta)c_\gamma\sqrt{n}\log n}{716\sqrt{n}\log n + \Theta(\sqrt{n})},$$

which is at least 4 if c_γ is chosen large enough but constant; here we use that α and β are constants in (0, 1). Then $1 \leq r^2 \leq \frac{\varepsilon(b-a)}{132\log(r/\varepsilon)}$ as demanded by the third condition.

To finally apply the drift theorem, similar calculations as before yield that

$$\frac{\varepsilon\ell}{132r^2} \geq \frac{\alpha(\beta-\alpha)(1-\beta)c_\gamma\sqrt{n}\log n}{528\cdot 11\sqrt{n}},$$

which is at least $\gamma\ln n$ if c_γ is chosen appropriately. By assumption $X_0 \geq b$. Hence, the theorem establishes that $\mathrm{P}[T \leq n^\gamma] = O(n^{-\gamma})$. □

With these lemmas, we now prove the main statement of this section.

Proof of Theorem 2 We assume in the following that $1/K$ divides $1/2-1/n$, implying that marginal probabilities are restricted to $\{1/n, 1/n+1/K, \ldots, 1/2, \ldots, 1-1/n-1/K, 1-1/n\}$.

Following [26, Theorem 3] we show that, starting with a setting where all probabilities are at least $1/2$ simultaneously, with probability $\Omega(1)$ after $O(\sqrt{n}K)$ iterations either the global optimum has been found or at least one probability has dropped below $1/3$. In the first case we speak of a success and in the latter case of a failure. The expected time until either a success or a failure happens is then $O(\sqrt{n}K)$.

Now choose a constant $\gamma > 0$ such that $n^\gamma \geq Kn^3$. According to Lemma 4 applied with $\alpha := 1/3$ and $\beta := 1/2$, the probability of a failure in n^γ iterations is at most $n^{-\gamma}$, provided the constant c in the condition $K \geq c\sqrt{n}\log n$ is large enough. In case of a failure we wait until the probabilities simultaneously reach values at least $1/2$ again and then we repeat the arguments from the preceding paragraph. It is easy to show (cf. Lemma 2 in [26]) that the expected time for one probability to reach the upper border is always bounded by $O(n^{3/2}K)$, regardless of the initial probabilities. By standard arguments on independent phases, the expected time until *all* probabilities have reached their upper border at least once is $O(n^{3/2}K\log n)$. Once a bit reaches the upper border, we apply Lemma 4 again with $\alpha := 1/2$ and $\beta := 2/3$ to show that the probability of a marginal probability decreasing below $1/2$ in time n^γ is at most $n^{-\gamma}$ (again, for large enough c). The probability that there is a bit for which this happens is at most $n^{-\gamma+1}$ by the union bound. If this does not happen, all bits attain value at least $1/2$ simultaneously, and we apply our above arguments again.

As the probability of a failure is at most $n^{-\gamma+1}$, the expected number of restarts is $O(n^{-\gamma+1})$ and considering the expected time until all bits recover to values at least $1/2$ only leads to an additional term of $n^{-\gamma+1}\cdot O((n^{3/2}\log n)K) \leq o(1)$ (as $n^{-\gamma} \leq n^{-3}/K$) in the expectation.

We only need to show that after $O(\sqrt{n}K)$ iterations without failure the probability of having found the global optimum is $\Omega(1)$. To this end, we consider a simple potential function that takes into account marginal probabilities for all bits. An important property of the potential is that once the potential has decreased to some constant value, the probability of generating the global optimum is constant.

Let $p_1, \dots, p_n$ be the current marginal probabilities and $q_i := 1 - 1/n - p_i$ for all i. Define the potential function $\varphi := \sum_{i=1}^n q_i$, which measures the distance to an ideal setting where all probabilities attain their maximum $1 - 1/n$. Let q_i' be the q_i-value in the next iteration and $p_i' = 1 - q_i'$. We estimate the expectation of $\varphi' := \sum_{i=1}^n q_i'$ and distinguish between two cases. If $p_i \le 1 - 1/n - 1/K$, by Lemma 3

$$\mathrm{E}\big(q_i' \mid q_i\big) \;\le\; q_i - \frac{p_i(1-p_i)}{K} \cdot \frac{2}{11} \cdot \left(\sum_{j \ne i} p_j(1-p_j)\right)^{-1/2}.$$

We bound $p_i(1-p_i)$ from below using $p_i \ge 1/3$ and $1 - p_i = q_i + 1/n$ and the sum from above using

$$\sum_{j \ne i} p_j(1-p_j) \le \sum_{j=1}^n (1-p_j) = \sum_{j=1}^n (q_j + 1/n) = 1 + \varphi.$$

Then

$$\begin{aligned}\mathrm{E}\big(q_i' \mid q_i\big) &\le\; q_i - \frac{q_i}{K} \cdot \frac{2}{33} \cdot \left(\frac{1}{1+\varphi}\right)^{1/2} \\ &\le\; q_i \left(1 - \frac{2}{33K} \cdot \frac{1}{1+\varphi^{1/2}}\right).\end{aligned}$$

If $p_i > 1 - 1/n - 1/K$, then $p_i = 1 - 1/n$ (as $1/K$ is a multiple of $1/2 - 1/n$) and p_i can only decrease. A decrease by $1/K$ happens with probability $1/n$, thus

$$\mathrm{E}\big(q_i' \mid q_i\big) \;\le\; q_i + \frac{1}{nK}.$$

To ease the notation we assume w. l. o. g. that the bits are numbered according to decreasing probabilities, i. e., increasing q-values. Let $m \in \mathbb{N}_0$ be the largest index such that $p_m = 1 - 1/n$. Observe that by definition of the q_i we have $\sum_{i=1}^m q_i = 0$ and $\sum_{i=m+1}^n q_i = \varphi$. It follows

$$\sum_{i=1}^m \mathrm{E}\big(q_i' \mid q_i\big) \;\le\; \sum_{i=1}^m q_i + \frac{m}{nK} \le \frac{1}{K}.$$

Putting everything together,

$$\begin{aligned}\mathrm{E}\big(\varphi' \mid \varphi\big) \;&=\; \sum_{i=1}^m \mathrm{E}\big(q_i' \mid q_i\big) + \sum_{i=m+1}^n \mathrm{E}\big(q_i' \mid q_i\big) \\ &\le\; \frac{1}{K} + \sum_{i=m+1}^n q_i \left(1 - \frac{2}{33K} \cdot \frac{1}{1+\varphi^{1/2}}\right)\end{aligned}$$

$$\begin{aligned} &= \frac{1}{K} + \varphi\left(1 - \frac{2}{33K} \cdot \frac{1}{1+\varphi^{1/2}}\right) \\ &= \varphi + \frac{1}{K} - \frac{2}{33K} \cdot \frac{\varphi^{1/2}}{\varphi^{-1/2}+1}. \end{aligned}$$

For $\varphi \geq 10000$ this can further be bounded using

$$\frac{2}{33K} \cdot \frac{\varphi^{1/2}}{\varphi^{-1/2}+1} \geq \frac{2}{33K} \cdot \frac{100}{101/100} > \frac{6}{K}$$

thus

$$\begin{aligned} \mathrm{E}\big(\varphi' \mid \varphi\big) &\leq \varphi + \frac{1}{K} - \frac{1}{6} \cdot \frac{2}{33K} \cdot \frac{\varphi^{1/2}}{\varphi^{-1/2}+1} - \frac{5}{6} \cdot \frac{2}{33K} \cdot \frac{\varphi^{1/2}}{\varphi^{-1/2}+1} \\ &\leq \varphi - \frac{5}{6} \cdot \frac{2}{33K} \cdot \frac{\varphi^{1/2}}{\varphi^{-1/2}+1} \\ &\leq \varphi - \frac{5}{6} \cdot \frac{2}{33K} \cdot \frac{100}{101} \cdot \varphi^{1/2} \\ &\leq \varphi - \frac{\varphi^{1/2}}{17K} \end{aligned}$$

where in the third inequality we used $\varphi \geq 10000$ again. We now apply the variable drift theorem (given by Theorem 18 in the Appendix) to bound the expected time for the potential φ to decrease from any initial value $\varphi \leq n$ to a value $\varphi \leq 10000$. To this end, we use the drift function $h(\varphi) := \varphi^{1/2}/(17K)$ as we just established that the expected change (drift) in one step is at least $h(\varphi)$ for all $\varphi \geq 10000$.

Since Theorem 18 only considers the hitting time of state 0 and the condition on the drift needs to hold for all states larger than 0, we consider a modified process instead where we merge all states with potentials $0 < \varphi < 10000$ with state 0: all steps reducing a potential of $\varphi \geq 10000$ to a value smaller than 10000 yield a potential of 0. In the modified process, the smallest state larger than 0 is $x_{\min} = 10000$. The modification can only increase the drift, hence the drift is still bounded from below by $h(\varphi)$ for all states $\varphi \geq x_{\min}$.

Now Theorem 18 yields that the expected time to reach state 0 in the modified process, or, equivalently, any state $\varphi < 10000$ in the original process, is at most

$$\frac{10000}{h(10000)} + \int_{10000}^{n} \frac{1}{h(\varphi)} \, \mathrm{d}\varphi \;=\; O(K) + O(K) \cdot \int_{10000}^{n} \varphi^{-1/2} \, \mathrm{d}\varphi \;=\; O(\sqrt{n}K).$$

Consider an iteration where $\varphi \leq 10000$. The probability of creating ones on all bits simultaneously, given that all marginal probabilities are at least $1/3$, is minimal in the extreme setting where a maximal number of bits has marginal probabilities at $1/3$ and all other bits, except at most one, have marginal probabilities at their upper border. Then the probability of creating the optimum in one step is at least

$\left(1-\frac{1}{n}\right)^{n-1} \cdot 3^{-\lceil \varphi \cdot 3/2 \rceil} = \Omega(1)$. Hence a successful phase finds the optimum with probability $\Omega(1)$. □

4.1 A Tail Bound on the Running Time

We further show that the upper bound from Theorem 2 holds with high probability. Along with the lower tail bounds to be presented in Sect. 5, this demonstrates that the runtime of cGA is highly concentrated, and that we have developed a very good understanding of its performance and dynamic behaviour. In the following result, the failure probability can be made an arbitrarily small polynomial.

Theorem 5 *For every $\kappa > 0$ there is a constant $c = c(\kappa)$ such that the upper bound $O(\sqrt{n}K)$ for the time of the cGA on* OneMax *from Theorem 2 holds with probability $1 - O(n^{-\kappa})$, provided $K \geq c\sqrt{n}\log n$ and $K = \mathrm{poly}(n)$.*

Throughout this section we re-use the notation from the proof of Theorem 2, in particular the potential function φ and variables p_i and $q_i := 1 - 1/n - p_i$ for $1 \leq i \leq n$.

We still consider the stochastic process w. r. t. the potential function φ from the proof of Theorem 2 and consider its drift. As done in said proof, we use that the probability that there exists a p_i whose value decreases below $1/3$ in n^γ steps is at most $n^{-\gamma+1}$ if the constant c in $K \geq c\sqrt{n}\log n$ is chosen large enough. Note that we can make γ larger to decrease the probability of such a failure; however, this dictates what values of c are appropriate. In the following, we assume that the probability of such a failure is at most $n^{-\kappa}$ and work under the assumption that no failure occurs.

To get a high-probability statement, we aim to apply drift analysis with tail bounds, stated as Theorem 19 in the Appendix.[3] To this end, we have to bound the moment-generating function (mgf.) of (a stochastic upper bound on) the absolute value of

$$\int_{\varphi_{t+1}}^{\varphi_t} \frac{1}{h(\max\{x, x_{\min}\})}\,\mathrm{d}x \leq \int_{\varphi_{t+1}}^{\varphi_t} \frac{K'}{\sqrt{x}}\,\mathrm{d}x,$$

where we use $K' = 17K$ to improve readability and $x_{\min} = 10000$.

The following lemma gives a tail bound for the time to reach a potential of at most $x_{\min}$.

Lemma 6 *Consider the potential φ and the drift function $h(\varphi) := \varphi^{1/2}/(17K)$ as defined in the proof of Theorem 2, and assume that no p_i decreases below $1/3$. Let T denote the random time for the potential to decrease below $x_{\min} = 10000$ for the first time, when starting with an initial value of φ_0. Then for every $t > 0$, conditional on the potential always being bounded by a maximum value $x_{\max}$,*

$$\mathrm{P}[T > t \mid \varphi_0, \ldots, \varphi_t \leq x_{\max}] \leq e^{\Omega((x_{\min}/h(x_{\min}) + \int_{x_{\min}}^{\varphi_0} 1/h(x)\,\mathrm{d}x - t/2)/x_{\max})}.$$

[3] To apply Theorem 19 we will again consider a slightly modified process, where potential values $0 < \varphi < 10000$ are being merged with state 0.

Proof For the purpose of bounding the tail of the first hitting time for potentials below 10000 we again consider a modified process where states $0 < \varphi < 10000$ are merged with state 0 (cf. proof of Theorem 18). The following calculations implicitly assume that $\varphi_t \geq 10000$ as otherwise we have reached a potential below 10000.

We first note that always $\varphi_{t+1} \geq \varphi_t(1-1/K) \geq \varphi_t/2$. This holds since a step of cGA in the worst case increases all frequencies by $1/K$ (except for those at the upper border), which decreases each q_i by $1/K$. Hence, we get

$$\Big|\int_{\varphi_{t+1}}^{\varphi_t} \frac{1}{h(\max\{x, x_{\min}\})}\,\mathrm{d}x\Big| \prec \int_{\varphi_{t+1}}^{\varphi_t} \frac{K'}{\sqrt{x}}\,\mathrm{d}x \prec \frac{K'|\varphi_{t+1}-\varphi_t|}{\sqrt{\varphi_t/2}},$$

and we are left with an analysis of $\Delta =: |\varphi_{t+1}-\varphi_t|$. Here we note that for any bit i, its frequency changes by an absolute value of at most $1/K$ with probability at most $q_i + 1/n \leq 2q_i$. Hence, $K\Delta$ is stochastically dominated by a Poisson-binomial distribution with parameters n and $2q_i$, where $1 \leq i \leq n$. Let A be the random variable describing this Poisson-binomial distribution. While we do not know the individual success probabilities, we know their average $p^* := \sum(2q_i/n) = 2\varphi_t/n$ and can bound A by a random variable B, where $B \sim np^* + \mathrm{Bin}(n, p^*) + 2$. To show this, we note that $\mathrm{P}[B \geq t] \geq \mathrm{P}[A \geq t]$ is trivial for $t \leq np^* + 2$ (as $\mathrm{P}[B \geq t] = 1$). For $t > np^* + 2$, even the dominance $\mathrm{P}\big[\mathrm{Bin}(n, p^*) \geq t\big] \geq \mathrm{P}[A \geq t]$ holds by the results of Gleser [14], see [23, p. 495] for a summary. Hence,

$$\begin{aligned}\Big|\int_{\varphi_{t+1}}^{\varphi_t} \frac{1}{h(\max\{x, x_{\min}\})}\,\mathrm{d}x\Big| &\prec \frac{1}{K}\,\frac{K'(np^* + 2 + \mathrm{Bin}(n, p^*))}{\sqrt{\varphi_t/2}}\\ &= \frac{c_1(np^* + 2 + \mathrm{Bin}(n, p^*))}{\sqrt{\varphi_t}} =: Z\end{aligned}$$

for some constant $c_1 > 0$.

We now bound the mgf. of Z. Looking up the mgf. of a binomial distribution, we obtain

$$\begin{aligned}\mathrm{E}\Big(e^{\lambda Z}\Big) &= \mathrm{E}\bigg(\Big(e^{\lambda(np^*+2+\mathrm{Bin}(n,p^*))}\Big)^{c_1/\sqrt{\varphi_t}}\bigg) = \Big(e^{\lambda np^*+2} \cdot \big(1 - p^* + p^* e^{\lambda}\big)^n\Big)^{c_1/\sqrt{\varphi_t}}\\ &= \Big(e^{\lambda(p^*+2/n)}\,\big(1 - p^* + p^* e^{\lambda}\big)\Big)^{nc_1/\sqrt{\varphi_t}}\end{aligned}$$

Assuming $\lambda \leq \min\{1, 1/(16c_1\sqrt{\varphi_t})\}$ and using $e^{\lambda} \leq 1 + \lambda + \lambda^2 \leq 1 + 2\lambda$, we bound the last expression from above by

$$\Big(e^{\lambda(p^*+2/n)}(1 - p^* + p^*(1+2\lambda))\Big)^{nc_1/\sqrt{\varphi_t}} = \Big(e^{\lambda(p^*+2/n)}(1 + 2p^*\lambda))\Big)^{nc_1/\sqrt{\varphi_t}},$$

which, since $1 + x \leq e^x$ for $x \in \mathbb{R}$, is at most

$$e^{\lambda(np^*+2)c_1/\sqrt{\varphi_t}} e^{\lambda 2p^* nc_1/\sqrt{\varphi_t}} \leq e^{\lambda 4np^* c_1/\sqrt{\varphi_t}} \leq e^{8\lambda c_1\sqrt{\varphi_t}},$$

since $p^* = 2\varphi_t/n$ and $np^* \geq 20000 \geq 2$ by our assumption on φ_t.

Using (again) $e^x \leq 1 + 2x$ for $x \leq 1$ and recalling that $\lambda \leq 1/(16c_1\sqrt{\varphi_t})$, we arrive at the bound

$$\mathrm{E}\left(e^{\lambda Z}\right) \leq 1 + 32c_1\lambda\sqrt{\varphi_t} \leq 1 + c_2\lambda\sqrt{\varphi_t} =: D,$$

for some some constant $c_2 > 0$. Hence, using the variable drift theorem with tail bounds, Theorem 19 in the Appendix, we get for any $\delta > 0$ and $\eta \leq \min\{\lambda, \delta\lambda^2/(D - 1 - \lambda)\}$ that

$$\mathrm{P}[T > t] \leq e^{\eta(x_{\min}/h(x_{\min}) + \int_{x_{\min}}^{\varphi_0} 1/h(x)\,\mathrm{d}x - (1-\delta)t)}. \tag{2}$$

We note that $\sqrt{\varphi_t} \geq 100$ if $\varphi_t \geq x_{\min} = 10000$. Hence, using our bound D, we satisfy

$$\frac{\delta\lambda^2}{D - 1 - \lambda} = \frac{\delta\lambda^2}{(c_2\sqrt{\varphi_t} - 1)\lambda} \geq \frac{\delta\lambda}{c_2\sqrt{\varphi_t}},$$

if c_2 is chosen large enough for $c_2\sqrt{\varphi_t} - 1 \geq 0$ to hold. Similarly, we show that $\delta\lambda^2/(D - 1 - \lambda) \leq \lambda$ if δ is sufficiently small, so that only the second argument of $\min\{\lambda, \delta\lambda^2/(D - 1 - \lambda)\}$ needs to be considered. We let $\delta := 1/2$. We choose $\lambda := 1/(16c_1\sqrt{x_{\max}})$ and $\eta := \delta\lambda/(c_2\sqrt{x_{\max}}) = c_3/x_{\max}$ for some constant c_3 to satisfy the requirements on λ and η. Substituting η and δ in (2) proves the claim. □

Reaching a small potential is not sufficient to show that the optimum is found with high probability. We also need to show that the algorithm spends a sufficiently large number of steps at a small potential. The following lemma shows that, after having reached a potential of at most $x_{\min}$, the algorithm quickly returns to this regime.

Lemma 7 *Consider the potential φ as defined in the proof of Theorem 2, where $K \geq c\sqrt{n}\log n$ for a sufficiently large $c > 0$ and $K = \mathrm{poly}(n)$. Whenever $\varphi_0 < 10000$, the time $R = \min\{t \geq 1 \mid \varphi_t < 10000, \varphi_0 < 10000\}$ to return to a potential below 10000 is at most $K\log^2 n$ with probability $1 - n^{-\omega(1)}$.*

Proof We first show that with high probability the potential never rises beyond $O(K)$ in any polynomial number of steps.

Consider p_i that are at the upper border initially. The probability that in one step more than $\log n$ variables move away from the upper border is at most $\binom{n}{\log n}(1/n)^{\log n} \leq 1/((\log n)!) = n^{-\omega(1)}$. Assuming this never happens within the next $K\log^2 n$ steps, during this time at most $K\log^3 n$ bits move away from the upper border. As every bit can only increase the potential by $1/K$ in one step, these bits only contribute at most $\log^3 n$ to the potential.

All bits that are not at the upper border initially can contribute up to 1 to the potential each. However, as they contribute at least $1/K$ (the minimum distance to the upper border), the number of such bits is bounded by $10000K$. Together, the potential is at most $\log^3 n + 10000K = O(K)$ with probability $1 - (K\log^2 n)\cdot n^{-\omega(1)} = 1 - n^{-\omega(1)}$ (as $K\log^2 n = \mathrm{poly}(n)$) throughout the first $K\log^2 n$ steps.

Now consider the potential φ_1 at time 1. If $\varphi_1 < 10000$, the return time is $R = 1$. Otherwise, by the same arguments as above, $\varphi_1 \le 10000 + O(1)$ with probability $1 - n^{-\omega(1)}$ as with this probability at most $\log n$ bits move away from the upper border, and at most $10000K$ bits that are away from the border initially only move by $\pm 1/K$ in one step.

Applying Lemma 6 with an initial potential (denoted by φ_0 in Lemma 6 but corresponding to φ_1 in the time scale of the present lemma) of at most $10000 + O(1)$, $t = K \log^2 n$, and $x_{\max} = \log^3 n + 10000K = O(K)$ yields that the probability of not returning to a potential below 10000 in $K \log^2 n$ steps is at most

$$e^{\left(x_{\min}/h(x_{\min}) + \int_{x_{\min}}^{10000+O(1)} 1/h(x)\,\mathrm{d}x - K \log^2 n\right)/O(K)}.$$

Note that

$$\begin{aligned} &\frac{x_{\min}}{h(x_{\min})} + \int_{x_{\min}}^{10000+O(1)} \frac{1}{h(x)}\,\mathrm{d}x \\ &= \frac{10000}{\sqrt{10000}/(17K)} + \int_{10000}^{10000+O(1)} \frac{1}{\sqrt{x}/(17K)}\,\mathrm{d}x = O(K), \end{aligned}$$

(still using the definition of h from the proof of Theorem 2), so that the probability under consideration is

$$e^{(O(K) - K \log^2 n)/O(K)} = e^{-\Omega(\log^2 n)} = n^{-\omega(1)}$$

as claimed. □

We now prove Theorem 5.

Proof of Theorem 5 Applying Lemma 4 as in the proof of Theorem 2, the probability of all p_i remaining above $1/3$ all the time for $n^{\gamma'}$ steps is at least $1 - n^{-\gamma'+1} \ge 1 - n^{-\kappa}$, where $\gamma' = \max\{\gamma, \kappa - 1\}$ and γ is chosen as in the proof of Theorem 2.

The aim is to apply Lemma 6 with $T^* := x_{\min}/h(x_{\min}) + \int_{x_{\min}}^{n} 1/h(x)\,\mathrm{d}x$, $t := 3T^*$ and $x_{\max} = n$. Note that T^* just represents the upper bound $O(K\sqrt{n})$ on the expected value derived from variable drift in the proof of Theorem 2. This bound is at least $T^* \ge \int_0^n 1/h(x)\,\mathrm{d}x = 17K \int_0^n x^{-1/2}\,\mathrm{d}x = 34K\sqrt{n}$. Invoking the lemma yields

$$\mathrm{P}\big[T > 3T^*\big] = \mathrm{P}[T > t] \le e^{\Omega((T^* - t/2)/n)} \le e^{-c'((T^*/2)/n)} \le e^{-c' 17K/\sqrt{n}}$$

for some constant $c' > 0$. As $K \ge c\sqrt{n}\ln n$, this means that the time is at most $3T^* = O(K\sqrt{n})$ with probability at least $1 - e^{-cc' 17 \ln n}$. This probability becomes at least $1 - n^{-\kappa}$ if c is chosen as a large enough constant.

Whenever the potential is at most 10000, we have a probability of $\Omega(1)$ to create the optimum (see proof of Theorem 2). By Lemma 7, the algorithm with high probability returns to such a state within $K \log^2 n$ steps. Applying these arguments $\log^2 n$ times (and considering failure probabilities for $\log^2 n$ applications of

Lemma 7), the probability that after $K\log^4 n$ steps the optimum has not been found is $(\log^2 n)\cdot e^{-\Omega(\log^2 n)} = n^{-\omega(1)}$.

Adding up all failure probabilities yields the claimed result. □

5 Large Update Strengths Lead to Genetic Drift

The bound $O(\sqrt{n}K)$ from Theorem 2 shows that larger update strengths (i. e., smaller K) result in smaller bounds on the runtime. However, the theorem requires that $K \geq c\sqrt{n}\log n$ so that the best possible choice results in $O(n\log n)$ runtime. An obvious question to ask is whether this is only a weakness of the analysis or whether there is an intrinsic limit that prevents smaller choices of K from being efficient.

In this section, we will show that smaller choices of K (i. e., larger update strengths) cannot give runtimes of lower orders than $n\log n$. In a nutshell, even though larger update strengths support faster exploitation of correct decisions at single bits by quickly reinforcing promising bit values they also increase the risk of genetic drift reinforcing incorrectly made decisions at single bits too quickly. Then it typically happens that several marginal probabilities reach their lower border $1/n$, from which it (due to so-called coupon collector effects) takes $\Omega(n\log n)$ steps to “unlearn” the wrong settings. The very same effect happens with 2-MMAS$_{\text{ib}}$ if its update strength ρ is chosen too large.

We now state the lower bounds we obtain for the two algorithms, see Theorems 8 and 9 below. Note that the statements are identical if we identify the update strength $1/K$ of cGA with the update strength ρ of 2-MMAS$_{\text{ib}}$. Also the proofs of these two theorems will largely follow the same steps. Therefore, we describe the proof approach in detail with respect to cGA in Sect. 5.1. In Sect. 5.2, we describe the few places where slightly different arguments are needed to obtain the result for 2-MMAS$_{\text{ib}}$.

Theorem 8 *The optimization time of cGA with* $K \leq \text{poly}(n)$ *is* $\Omega(\sqrt{n}K + n\log n)$ *with probability* $1 - \text{poly}(n)\cdot 2^{-\Omega(\min\{K, n^{1/2-o(1)}\})}$ *and in expectation.*

Theorem 9 *The optimization time of 2-MMAS*$_{ib}$ *with* $1/\rho \leq \text{poly}(n)$ *is* $\Omega(\sqrt{n}/\rho + n\log n)$ *with probability* $1 - \text{poly}(n)\cdot 2^{-\Omega(\min\{1/\rho, n^{1/2-o(1)}\})}$ *and in expectation.*

We first describe at an intuitive level why large update strengths can be risky. In the upper bounds from Theorems 1 and 2, we have shown that for sufficiently small update strengths, the positive stochastic drift by b-steps is strong enough such that even in the presence of rw-steps *all* bits never reach marginal probabilities below $1/3$, with high probability. Then no “incorrect” decision is made.

With larger update strengths than $1/(\sqrt{n}\log n)$ the effect of rw-steps is strong enough such that with high probability *some* bits will make an incorrect decision and reach the lower borders of marginal probabilities.

More specifically, the lower bounds of $\Omega(n\log n)$ in Theorems 8 and 9 will be established from the following arguments. We show that many marginal probabilities will remain close to their initial values during the early stages of a run (Lemmas 13 and 15). This then implies that b-steps will be rare (Lemma 12) throughout this time, and thus genetic drift dominates. Through a detailed analysis of the distribution of first hitting times in rw-steps we show that then some marginal probabilities will hit the lower border (Lemmas 10 and 16). Finally, we show that once sufficiently many marginal probabilities have reached the lower border, then this implies a lower bound of $\Omega(n \log n)$ as claimed (Lemma 14).

5.1 Proof of Lower Bound for cGA

We start with a detailed analysis of the hitting time for a marginal probability to reach the lower border $1/n$ and the distribution hitting times.

To illustrate this setting, fix one bit and imagine that all steps were rw-steps (we will explain later how to handle b-steps), and that all rw-steps change the current value of the bit's marginal probability (i. e., there are no self-loops). Then the process would be a fair random walk on $\{0, 1/K, 2/K, \dots, (K-1)/K, 1\}$, started at $1/2$. This fair random walk is well understood (see, e. g., Chapter 14.3 in [9]) and it is well known that the hitting time is not sharply concentrated around the expectation. More precisely, there is still a polynomially in K small probability of hitting a border within at most $O(K^2/\log K)$ steps and also of needing at least $\Omega(K^2 \log K)$ steps. The underlying idea is that the central limit theorem (CLT) approximates the progress within a given number of steps.

The real process is more complicated because of self-loops. Recall from the definition of F_t that the process only changes its current state by $\pm 1/K$ with probability $2p_{t,i}(1-p_{t,i})$, hence with probability $1-2p_{t,i}(1-p_{t,i})$ a self-loop occurs on this bit. The closer the process is to one of its borders $\{1/n, 1-1/n\}$, the larger the self-loop probability becomes and the more the random walk slows down. Hence the actual process is clearly slower in reaching a border since every looping step is just wasted. One might conjecture that the self-loops will asymptotically increase the expected hitting time. But interestingly, as we will show, the expected hitting time in the presence of self-loops is still of order $\Theta(K^2)$. Also the CLT (in a generalized form) is still applicable despite the self-loops, leading to a similar distribution as above.

The distribution of the hitting time of the random walk with self-loops will be analyzed in Lemma 10 below. In order to deal with self-loops, in its proof, we use a potential function mapping the actual process to a process on a scaled state space with nearly position-independent variance. Unlike the typical applications of potential functions in drift analysis, the purpose of the potential function is not to establish a position-independent first-moment stochastic drift but a (nearly) position-independent variance, i. e., the potential function is designed to analyze a second moment. This argument seems to be new in the theory of drift analysis and may be of independent interest.

Lemma 10 *Consider a bit of cGA on* OneMax *and let p_t be its marginal probability at time t. Let $t_1, t_2, \dots$ be the times where cGA performs an rw-step (before hitting one of the borders $1/n$ or $1-1/n$) and let $\Delta_i := p_{t_i+1} - p_{t_i}$. For $s \in \mathbb{R}$, let T_s be the smallest t such that $\operatorname{sgn}(s)\left(\sum_{i=0}^{t} \Delta_i\right) \geq |s|$ holds.*

Choosing $0 < \alpha < 1$*, where* $1/\alpha = o(K)$*, and* $-1 \le s < 0$ *constant, we have*

$$\mathrm{P}\Big[T_s \le \alpha(sK)^2 \textit{ or } \ p_t \ \textit{ exceeds } \ 5/6 \textit{ or reaches } 1/n \textit{ before } t_{T_s}\Big]$$
$$\Big(\frac{1}{13\sqrt{1/(|s|\alpha)}} - \frac{1}{(13\sqrt{1/(|s|\alpha)})^3}\Big)\frac{1}{\sqrt{2\pi}}e^{-\frac{169}{2|s|\alpha}} - O\Big(\frac{1}{|s|\sqrt{\alpha K}}\Big).$$

Moreover, for any $\alpha > 0$ *and* $s \in \mathbb{R}$*,*

$$\mathrm{P}\Big[T_s \ge \alpha(sK)^2 \textit{ or a border is reached until time } \ t_{\alpha(sK)^2}\Big] \ge 1 - e^{-1/(4\alpha)}.$$

Informally, the lemma means that every deviation of the hitting time T_s by a constant factor from its expected value (which turns out as $\Theta(s^2K^2)$) still has constant probability, and even deviations by logarithmic factors have a polynomially small probability. We will mostly apply the lemma for $\alpha < 1$, especially $\alpha \approx 1/\log n$, to show that there are marginal probabilities that quickly approach the lower border; in fact, this effect implies that the smallest possible update strength $K \sim \sqrt{n}\log n$ in Theorem 2 necessarily involves a $\log n$-term. Note that the second statement of the lemma also holds for $\alpha \ge 1$; however, in this realm also Markov's inequality works. Then, by the inequality $e^{-x} \le 1 - x/2$ for $x \le 1$, we get $\mathrm{P}\big[T_s \ge \alpha(sK)^2\big] \ge 1/(8\alpha)$, which means that Markov's inequality for deviations above the expected value is asymptotically tight in this case.

We start with the proof of the second statement, which is can be obtained by a relatively straightforward analysis of a fair random walk.

Proof of Lemma 10, 2nd statement Throughout this proof, to ease notation we consider the scaled process on the state space $S := \{0, 1, \dots, K\}$ obtained by multiplying all marginal probabilities by K; the random variables $X_t = Kp_t$ will live on this scaled space. Note that we also remove the borders (K/n and $K - K/n$), which is possible as all considerations are stopped when such a border is reached. For the same reason, we only consider current states from $\{1, \dots, K-1\}$ in the remainder of this proof.

The first hitting time T_s becomes only stochastically larger if we ignore all self-loops. Formally, recalling the trivial scaling of the state space, we consider the fair random walk where $\mathrm{P}\big[X_{t_i+1} = j-1\big] = \mathrm{P}\big[X_{t_i+1} = j+1\big] = 1/2$ if $X_{t_i} = j \in \{1, \dots, K-1\}$. We write $Y_t = \sum_{i=0}^{t-1} \Delta_{t_i}$. Clearly, Δ_i is uniform on $\{-1, 1\}$, $\mathrm{E}\big(\Delta_i \mid 0 < X_{t_i} < K\big) = 0$, $\mathrm{Var}(\Delta_i \mid 0 < X_{t_i} < K) = 1$ and Y_t is a sum of independent, identically distributed random variables. It is well known that $(Y_t - \mathrm{E}(Y_t))/\sqrt{\mathrm{Var}(Y_t)}$ converges in distribution to a standard normally distributed random variable (see, e. g., Chapter 10 in [9]). However, we do not use this fact directly here. Instead, to bound the deviation from the expectation, we use a classical Hoeffding bound. We assume $s \ge 0$ now and will see that the case $s < 0$ can be handled symmetrically.

Theorem 1.11 in [4] yields, with $c_i = 2$ as the size of the support of Δ_i, that

$$\mathrm{P}\big[Y_{\alpha s^2K^2} \ge sK\big] \le e^{-(sK)^2/(4\alpha s^2K^2)} = e^{-1/(4\alpha)}.$$

Moreover, according to Theorem 1.13 in [4], the bound also holds for all $k \le \alpha s^2 K^2$ together, more precisely,

$$\mathrm{P}\Big[\exists k \le \alpha s^2 K^2 \colon Y_k \ge sK\Big] \le e^{-1/(4\alpha)}.$$

Symmetrically, we obtain

$$\mathrm{P}\Big[\exists k \le \alpha s^2 K^2 \colon Y_k \le -sK\Big] \le e^{-1/(4\alpha)}.$$

Hence, a distance that is strictly smaller than sK is bridged through $\alpha(sK)^2$ rw-steps (or the process reaches a border before) with probability at least $1 - e^{-1/(4\alpha)}$. □

To illustrate the main idea for the proof of the first statement Lemma 10, we ignore b-steps for a while and recall that we are confronted with a fair random walk then. However, the random walk is not homogeneous with respect to place as the self-loops slow the process down in the vicinity of a border. Unlike the classical fair random walk, the random variables describing the change of position from time t to time $t+1$ (formally, $\Delta_t := p_{t+1} - p_t$) are not identically distributed. In fact, the variance of Δ_t becomes smaller the closer p_t is to one of the borders.

In more detail, the potential function used in the proof of Lemma 10 will essentially use the self-loop probabilities to construct extra distances to bridge. For instance, states with low self-loop probability (e. g., $1/2$), will have a potential that is only by $\Theta(1)$ larger or smaller than the potential of its neighbors. On the other hand, states with a large self-loop probability, say $1/K$, will have a potential that can differ by as much as $2\sqrt{K}$ from the potential of its neighbors. Interestingly, this choice leads to variances of the one-step changes that are basically the same on the whole state space (very roughly, this is true since the squared change $(2\sqrt{K})^2 = \Theta(K)$ is observed with probability $\Theta(1/K)$). However, using the potential for this trick is at the expense of changing the support of the underlying random variables, which then will depend on the state. Nevertheless, as the support is not changed too much, the Central Limit Theorem (CLT) still applies and we can approximate the progress made within T steps by a normally distributed random variable. This approximation is made precise in the following lemma, along with a bound on the absolute error.

Lemma 11 (CLT with Lyapunov condition, Berry-Esseen inequality [10], p. 544). *Let $X_1, \dots, X_m$ be a sequence of independent random variables, each with finite expected value μ_i and variance σ_i^2. Define*

$$s_m^2 := \sum_{i=1}^{m} \sigma_i^2 \quad \textit{and} \quad C_m := \frac{1}{s_m} \sum_{i=1}^{m} (X_i - \mu_i)\,.$$

If there exists a $\delta > 0$ such that

$$\lim_{m\to\infty} \frac{1}{s_m^{2+\delta}} \sum_{i=1}^{m} \mathrm{E}\Big(|X_i - \mu_i|^{2+\delta}\Big) = 0$$

(assuming all the moments of order $2 + \delta$ to be defined), then C_m converges in distribution to a standard normally distributed random variable.

Moreover, the approximation error is bounded as follows: for all $x \in \mathbb{R}$,

$$|P[C_m \le x] - \Phi(x)| \le C \cdot \frac{\sum_{i=1}^{m} \mathrm{E}\big(|X_i - \mu_i|^3\big)}{s_m^3}$$

where C is an absolute constant and $\Phi(x)$ denotes the cumulative distribution function of the standard normal distribution.

We now turn to the formal proof of the outstanding 1st statement of Lemma 10.

Proof of Lemma 10, 1st statement As in the proof of the 2nd statement of Lemma 10 above, we consider the scaled search space $\{1, \dots, K-1\}$. Here we will essentially use an approximation of the accumulated state within $\alpha s^2 K^2$ steps by the normal distribution, but have to be careful to take into account steps describing self-loops. To analyze the hitting time T_s for the X_{t_i}-process, we now define a potential function $g \colon S \to \mathbb{R}$. Unlike the typical applications of potential functions, the purpose of g is not to establish a position-independent first-moment drift (in fact, there is no drift within S since the original process is a martingale) but a (nearly) position-independent variance, i. e., the potential function is designed to analyze a second moment.

Potential function We proceed with the formal definition of the potential function, the analysis of its expected first-moment change and the corresponding variance, and a proof that the Lyapunov condition holds for the accumulated change within $\alpha s^2 K^2$ steps. The potential function g is monotonically decreasing on $\{1, \dots, K/2\}$ and centrally symmetric around $K/2$. We define it as follows:

$$\begin{aligned} g(K/2) &:= 0 && (3)\\ g(i) - g(i+1) &:= \sqrt{2K/(i+1)} \text{ for } 1 \le i \le K/2 - 1, && (4)\\ g(K-i) &:= -g(i) \text{ for } i \ge K/2. && (5) \end{aligned}$$

Inductively, we have for $1 \le i \le K/2$ that

$$g(i) = g(i) - g(K/2) = \sum_{j=i}^{K/2-1} (g(j) - g(j+1)) = \sum_{j=i}^{K/2-1} \sqrt{2K/(j+1)},$$

where the second equality holds since the sum is telescoping. We also note that $g(0) = O(K)$, more precisely it holds that

$$\begin{aligned} g(0) &= \sqrt{2K} \left(\sum_{j=1}^{K/2} \sqrt{1/j} \right) \le \sqrt{2K} \left(1 + \int_1^{K/2-1} \frac{1}{\sqrt{x}} \mathrm{d}x \right) \\ &\le \sqrt{2K} (2\sqrt{K/2}) = 2K, \end{aligned}$$

where the first inequality used $\sum_{j=2}^{K/2-1}\sqrt{1/j}$ as a lower sum of the integral. More generally, using the monotonicity of g and the same kind of estimations as before, we obtain for $i < j \leq K/2$ that

$$g(i) - g(j) \leq g(0) - g(j-i) = \sqrt{2K}\sum_{k=1}^{j-i}\sqrt{1/k} \leq 2\sqrt{2K}(\sqrt{j-i}) \tag{6}$$

Informally, the potential function stretches the whole state space by a factor of at most 4 but adjacent states in the vicinity of borders can be by $2\sqrt{K}$ apart in potential.

Let $Y_t := g(X_t)$. We consider the one-step differences $\Psi_i := Y_{t_i+1} - Y_{t_i}$ at the times i where rw-steps occur, and we will show via the representation $Y_{t_i} := \sum_{j=0}^{i-1}\Psi_j$ that Y_{t_i} approaches a normally distributed variable. Note that Y_{t_i} is not necessarily the same as $g(X_{t_i}) - g(X_{t_0})$ since only the effect of rw-steps is covered by Y_{t_i}.

In the following, we assume $1 \leq X_{t_i} \leq K/2$ and note that the case $X_{t_i} > K/2$ can be handled symmetrically with respect to $-\Psi_i$. We proceed with the announced analysis of different moments of Ψ_i.

Analysis of expected change of potential We claim that for all $i \geq 0$

$$0 \leq \mathrm{E}\big(\Psi_i \mid X_{t_i}\big) \leq \sqrt{2/(X_{t_i}K)} \leq o(1), \tag{7}$$

where the o-notation is with respect to K.

The lower bound $\mathrm{E}\big(\Psi_i \mid X_{t_i}\big) \geq 0$ is easy to see since X_{t_i} is a fair random walk and $g(j-1) - g(j) \geq g(j) - g(j+1)$ holds for all $j \leq K/2$. To prove the upper bound, we note that $X_{t_i+1} \in \{X_{t_i} - 1, X_{t_i}, X_{t_i} + 1\}$ so that

$$\begin{aligned}\mathrm{E}\big(\Psi_i \mid X_{t_i}\big) = {} & \mathrm{P}\big[X_{t_i+1} < X_{t_i}\big](g(X_{t_i} - 1) - g(X_{t_i}))\\ & + \mathrm{P}\big[X_{t_i+1} > X_{t_i}\big](g(X_{t_i} + 1) - g(X_{t_i}))\end{aligned}$$

Using the properties of rw-steps, we have that $\mathrm{P}\big[Y_{t_i+1} \neq Y_{t_i}\big] = 2\frac{(K-X_{t_i})X_{t_i}}{K^2}$. Moreover, on $Y_{t_i+1} \neq Y_{t_i}$, Y_{t_i+1} takes each of the two values $g(X_{t_i} - 1)$ and $g(X_{t_i} + 1)$ with the same probability. Hence

$$\begin{aligned}\mathrm{E}\big(\Psi_i \mid X_{t_i}\big) &= \frac{(K - X_{t_i})X_{t_i}}{K^2}\big((g(X_{t_i} - 1) - g(X_{t_i})) + (g(X_{t_i} + 1) - g(X_{t_i}))\big)\\ &= \frac{(K - X_{t_i})X_{t_i}}{K^2}\big((g(X_{t_i} - 1) - g(X_{t_i})) - (g(X_{t_i}) - g(X_{t_i} + 1))\big)\\ &= \frac{(K - X_{t_i})X_{t_i}}{K^2}\cdot\sqrt{2K}\left(\frac{1}{\sqrt{X_{t_i}}} - \frac{1}{\sqrt{X_{t_i} + 1}}\right)\\ &\leq \frac{X_{t_i}}{K}\cdot\sqrt{2K}\left(\frac{1}{\sqrt{X_{t_i}}} - \frac{1}{\sqrt{X_{t_i} + 1}}\right),\end{aligned}$$

where the last equality used (4).

We estimate the bracketed terms using

$$\frac{1}{\sqrt{X_{t_i}}} - \frac{1}{\sqrt{X_{t_i}+1}} = \frac{\sqrt{X_{t_i}+1} - \sqrt{X_{t_i}}}{\sqrt{X_{t_i}}\sqrt{X_{t_i}+1}} \le \frac{1/(2\sqrt{X_{t_i}})}{X_{t_i}} \le \frac{1}{\left(X_{t_i}\right)^{3/2}},$$

where the penultimate inequality exploited that $f(x+h) - f(x) \le hf'(x)$ for any concave, differentiable function f and $h \ge 0$; here using $f(x) = \sqrt{x}$ and $h = 1$. Altogether,

$$\mathrm{E}\big(\Psi_i \mid X_{t_i}\big) \le \frac{X_{t_i}}{K} \cdot \frac{\sqrt{2K}}{\left(X_{t_i}\right)^{3/2}} = \frac{\sqrt{2}X_{t_i}}{\sqrt{K}\left(X_{t_i}\right)^{3/2}} = \sqrt{\frac{2}{X_{t_i}K}},$$

which proves (7) since $X_{t_i} \ge 1$ and $K = \omega(1)$.
Analysis of the variance of the change of potential We claim that for all $i \ge 0$

$$\mathrm{Var}(\Psi_i \mid X_{t_i}) \ge 1/4. \tag{8}$$

To show this, note that

$$\begin{aligned}\mathrm{Var}(\Psi_i \mid X_{t_i}) &\ge \mathrm{E}\Big((\Psi_i - \mathrm{E}\big(\Psi_i \mid X_{t_i}\big))^2 \cdot \mathbb{1}\{\Psi_i \le 0\} \mid X_{t_i}\Big)\\ &\ge \mathrm{E}\Big((\Psi_i)^2 \cdot \mathbb{1}\{\Psi_i \le 0\} \mid X_{t_i}\Big)\end{aligned}$$

since $\mathrm{E}\big(\Psi_i \mid X_{t_i}\big) \ge 0$. Now, as $0 < X_{t_i} \le K/2$, we have $\mathrm{P}\big[Y_{t_i+1} < Y_{t_i}\big] = \frac{(K-X_{t_i})X_{t_i}}{K^2} \ge \frac{X_{t_i}}{2K}$. Moreover, $Y_{t_i+1} < Y_{t_i}$ implies that $X_{t_i+1} = X_{t_i} + 1$ since g is monotone decreasing on $\{1, \dots, K/2\}$ and the X_{t_i}-value can change by either -1, 0, or 1. Hence, if $Y_{t_i+1} < Y_{t_i}$ then $Y_{t_i+1} - Y_{t_i} = g(X_{t_i}+1) - g(X_{t_i}) = -\sqrt{2K/(X_{t_i}+1)}$. Altogether,

$$\mathrm{Var}(\Psi_i \mid X_{t_i}) \ge \frac{X_{t_i}}{2K} \cdot \left(-\sqrt{2K/(X_{t_i}+1)}\right)^2 \ge 1/4,$$

where we used $X_{t_i}/(X_{t_i}+1) \ge 1/2$. This proves the lower bound on the variance.

Approximating the accumulated change of potential by a Normal distribution We are almost ready to prove that $Y_{t_i} := \sum_{j=0}^{i-1} \Psi_j$ can be approximated by a normally distributed random variable for sufficiently large t. We denote by $s_i^2 := \sum_{j=0}^{i-1} \mathrm{Var}(\Psi_j \mid X_{t_j})$ and note that $s_i^2 \ge i/4$ by our analysis of variance from above. The so-called Lyapunov condition, which is sufficient for convergence to the normal distribution (see Lemma 11), requires the existence of some $\delta > 0$ such that

$$\lim_{i\to\infty} \frac{1}{s_i^{2+\delta}} \sum_{j=0}^{i-1} \mathrm{E}\Big(|\Psi_j - \mathrm{E}\big(\Psi_j \mid X_{t_j}\big)|^{2+\delta} \mid X_{t_j}\Big) = 0. \tag{9}$$

We will show that the condition is satisfied for $\delta = 1$ (smaller values could be used but do not give any benefit) and $i = \omega(K)$ (which, as $i = \alpha s^2 K^2$, holds due to our assumptions $1/\alpha = o(K)$ and $|s| = \Omega(1)$). We argue that

$$\begin{aligned} |\Psi_i - \mathrm{E}(\Psi_i \mid X_{t_i})| &\le |\Psi_i| + |\mathrm{E}(\Psi_i \mid X_{t_i})| \\ &\le \left|\max\left\{k \mid \mathrm{P}\left[|\Psi_i| \ge k \mid X_{t_i}\right] > 0\right\}\right| + o(1), \end{aligned}$$

where we have used the bound on $|\mathrm{E}(\Psi_i \mid X_{t_i})|$ from (7). As the X_{t_i}-value can only change by $\{-1, 0, 1\}$, we get, by summing up all possible changes of the g-value, that

$$\begin{aligned} |\Psi_i - \mathrm{E}(\Psi_i \mid X_{t_i})| &\le (g(X_{t_i} - 1) - g(X_{t_i})) + (g(X_{t_i}) - g(X_{t_i} + 1)) + o(1) \\ &\le g(X_{t_i} - 1) - g(X_{t_i} + 1) + o(1) \\ &\le \left(2 \cdot \sqrt{2K/(X_{t_i} - 1)}\right) + o(1) \end{aligned}$$

for K large enough.

Hence, plugging this in the Lyapunov condition (9) for $\delta = 1$, we obtain

$$\begin{aligned} &\mathrm{E}\left(|\Psi_j - \mathrm{E}(\Psi_j \mid X_{t_j})|^3 \mid X_{t_j}\right) \\ &\le \frac{2X_{t_j}}{K}\left(2 \cdot \sqrt{2K/(X_{t_j} - 1)}\right)^3 (1 + o(1)) + o(1) = O(\sqrt{K}), \end{aligned}$$

implying that

$$\frac{1}{s_i^3} \sum_{j=0}^{i-1} \mathrm{E}\left(|\Psi_j - \mathrm{E}(\Psi_j)|^3 \mid X_{t_j}\right) \le \frac{1}{(i/4)^{1.5}} O(i\sqrt{K}) = O(\sqrt{K/i}), \tag{10}$$

which goes to 0 as $i = \omega(\sqrt{K})$. Hence, for the value $i := \alpha s^2 K^2$ considered in the lemma we obtain that

$$\frac{Y_{t_i} - \mathrm{E}(Y_{t_i} \mid X_0)}{s_i} \tag{11}$$

converges in distribution to $N(0, 1)$ according to Lemma 11. The absolute error of this approximation is also $O(\sqrt{K/i})$ by reusing (10).

Estimating the accumulated progress Recall that our aim is to show that the event $\sum_{j=0}^{i-1} \Delta_j \le s$ (where s is negative and $i = \alpha s^2 K^2$) happens with at least the probability stated in the lemma. Since we analyzed the change of the potential function g, we establish a sufficient *increase* of the g-value (corresponding to a decrease of marginal probability) that implies $\sum_{j=0}^{i-1} \Delta_j \le s$. By (6), we know that $g(X_{t_i}) - g(X_0) \ge 2\sqrt{|s|}K$ implies $X_{t_i} - X_0 \le sK < 0$ and therefore also $\sum_{j=0}^{i-1} \Delta_j \le s$. Hence, in the following it suffices to study the event $g(X_{t_i}) - g(X_0) \ge 2\sqrt{|s|}K$ and to show that it happens with the required probability.

As already mentioned, the random variable Y_{t_i} denotes the accumulated progress (in terms of g-value) due to rw-steps up to time t_i. To show that Y_{t_i} is at least $2\sqrt{|s|}K$ with the claimed probability bounds, we exploit the above-established property that (11) converges in distribution to $N(0, 1)$. Hence, we need to estimate the variance s_i and the expected value $\mathrm{E}\big(Y_{t_i}\big)$.

Note that $s_i^2 \geq \alpha s^2 K^2/4$ by our analysis of variance above and therefore $s_i \geq \sqrt{\alpha}|s|K/2$. We have to be more careful when computing $\mathrm{E}\big(Y_{t_i}\big)$ since $\mathrm{E}\big(\Psi_i \mid X_{t_i}\big)$ is negative for $X_{t_i} > K/2$. Note, however, that considerations are stopped when the marginal probability exceeds 5/6, i. e., when $X_{t_i} > 5K/6$. Using (7), we hence have that $\mathrm{E}\big(\Psi_i \mid X_{t_i}\big) \geq -\sqrt{2/(5K^2/6)} \geq -1.55/K$. Therefore, $\mathrm{E}\big(Y_{t_i}\big) \geq i\cdot(-1.55/K) = -1.55\alpha s^2 K$ and $\mathrm{E}\big(Y_{t_i}/s_i\big) \geq -3.1|s|\sqrt{\alpha}$.

We study the event $Y_{t_i} \geq rK$ for general $r \geq 0$, which is equivalent to $\frac{Y_{t_i}-\mathrm{E}(Y_{t_i}|X_0)}{s_i} \geq rK/s_i - \mathrm{E}\big(Y_{t_i}/s_i\big)$. If (11) was really $N(0, 1)$-distributed, the probability of the event would be $\Phi(rK/s_i - \mathrm{E}\big(Y_{t_i}/s_i\big))$, where Φ denotes the cumulative distribution function of the standard normal distribution. Taking into account the approximation error $O(\sqrt{K/i})$ computed above and plugging in our estimates for expected value and variance, we altogether have that

$$
\begin{aligned}
&\mathrm{P}\big[Y_{t_i} \geq rK\big] \\
&\quad \geq \Big(1 - \Phi\big(rK/s_i - \mathrm{E}\big(Y_{t_i}/s_i\big)\big)\Big) - O(\sqrt{K}/i) && (12)\\
&\quad = 1 - \Phi\Big(r/(|s|\sqrt{\alpha/4}) + 3.1|s|\sqrt{\alpha}\Big) - O(\sqrt{K}/i) && (13)
\end{aligned}
$$

for any r leading to a positive argument of Φ,

Using $r = 3\sqrt{|s|}$ in (13) , we compute

$$\frac{r}{|s|\sqrt{\alpha/4}} + 3.1|s|\sqrt{\alpha} \leq \frac{3\sqrt{|s|}}{|s|\sqrt{\alpha/4}} + 3.1|s|\sqrt{\alpha} \leq \frac{13}{\sqrt{|s|\alpha}}.$$

Using Lemma 21 (in the Appendix) we can now bound the term $1 - \Phi\Big(r/(|s|\sqrt{\alpha/4}) + 3.1|s|\sqrt{\alpha}\Big)$ from (13) below and obtain

$$\left(\frac{1}{13\sqrt{1/(|s|\alpha)}} - \frac{1}{(13\sqrt{1/(|s|\alpha)})^3}\right)\frac{1}{\sqrt{2\pi}}e^{-169/(2|s|\alpha)} =: p(\alpha, s),$$

using $|s| \leq 1$ and $\alpha \leq 1$. This means that distance sK (in negative direction) is bridged by the rw-steps before or at time t_i, where $i = \alpha s^2 K^2$, with probability at least $p(\alpha, s) - O(\sqrt{K/i}) = p(\alpha, s) - O(\alpha^{-1/2} s^{-1} K^{-1/2})$, where the O-term is the bound on the approximation error computed above. Undoing the scaling of the state space introduced at the beginning of this proof, this corresponds to an accumulated change of the actual state of cGA in rw-steps by s; more formally, $\big(\sum_{i=0}^{t} \Delta_i\big) \leq s$ in terms of the original state space. This establishes also the first statement of the lemma and completes the proof. □

As rw-steps are interleaved with b-steps, Lemma 10 alone is not sufficient to analyze the overall movement of a marginal probability. We also requires a bounded number of b-steps within a given period of time. To establish this, we first show that, during the early stages of a run, the probability of a b-step is only $O(1/\sqrt{n})$. Intuitively, during early stages of the run many bits will have marginal probabilities in the interval $[1/6, 5/6]$. Then the standard sampling deviation of the OneMax-value is of order $\Theta(\sqrt{n})$, and the probability of a b-step is $1 - \mathrm{P}[R_t] = O(1/\sqrt{n})$. The link between $1 - \mathrm{P}[R_t]$ and the standard deviation already appeared in Lemma 3 above; roughly, it says that every step is a b-step for bit i with probability at least $(\sum_{j \neq i} p_j(1-p_j))^{-1/2}$, which is the reciprocal of the standard deviation in terms of the other bits.

The following Lemma 12 represents a kind of counterpart of Lemma 3, but here we seek an *upper* bound on $1 - \mathrm{P}[R_t]$.

Lemma 12 *Assume that at time t there are γn bits for some constant $\gamma > 0$ bits whose marginal probabilities are within* $[1/6, 5/6]$. *Then the probability of having a b-step on any fixed bit position is*

$$1 - \mathrm{P}[R_t] = O(1/\sqrt{n}),$$

regardless of the decisions made in this step on all other $n - \gamma n - 1$ bits.

Proof We know from our earlier discussion that a b-step at bit i requires $D_t \in \{-1, 0\}$ where $D_t := |x_t| - |x_{t,i}| - (|y_t| - |y_{t,i}|)$ is the change of the OneMax-value at bits other than i in the two solutions x_t and y_t sampled at time t.

We apply the principle of deferred decisions and fix all decisions for creating x_t as well as decisions for y_t on all but the $m := \gamma n$ selected bits with marginal probabilities in $[1/6, 5/6]$. Let $p_1, p_2, \dots, p_m$ denote the corresponding marginal probabilities after renumbering these bits, and let S denote the random number of these bits set to 1. Note that there are at most 2 values for S which lead to the algorithm making a b-step.

Since S is determined by a Bernoulli trial with success probabilities $p_1, \dots, p_m$, Theorem 22 in the Appendix implies that the probability of S attaining any particular value is at most

$$\frac{1}{2\sqrt{\sum_{i=1}^m p_i(1-p_i)}} \le \frac{1}{2\sqrt{\sum_{i=1}^m (1/6) \cdot (5/6)}} = O(1/\sqrt{m}) = O(1/\sqrt{n}).$$

Taking the union bound over 2 values proves the claim. □

Even though one main aim is to show that rw-steps make certain marginal probabilities reach their lower border, we will also ensure that with high probability, $\Omega(n)$ marginal probabilities do not move by too much, resulting in a large sampling variance and a small probability of b-steps. The following lemma serves this purpose. Its proof is a straightforward application of Hoeffding's inequality since it is pessimistic here to ignore the self-loops.

Lemma 13 *For any bit, with probability $\Omega(1)$ for any $t \le \kappa K^2$, $\kappa > 0$ a small enough constant, the first t rw-steps lead to a total change of the bit's marginal probability within* $[-1/6, 1/6]$. *This fact holds independently of all other bits.*

The probability that the above holds for less than γn bits amongst the first $n/2$ bits is $2^{-\Omega(n)}$, regardless of the decisions made on the last $n/2$ bits.

Proof Note that the probability of exceeding $[-1/6, 1/6]$ increases with the number of rw-steps that do increase or decrease the marginal probability (as opposed to self-loops). We call these steps *relevant* and pessimistically assume that all t steps are relevant steps.

Now defining $X_j := \sum_{i=1}^{j} X_i$ as the total progress in the first j relevant steps, we have $\mathrm{E}\big(X_j\big) = 0$, for all $j \le t$, and the total change in these j steps exceeds $1/6$ only if $X_j \ge K/6$. Applying a Hoeffding bound, Theorem 1.13 in [4], the maximum total progress is bounded as follows:

$$\mathrm{P}\left[\max_{j \le t} X_j \le K/6\right] \le \exp\left(\frac{-2(K/6)^2}{4t}\right) \le \exp\left(-\frac{1}{12\kappa}\right).$$

By symmetry, the same holds for the total change reaching values less or equal to $-1/6$. By the union bound, the probability that the total change always remains within the interval $[-1/6, 1/6]$ is thus at least

$$1 - 2\exp\left(-\frac{1}{12\kappa}\right).$$

Assuming $\kappa < 1/(12\ln 2)$ gives a lower bound of $\Omega(1)$.

Note that due to our pessimistic assumption of all steps being relevant, all bits are treated independently. Hence we may apply standard Chernoff bounds to derive the second claim. □

The following lemma shows that whenever a small number of bits has reached the lower border for marginal probabilities, the remaining optimization time is $\Omega(n \log n)$ with high probability. The proof is similar to the well known coupon collector's theorem [24].

Lemma 14 *Assume cGA reaches a situation where at least $\Omega(n^{\varepsilon})$ marginal probabilities attain the lower border $1/n$. Then with probability $1 - e^{-\Omega(n^{\varepsilon/2})}$, and in expectation, the remaining optimization time is $\Omega(n \log n)$.*

Proof Let $m = \Omega(n^{\varepsilon})$ be the number of bits that have reached the lower border $1/n$. A necessary condition for reaching the optimum within $t := (n/2 - 1) \cdot (\varepsilon/2) \ln n$ iterations is that during this time each of these m bits is sampled at value 1 in at least one of the two search points constructed. The probability that one bit never samples a 1 in t iterations is at least $(1 - 2/n)^t$. The probability that all m bits sample a 1 during t steps is at most, using $(1 - 2/n)^{n/2-1} \ge 1/e$ and $1 + x \le e^x$ for $x \in \mathbb{R}$,

$$\left(1 - \left(1 - \frac{2}{n}\right)^t\right)^m \le \left(1 - n^{-\varepsilon/2}\right)^m \le \left(\exp\left(-n^{-\varepsilon/2}\right)\right)^m \le \exp(-\Omega(n^{\varepsilon/2})).$$

Hence with probability $1-\exp(-\Omega(n^{\varepsilon/2}))$ the remaining optimization time is at least $t=\Omega(n\log n)$. As $1-\exp(-\Omega(n^{\varepsilon/2}))=\Omega(1)$, the expected remaining optimization time is of the same order. □

We have collected most of the machinery to prove Theorem 8. The following lemma identifies a set of bits that stay centered in a phase of $\Theta(K\min\{K,\sqrt{n}\})$ steps, resulting in a low probability of b-steps. Basically, the idea is to bound the accumulated effect of b-steps in the phase using Chernoff bounds: given $K/6$ b-steps, a marginal probability cannot change by more than $1/6$. Note that this applies to many, but not all bits. Later, we will see that within the phase, some of the remaining bits will reach their lower border with not too low probability.

Lemma 15 *Let $\kappa>0$ be a small constant. There exists a constant γ, depending on κ, and a selection S of γn bits among the first $n/2$ bits such that the following properties hold regardless of the last $n/2$ bits throughout the first $T:=\kappa K\cdot\min\{K,\sqrt{n}\}$ steps of cGA with $K\leq\mathrm{poly}(n)$, with probability $1-\mathrm{poly}(n)\cdot 2^{-\Omega(\min\{K,n\})}$:*

1. *the marginal probabilities of all bits in S is always within $[1/6,5/6]$ during the first T steps,*
2. *the probability of a b-step at any bit is always $O(1/\sqrt{n})$ during the first T steps, and*
3. *the total number of b-steps for each bit is bounded by $K/6$, leading to a displacement of at most $1/6$.*

Proof The first property is trivially true at initialization, and we show that an event of exponentially small probability needs to occur in order to violate the property. Taking a union bound over all T steps ensures that the property holds throughout the whole phase of T steps with the claimed probability.

By Lemma 13, with probability $1-2^{-\Omega(n)}$, for at least γn of these bits the total effect of all rw-steps is always within $[-1/6,+1/6]$ during the first $T\leq\kappa K^2$ steps. We assume in the following that this happens and take S as a set containing exactly γn of these bits.

It remains to show that for all bits in S the total effect of b-steps is bounded by $1/6$ with high probability. Note that, while this is the case, according to Lemma 12, the probability of a b-step at every bit in S is at most $c_2/\sqrt{n}$ for a positive constant c_2. This corresponds to the second property, and so long as this holds, the expected number of b-steps in $T\leq\kappa K^2$ steps is at most $\kappa\cdot c_2 K$. Each b-step changes the marginal probability of the bit by $1/K$. A necessary condition for increasing the marginal probability by a total of at least $1/6$ is that we have at least $K/6$ b-steps amongst the first T steps. Choosing κ small enough to make $\kappa\cdot c_2 K\leq 1/2\cdot K/6$, by Chernoff bounds the probability to get at least $K/6$ b-steps in T steps is $e^{-\Omega(K)}$. In order for the first property to be violated, an event of probability $e^{-\Omega(K)}$ is necessary for any bit in S and any length of time $t\leq T$; otherwise all properties hold true.

Taking the union bound over all $T\leq\kappa K^2$ steps and all γn bits gives a probability bound of $\kappa K^2\cdot\gamma n\cdot e^{-\Omega(K)}\leq\mathrm{poly}(n)\cdot 2^{-\Omega(K)}$ for a property being violated. This proves the claim. □

Finally, we put everything together to prove our lower bound for cGA.

Proof of Theorem 8 If $K = O(1)$ then it is easy to show, similarly to Lemma 17, that each bit independently hits the lower border with probability $\Omega(1)$ by sampling only zeros. Then the result follows easily from Chernoff bounds and Lemma 14. Hence we assume in the following $K = \omega(1)$.

For $K \geq \sqrt{n}$, Lemma 15 implies a lower bound of $\Omega(K\sqrt{n})$ as then the probability of sampling the optimum in any of the first $T := \kappa K \cdot \min\{K, \sqrt{n}\}$ steps is at most $(5/6)^{\gamma n} = 2^{-\Omega(n)}$. Taking a union bound over the first T steps and adding the error probability from Lemma 15 proves the claim for a lower bound of $\Omega(K\sqrt{n})$ with the claimed probability. This proves the theorem for $K = \Omega(\sqrt{n} \log n)$ as then the $\Omega(\sqrt{n}K)$ term dominates the runtime. Hence we may assume $K = o(\sqrt{n} \log n)$ in the following and note that in this realm proving a lower bound of $\Omega(n \log n)$ is sufficient as here this term dominates the runtime.

We still assume that the events from Lemma 15 apply to the first $n/2$ bits. We now use Lemma 10 to show that some marginal probabilities amongst the last $n/2$ bits are likely to walk down to the lower border. Note that Lemma 10 applies for an arbitrary (even adversarial) mixture of rw-steps and b-steps over time. This allows us to regard the progress in rw-steps as independent between bits.

In more detail, we will apply both statements of Lemma 10 to a fresh marginal probability from the last $n/2$ bits, to prove that it walks to its lower border with a not too small probability. First we apply the second statement of the lemma for a positive displacement of $s := 1/6$ within T steps, using $\alpha := T/(sK)^2$. The random variable T_s describes the first point of time where the marginal probability reaches a value of at least $1/2 + 1/6 + s = 5/6$ through a mixture of b- and rw-steps. This holds since we work under the assumption that the b-steps only account for a total displacement of at most $1/6$ during the phase. Lemma 10 now gives us a probability of at least $1 - e^{-1/(4\alpha)} = \Omega(1)$ (using $\alpha = O(1)$) for the event that the marginal probability does not exceed $5/6$. In the following, we condition on this event.

We then revisit the same stochastic process and apply Lemma 10 again to show that, under this condition, the random walk achieves a negative displacement. Note that the event of not exceeding a certain positive displacement is positively correlated with the event of reaching a given negative displacement (formally, the state of the conditioned stochastic process is always stochastically smaller than of the unconditioned process), allowing us to apply Lemma 10 again despite dependencies between the two applications.

We now apply the first statement of Lemma 10 for a negative displacement of $s := -1$ through rw-steps within T steps, using $\alpha := T/((sK)^2)$. Since we still work under the assumption that the b-steps only account for a total displacement of at most $1/6$ during the phase, the displacement is then altogether no more than $s + 1/6 \leq -5/6$, implying that the lower border is hit as the marginal frequency does not exceed $5/6$.

The conditions on α in of Lemma 10 hold as $0 < \alpha < 1$ choosing κ small enough, and $1/\alpha = O(K/\min\{\sqrt{n}, K\}) = o(K)$ for $K = \omega(1)$. Also note that $1/\alpha = O(K/\min\{\sqrt{n}, K\}) = o(\log n)$ since $K = o(n \log n)$. Now the lemma states that the probability of the random walk reaching a displacement through rw-steps of at most s (or hitting the lower border before) is at least

$$\left(\frac{1}{13\sqrt{1/(|s|\alpha)}}-\frac{1}{(13\sqrt{1/(|s|\alpha)})^3}\right)\frac{1}{\sqrt{2\pi}}e^{-\frac{169}{2|s|\alpha}}-O\big(1/(s\sqrt{\alpha K})\big) \qquad (14)$$

To bound the last expression from below, we distinguish between two cases. If $K \le \sqrt{n}$, then $\alpha = \Omega(1)$ and (14) is at least

$$\Omega(1) - O\left(\frac{1}{\sqrt{K}}\right) = \Omega(1) - \frac{1}{\omega(1)} = \Omega(1)$$

since $K = \omega(1)$ and $s = \Theta(1)$. If $K \ge \sqrt{n}$, then with $1/\alpha = o(\log n)$ we estimate (14) from below by

$$\begin{aligned}&\Omega\left(\frac{1}{o(\sqrt{\log n})}\cdot e^{-o(\ln n)}\right) - O\big(1/(s\sqrt{\alpha K})\big)\\ &= \Omega\left(\frac{1}{o(\sqrt{\log n})}\cdot e^{-o(\ln n)}\right) - o\left(\frac{\sqrt{\log n}}{n^{1/4}}\right) \ge n^{-\beta}\,,\end{aligned}$$

for some $\beta = \beta(n) = o(1)$.

Combining with the probability of not exceeding $5/6$, which we have proved to be constant, the probability of the bit's marginal probability hitting the lower border within T steps is $\Omega(n^{-\beta})$. Hence by Chernoff bounds, with probability $1 - 2^{-\Omega(n^{1-\beta})}$, the final number of bits hitting the lower border within T steps is $\Omega(n^{1-\beta}) = \Omega(n^{1-o(1)})$.

Once a bit has reached the lower border, while the probability of a b-step is $O(1/\sqrt{n})$, the probability of leaving the bound again is $O(n^{-3/2})$ as it is necessary that either the bit is sampled as 1 at one of the offspring and a b-step happens, or in both offspring the bit is sampled at 1. So the probability that this does not happen until the $T = O(n \log n)$ steps are completed is $(1 - O(n^{-3/2}))^T \le e^{-O(\log(n)/\sqrt{n})} = o(1)$. Again applying Chernoff bounds leaves $\Omega(n^{1-o(1)})$ bits at the lower border at time T with probability $1 - 2^{-\Omega(n^{1-o(1)})}$.

Then Lemma 14 implies a lower bound of $\Omega(n \log n)$ that holds with probability $1 - 2^{-\Omega(n^{1/2-o(1)})}$. □

5.2 Proof of Lower Bound for 2-MMAS$_{\text{ib}}$

We will use, to a vast extent, the same approach as in Sect. 5.1 to prove Theorem 9. Most of the lemmas can be applied directly or with very minor changes. In particular, Lemmas 13, 14 and 15 also apply to 2-MMAS$_{\text{ib}}$ by identifying $1/K$ with ρ. Intuitively, this holds since the analyses of b-steps always pessimistically bound the absolute change of a marginal probability by the update strength ($1/K$ for cGA). This also holds with respect to the update strength ρ for 2-MMAS$_{\text{ib}}$.

To prove lower bounds on the time to hit a border through rw-steps, the next lemma is used. It is very similar to Lemma 10, except for two minor differences: first, also the accumulated effect of b-steps is included in the quantity $p_t - p_0$ analyzed in the lemma. Second, considerations are stopped when the marginal probability becomes less than ρ or more than $1-\rho$. This has technical reasons but is not a crucial restriction.

We supply an additional lemma, Lemma 17 below, that applies when the marginal probability is less than ρ. The latter lemma uses known analyses similar to so-called landslide sequences defined in [26, Section 4].

Lemma 16 *Consider a bit of 2-MMAS$_{ib}$ on* OneMax *and let p_t be its marginal probability at time t. We say that the process breaks a border at time t if* $\min\{p_t, 1-p_t\} \leq \max\{1/n, \rho\}$. *Given $s \in \mathbb{R}$ and arbitrary starting state p_0, let T_s be the smallest t such that* $\operatorname{sgn}(s)(p_t - p_0) \geq |s|$ *holds or a border is reached.*

Choosing $0 < \alpha < 1$, where $1/\alpha = o(\rho^{-1})$, and $-1 \leq s < 0$ constant, and assuming that every step is a b-step with probability at most $\rho/(4\alpha)$, we have

$$\mathrm{P}\Big[T_s \leq \alpha(s/\rho)^2 \textit{ or } p_t \textit{ exceeds } 5/6 \textit{ before } T_s\Big] \\ \geq \left(\frac{1}{24\sqrt{(1/(|s|\alpha)}} - \frac{1}{(24\sqrt{1/(|s|\alpha)})^3}\right)\frac{1}{\sqrt{2\pi}}e^{-288/(|s|\alpha)} - O\left(\frac{\sqrt{\rho}}{|s|\sqrt{\alpha}}\right).$$

Moreover, for any $\alpha > 0$ and constant $0 < s \leq 1$, if there are at most $s/(2\alpha\rho)$ b-steps until time $\alpha(s/\rho)^2$, then

$$\mathrm{P}\Big[T_s \geq \alpha(s/\rho)^2 \textit{ or a border is reached until time } \alpha(s/\rho)^2\Big] \geq 1 - e^{-1/(16\alpha)}.$$

Proof We follow similar ideas as in the proof of Lemma 10. Again, we start with the second statement, where $s \geq 0$ is assumed, and aim for applying a Hoeffding bound. We note that a marginal probability of 2-MMAS$_{\mathrm{ib}}$ can only change by an absolute amount of at most ρ in a step. Hence, the b-steps until time $\alpha(s/\rho^2)$ account for an increase of the X_t-value by at most $s/2$. With respect to the rw-steps, Theorem 1.11 from [4] can be applied with $c_i = 2\rho$ and $\lambda = s/2$.

Also for the first statement, we follow the ideas from the proof of Lemma 10. In particular, the borders stated in the lemma will be ignored as all considerations are stopped when they are reached. We will apply a potential function and estimate its first and second moment separately with respect to rw-steps and non-rw steps.

Definition of potential function Our potential function is

$$g(x) := \int_x^{1/2} \frac{1}{\rho\sqrt{z}}\,\mathrm{d}z,$$

which can be considered the continuous analogue of the function g used in the proof of Lemma 10. For $r > 0$ and $x \leq 1/2$, we note that

$$g(x-r) - g(x) = \frac{2}{\rho}\big(\sqrt{x} - \sqrt{x-r}\big). \tag{15}$$

For better readability, we denote by $X_t := p_t, t \geq 0$, the stochastic process obtained by listing the marginal probabilities of the considered bit over time. Let $Y_t := g(X_t)$ and $\Delta_t := Y_{t+1} - Y_t$. In the remainder of this proof, we assume $X_t \leq 1/2$; analyses for the case $X_t > 1/2$ are symmetrical by switching the sign of Δ_t. We also assume $X_t \geq \rho$ as we are only interested in statements before the first point of time where a border is reached. As mentioned, following the structure of the proof of Lemma 10, we now analyze several moments of Δ_t, with the final aim of establishing the Lyapunov condition in Lemma 11.

Analysis of expected change of potential We claim for all $t \geq 0$ where rw-steps occur (hence, formally we enter the conditional probability space on R_t, the event that an rw-step occurs at time t) that

$$0 \leq \mathrm{E}(\Delta_t \mid X_t; R_t) \leq \frac{3\rho}{2\sqrt{X_t}} = o(1) \tag{16}$$

Moreover, we claim for the unconditional expected value that

$$\mathrm{E}(\Delta_t \mid X_t) \geq -\frac{\rho}{2\alpha}. \tag{17}$$

For a proof of (16), we exploit the martingale property

$$\mathrm{E}(X_{t+1} \mid X_t; R_t) = (1 - X_t)(X_t - \rho X_t) + X_t(X_t + \rho(1 - X_t)) = X_t.$$

that holds in rw-steps of 2-MMAS$_{\text{ib}}$, where there are two possible successor states different from X_t. Since $g(x)$ is a convex function on $[0, 1/2]$, we have by Jensen's inequality

$$\mathrm{E}(\Delta_t \mid X_t; R_t) = \mathrm{E}(g(X_{t+1}) \mid X_t) - g(X_t) \geq g(\mathrm{E}(X_{t+1} \mid X_t)) - g(X_t) = 0.$$

To bound the expected value from above, we carefully estimate the error introduced by the convexity. Note that

$$g(x - x\rho) - g(x) = \int_{x-x\rho}^{x} \frac{1}{\rho\sqrt{z}}\,\mathrm{d}z \leq \frac{x}{\sqrt{x - x\rho}} \tag{18}$$

since the integrand is non-increasing. Analogously,

$$\frac{1-x}{\sqrt{x + (1-x)\rho}} \leq g(x) - g(x + (1-x)\rho) \leq \frac{1-x}{\sqrt{x}} \tag{19}$$

Inspecting the g-values of two possible successor states of $x := X_t$, we get that

$$\begin{aligned}
&\mathrm{E}(\Delta_t \mid X_t = x; R_t) = \mathrm{E}(g(X_{t+1}) - g(x) \mid X_t = x; R_t) &(20)\\
&\quad\leq (1-x)\frac{x}{\sqrt{x - x\rho}} - x\frac{1-x}{\sqrt{x + (1-x)\rho}}\\
&\quad= (1-x)x\left(\frac{1}{\sqrt{x - x\rho}} - \frac{1}{\sqrt{x + (1-x)\rho}}\right)\\
&\quad= (1-x)x \cdot \frac{\sqrt{x + (1-x)\rho} - \sqrt{x - x\rho}}{\sqrt{x + (1-x)\rho} \cdot \sqrt{x - x\rho}} \leq \frac{(1-x)x\frac{\rho}{2\sqrt{x - x\rho}}}{x - x\rho} \leq \frac{x\rho}{2(x/2)^{3/2}}\\
&\quad\leq \frac{3\rho}{2\sqrt{x}}, &(21)
\end{aligned}$$

where the third-last inequality estimated $1-x \leq 1$ and used that $f(z+\rho)-f(z) \leq \rho f'(z)$ for any concave, differentiable function f and $\rho \geq 0$; here using $f(z)=\sqrt{z}$ and $z=x-\rho$. The penultimate used $\rho \leq 1/2$. Since the final bound is $O(\rho/\sqrt{x}) = o(1)$ due to our assumption on $X_t \geq \rho$, we have proved (16).

We now consider the case that a b-step occurs at time t. We are only interested in bounding $\mathrm{E}(\Delta_t \mid X_t)$ from below now. Given $X_t = x$, we have $X_{t+1} > x$ (which means $\Delta_t < 0$) with probability at most $1-(1-x)^2 = 1-(1-2x+x^2) \leq 2x$. With the remaining probability, $X_{t+1} < x$. Since $X_{t+1} \leq x+\rho$, we get

$$\mathrm{E}\big(\Delta_t \mid X_t = x; \overline{R_t}\big) \geq -2x \int_x^{x+\rho} \frac{1}{\rho\sqrt{z}} \mathrm{d}z \geq -2\sqrt{x}. \tag{22}$$

Now, since by assumption a b-step occurs with probability at most $\rho/(4\alpha)$, the unconditional expected value of Δ_t can be computed using the superposition equality. Combining (16) and (22), we get

$$\mathrm{E}(\Delta_t \mid X_t = x) \geq 0 - \frac{\rho}{4\alpha} 2\sqrt{x} \geq -\frac{\rho}{2\alpha}. \tag{23}$$

since $x \leq 1$, proving (17).

Analysis of variance of change of potential Regarding the variance of Δ_t, we claim that

$$\mathrm{Var}(\Delta_t \mid X_t; R_t) \geq 1/16 \tag{24}$$

and, without the condition of having an rw-step,

$$\mathrm{Var}(\Delta_t \mid X_t = x) \geq \frac{1}{32}. \tag{25}$$

To prove this, we expand the definition of variance to estimate

$$\begin{aligned}\mathrm{Var}(\Delta_t \mid X_t) &\geq \mathrm{E}\Big((\Delta_t - \mathrm{E}(\Delta_t \mid X_t = x))^2 \cdot \mathbb{1}\{\Delta_t \leq 0\} \mid X_t = x\Big)\\ &\geq \mathrm{E}\Big((\Delta_t)^2 \cdot \mathbb{1}\{\Delta_t \leq 0\} \mid X_t = x\Big)\end{aligned}$$

since $\mathrm{E}(\Delta_t \mid X_t) \geq 0$. We note that for $X_t = x$, we have $\mathrm{P}\big[X_{t+1} \geq x\big] = x$. On $X_{t+1} \geq x$, we have $\Delta_t < 0$, which means $\mathrm{P}[\Delta_t < 0] = x$. Now,

$$|\Delta_t| = g(x+(1-x)\rho) - g(x) \geq \frac{1-x}{\sqrt{x+\rho(1-x)}} \geq \frac{1-x}{\sqrt{x+x(1-x)}} \geq \frac{1}{4\sqrt{x}}, \tag{26}$$

where the penultimate inequality used $\rho \leq x$ and the last one $x \leq 1/2$. Plugging this in, we get

$$\mathrm{Var}(\Delta_t \mid X_t = x\ R_t) \geq x \cdot \left(\frac{1}{4\sqrt{x}}\right)^2 \geq \frac{1}{16},$$

which completes the proof of (24).

By the law of total probability, we get for the unconditional variance that

$$\mathrm{Var}(\Delta_t \mid X_t) = \mathrm{Var}(\Delta_t \mid X_t; R_t)\mathrm{P}[R_t] + \mathrm{Var}(\Delta_t \mid X_t; \overline{R_t})(1 - \mathrm{P}[R_t]),$$

Since $\mathrm{P}[R_t] \geq 1/2$, we altogether have for the unconditional variance that

$$\mathrm{Var}(\Delta_t \mid X_t = x) \geq 1/32,$$

as claimed in (25).

Approximating the accumulated change of potential by a Normal distribution The aim is to apply the central limit theorem (Lemma 11) on the sum of the Δ_t. To this end, we will verify the Lyapunov condition for $\delta = 1$ (smaller values could be used but do not give any benefit) and $t = \omega(1/\rho)$ (which, as $t = \alpha(s/\rho)^2$, holds due to our assumptions $1/\alpha = o(\rho^{-1})$ and $|s| = \Omega(1)$). We compute

$$\begin{aligned}
&\mathrm{E}\Big(|\Delta_t - \mathrm{E}(\Delta_t \mid X_t)|^3 \mid X_t\Big)\\
&\quad\leq \mathrm{P}[\Delta_t > 0] \cdot (\Delta_t - \mathrm{E}(\Delta_t \mid X_t))^3 + \mathrm{P}[\Delta_t < 0] \cdot (|\Delta_t| + |\mathrm{E}(\Delta_t \mid X_t)|)^3\\
&\quad\leq (1-x)\left(\frac{x}{\sqrt{x - x\rho}}\right)^3 + x \cdot \left(\frac{1-x}{\sqrt{x}} + \frac{3\rho}{2\sqrt{x}} + \frac{\rho}{2\alpha}\right)^3,
\end{aligned}$$

where we again have used (18) and the upper bound from (19) with respect to the two outcomes of X_{t+1}. Moreover, we have used the bound $\mathrm{E}(\Delta_t \mid X_t) \geq 0$ in the first term and $\mathrm{E}(|\Delta_t| \mid X_t) \leq 3\rho/(2\sqrt{x}) + \rho/(2\alpha)$ in the second term, which is a crude combination of (21) and (17). As $\rho \leq 1/2$ and $\rho \leq x$ as well as $\alpha \geq \rho$, the expected value satisfies

$$\begin{aligned}
&\mathrm{E}\Big(|\Delta_t - \mathrm{E}(\Delta_t \mid X_t)|^3 \mid X_t\Big) \leq \left(\frac{x}{\sqrt{x/2}}\right)^3 + x\left(O\left(\frac{1}{\sqrt{x}} + 3\sqrt{x} + \frac{1}{2}\right)^3\right)\\
&\quad\leq 1 + x\left(O\left(\frac{1}{\sqrt{x}}\right)^3\right) = O(1/\sqrt{x}) = O(1/\sqrt{\rho}),
\end{aligned}$$

where we used $x \leq 1$ and $x \geq \rho$. Using $s_t^2 := \sum_{j=0}^{t-1} \mathrm{Var}(\Delta_j \mid X_j)$ in the notation of Lemma 11 and using that $\mathrm{Var}(\Delta_j \mid X_j) \geq 1/32$ by (25), we get

$$\frac{1}{s_t^3}\sum_{j=0}^{t-1} \mathrm{E}\Big(|\Psi_j - \mathrm{E}\big(\Psi_j\big)|^3 \mid X_j\Big) \leq \frac{182}{t^{1.5}} O(t/\sqrt{\rho}) = O(\sqrt{1/(t\rho)}), \tag{27}$$

which goes to 0 as $t = \omega(1/\rho)$. This establishes the Lyapunov condition. Hence, for the value $t := \alpha(s/\rho)^2$ considered in the lemma, we obtain that $\frac{Y_t - \mathrm{E}(Y_t \mid X_0)}{s_t}$ converges in distribution to the normal distribution $N(0, 1)$.

Estimating the accumulated progress Note that $s_t^2 \geq \alpha(s/\rho)^2/32$ since $\mathrm{Var}(\Delta_t \mid X_t) \geq 1/32$ by (25). Hence, $s_t = \sqrt{\alpha/32}(|s|/\rho)$, recalling that $s < 0$. Moreover, as $x \leq 5/6$ is assumed in this part of the lemma, by combining (21) and (17), we get $\mathrm{E}(\Delta_t \mid X_t) \geq -\rho/(2\alpha) - \rho \cdot (3/2)\sqrt{6/5} \geq -\rho/(2\alpha) - 1.7\rho \geq -2.2\rho/\alpha$ and $\mathrm{E}(Y_t) = Y_0 + \sum_{i=0}^{t-1} \mathrm{E}(\Delta_i \mid X_i) \geq 0 + t(-2.2\rho/\alpha) \geq -2.2s^2/\rho$. Together, this means $\frac{\mathrm{E}(Y_t)}{s_t} \geq -\frac{2.2s^2/\rho}{\sqrt{\alpha/32}(|s|/\rho)} \geq -\sqrt{155/\alpha}|s| \geq -\sqrt{155/\alpha}$ since $|s| \leq 1$ and $\alpha \leq 1$. By the normalization to $N(0, 1)$, we have that

$$\mathrm{P}[Y_t \geq r] = \mathrm{P}\left[\frac{Y_t}{s_t} - \frac{\mathrm{E}(Y_t \mid X_0)}{s_t} \geq \frac{r}{s_t} - \frac{\mathrm{E}(Y_t \mid X_0)}{s_t}\right],$$

hence

$$\mathrm{P}[Y_t \geq r] \geq \left(1 - \Phi(r\rho/(|s|\sqrt{\alpha/32}) + \sqrt{155/\alpha})\right) - O(\sqrt{1/(t\rho)})$$

for any r leading to a positive argument of Φ, where Φ denotes the cumulative distribution function of the standard normal distribution and $O(\sqrt{1/(t\rho)})$ the approximation error derived in (27).

We are interested in the event that $Y_t \geq 2\sqrt{|s|}/\rho$, recalling that $s < 0$ and $X_{t+1} \geq X_t \iff Y_{t+1} \leq Y_t$. We made this choice because the event $Y_t = g(X_t) - g(X_0) \geq 2\sqrt{|s|}/\rho$ implies that $X_t - X_0 \leq s$ by (15).

To compute the probability of the event $Y_t \geq 2\sqrt{|s|}/\rho$, we choose $r = 2\sqrt{|s|}/\rho$ and get $r\rho/(|s|\sqrt{\alpha/32}) + \sqrt{155/\alpha}) \leq 24/\sqrt{|s|\alpha}$. We get

$$\mathrm{P}\left[Y_t \geq 2\sqrt{|s|}/\rho\right] \geq \left(1 - \Phi(24/\sqrt{|s|\alpha})\right) - O(\sqrt{1/(t\rho)}).$$

By Lemma 21,

$$1 - \Phi(24/\sqrt{|s|\alpha}) \geq \left(\frac{1}{24/\sqrt{|s|\alpha}} - \frac{1}{(24/\sqrt{|s|\alpha})^3}\right)\frac{1}{\sqrt{2\pi}} e^{-288/(|s|\alpha)} =: p(\alpha, s),$$

which means that distance s is bridged (in negative direction) before or at time $t = \alpha(s/\rho)^2$ with probability at least $p(\alpha, s) - O(\sqrt{1/(t\rho)}) = p(\alpha, s) - O(\sqrt{\rho}/(|s|\sqrt{\alpha}))$. □

The following lemma shows that a marginal probability of less than ρ is unlikely to be increased again.

Lemma 17 *In the setting of Lemma 16, if* $\min\{p_0, 1 - p_0\} \leq \rho$*, the marginal probability will reach the closer border from* $\{1/n, 1 - 1/n\}$ *in* $O((\log n)/\rho)$ *steps with probability at least* $e^{-2/(1-e)}$*. This even holds if each step is a b-step.*

Proof We consider only the case $X_0 \le \rho$ as the other case is symmetrical. The idea is to consider $O(\log n)$ phases and prove that the X_t-value only decreases throughout all phases with the stated probability. Phase i, where $i \ge 0$, starts at the first time where $X_t \le \rho e^{-i}$. Clearly, as $\rho \le 1$, at the latest in phase $\ln n$ the border $1/n$ has been reached. We note that phase i ends after $1/\rho$ steps if all these these steps decrease the value; here we use that each step decreases by a relative amount of $1-\rho$ and that $(1-\rho)^{1/\rho} \le e^{-1}$.

The probability of decreasing the X_t-value in a step of phase i is at least $(1-\rho e^{-i})^2 \ge 1-2e^{-i}\rho$ even if the step is a b-step. Hence, the probability of all steps of phase i being decreasing is at least $(1-2e^{-i}\rho)^{1/\rho} \ge e^{-2e^{-i}}$. For all phases together, the probability of only having decreasing steps is still at least

$$\prod_{i=0}^{\ln n} e^{-2e^{-i}} \ge e^{-2\sum_{i=0}^{\infty} e^{-i}} = e^{-2/(1-e)}$$

as suggested. □

We have now collected all tools to prove the lower bound for 2-MMAS$_{\text{ib}}$.

Proof of Theorem 9 This follows mostly the same structure as the proof of Theorem 8. Every occurrence of the update strength $1/K$ should be replaced by ρ.

There is a minor change in the analysis of rw-steps. The two applications of Lemma 10 are replaced with Lemma 16, followed by an additional application of Lemma 17. The slightly different constants in the statement of Lemma 10 do not affect the asymptotic bound $\Omega(n^{-\beta})$ obtained. Neither does the additional application of Lemma 17, which gives a constant probability. We do not care about the time $O((\log n)/\rho)$ stated in Lemma 17, since we are only interested in a lower bound on the hitting time.

There is a difference in how b-steps are being handled. While Lemma 10 only considers the accumulated effect of rw-steps (leaving the consideration of b-steps to the proof of Theorem 8), Lemma 16 also includes the effect of b-steps, assuming bounds on the probability of b-steps and on the number of b-steps, respectively. We still have to verify that these assumptions are met.

Lemma 16 requires in its first statement that the probability of a b-step is at most $\rho/(4\alpha)$. Recall that such a step has probability $O(1/\sqrt{n})$. We argue that $\rho/(4\alpha) \ge c/\sqrt{n}$ for any constant $c > 0$ if κ is small enough. To see this, we simply recall that $\alpha = \kappa\sqrt{n}\rho/(3s^2)$ by definition and $|s| = \Omega(1)$.

Finally, the second statement of Lemma 16 restricts the number of b-steps until time $\alpha(s/\rho)^2$ to at most $s/(2\alpha\rho)$. Reusing that $\rho = O(\alpha/(\kappa\sqrt{n}))$, this holds by Chernoff bounds with high probability if κ is a sufficiently small constant. Hence, the application of the lemma is possible. □

6 Conclusions

We have performed a runtime analysis of two probabilistic model-building Genetic Algorithms, namely cGA and 2-MMAS$_{\text{ib}}$, on OneMax. The expected runtime of these algorithms was analyzed in dependency of the so-called update strength $S = 1/K$ and $S = \rho$, respectively, resulting in the upper bound $O(\sqrt{n}/S)$ for $S = O(1/\sqrt{n}\log n)$ and $\Omega(\sqrt{n}/S + n\log n)$. Hence, $S \sim 1/\sqrt{n}\log n$ was identified as the choice for the update strength leading to asymptotically smallest expected runtime $\Theta(n\log n)$.

Our analyses of update strength reveal a general trade-off between the speed of learning and genetic drift. High update strengths imply globally a fast adaptation of the probabilistic model but impact the overall correctness of the model negatively, resulting in increased risk of adapting to samples that are locally incorrect. We think that this constitutes a universal limitation of the algorithms that extends to more general classes of functions. As even on the simple OneMax the update strength should not be bigger than $1/(\sqrt{n}\log n)$, we propose this setting as a general rule of thumb.

Our analyses have developed a quite technical machinery for the analysis of genetic drift. These techniques are not necessarily limited to cGA and 2-MMAS$_{\text{ib}}$ on OneMax. Very recently, they have been used in [19] to analyze the so-called UMDA, which is a more complicated EDA. We also believe that the techniques will lead to improved results for classical Genetic Algorithms such as the simple Genetic Algorithm [27], where currently only quite restricted lower bounds on the runtime are available.

Acknowledgements This research was initiated at Dagstuhl seminar 15211 “Theory of Evolutionary Algorithms” and also benefitted from Dagstuhl seminars 16011 “Evolution and Computing” and 17191 “Theory of Randomized Optimization Heuristics”. The authors thank the organisers and participants of all three seminars. The research leading to these results has received funding from the European Union Seventh Framework Programme (FP7/2007-2013) under Grant Agreement No. 618091 (SAGE) and from the Danish Research Council (DFF-FNU) under Grant 4002-00542. This article is based upon work from COST Action CA15140 ‘Improving Applicability of Nature-Inspired Optimisation by Joining Theory and Practice (ImAppNIO)’ supported by COST (European Cooperation in Science and Technology).

A General Tools

A.1 Drift Theorems

The term *variable drift analysis* was coined by Johannsen [17] to describe a stochastic process on non-negative real values where the expected change towards an absorbing target state 0 can be bounded by a positive and monotone increasing function h. His variable drift theorem was subsequently refined and generalized (see also [28] for a broader class of functions h). The following variant is due to Lehre and Witt [22], who allow variable drift for continuous spaces.

Theorem 18 (Variable drift, upper bound; Theorem 16 in [22]). *Let $(X_t)_{t\in\mathbb{N}_0}$, be a stochastic process over some state space $S \subseteq \{0\}\cup[x_{\min}, x_{\max}]$, adapted to a filtration $(\mathcal{F}_t)_{t\in\mathbb{N}_0}$, where $x_{\min} > 0$. Let $h(x)\colon [x_{\min}, x_{\max}] \to \mathbb{R}^+$ be a monotone increasing function such that $1/h(x)$ is integrable on $[x_{\min}, x_{\max}]$ and $\mathrm{E}(X_t - X_{t+1} \mid \mathcal{F}_t) \geq h(X_t)$ if $X_t \geq x_{\min}$. Then it holds for the first hitting time $T := \min\{t \mid X_t = 0\}$ that*

$$\mathrm{E}(T \mid X_0) \leq \frac{x_{\min}}{h(x_{\min})} + \int_{x_{\min}}^{X_0} \frac{1}{h(x)}\,\mathrm{d}x.$$

The next theorem gives tail bounds on variable drift bounds.

Theorem 19 (Tail bounds for variable drift [21], see also Th. 4 in [22]). *Let $(X_t)_{t\in\mathbb{N}_0}$, be a stochastic process, adapted to a filtration $(\mathcal{F}_t)_{t\in\mathbb{N}_0}$, over some state space $S \subseteq \{0\} \cup [x_{\min}, x_{\max}]$, where $x_{\min} \geq 0$. Let $h\colon [x_{\min}, x_{\max}] \to \mathbb{R}^+$ be a function such that $1/h(x)$ is integrable on $[x_{\min}, x_{\max}]$. Suppose there exist a random variable Z and some $\lambda > 0$ such that $|\int_{X_{t+1}}^{X_t} 1/h(\max\{x, x_{\min}\})\,\mathrm{d}x| \prec Z$ for $X_t \geq x_{\min}$ and $E(e^{\lambda Z}) = D$ for some $D > 0$. Then the following two statements hold for the first hitting time $T := \min\{t \mid X_t = 0\}$.*

(i) *If $\mathrm{E}(X_t - X_{t+1} \mid \mathcal{F}_t; X_t \geq x_{\min}) \geq h(X_t)$ then for any $\delta > 0$, and $\eta := \min\{\lambda, \delta\lambda^2/(D-1-\lambda)\}$ and $t > 0$ it holds that*

$$\mathrm{P}[T > t \mid X_0] \leq \exp\left(\eta\left(\frac{x_{\min}}{h(x_{\min})} + \int_{x_{\min}}^{X_0} \frac{1}{h(x)}\,\mathrm{d}x - (1-\delta)t\right)\right).$$

(ii) *If $\mathrm{E}(X_t - X_{t+1} \mid \mathcal{F}_t; X_t \geq x_{\min}) \leq h(X_t)$ then for any $\delta > 0$, $\eta := \min\{\lambda, \delta\lambda^2/(D-1-\lambda)\}$ and $t > 0$ it holds*

$$\begin{aligned}&\mathrm{P}[T < t \mid X_0]\\ &\quad\leq \exp\left(\eta\left((1+\delta)t - \frac{x_{\min}}{h(x_{\min})} - \int_{x_{\min}}^{X_0} \frac{1}{h(x)}\,\mathrm{d}x\right)\right)\frac{1}{\eta(1+\delta)}.\end{aligned}$$

If state 0 is absorbing then

$$\mathrm{P}[T < t \mid X_0] \leq \exp\left(\eta\left((1+\delta)t - \frac{x_{\min}}{h(x_{\min})} - \int_{x_{\min}}^{X_0} \frac{1}{h(x)}\,\mathrm{d}x\right)\right).$$

Finally, we will need the following theorem concerned with drift away from the target. It is taken from [27].

Theorem 20 (Negative Drift with Scaling (Theorem 2 in [27])). *Let $(X_t)_{t\in\mathbb{N}_0}$, be a stochastic process, adapted to a filtration $(\mathcal{F}_t)_{t\in\mathbb{N}_0}$, over some state space $S \subseteq \mathbb{R}_0^+$. Suppose there exist an interval $[a, b] \subseteq \mathbb{R}$ and, possibly depending on $\ell := b - a$, a drift bound $\varepsilon := \varepsilon(\ell) > 0$ as well as a scaling factor $r := r(\ell)$ such that for all $t \geq 0$ the following three conditions hold:*

1. $\mathrm{E}(X_{t+1} - X_t \mid \mathcal{F}_t\,;\, a < X_t < b) \geq \varepsilon$,

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

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

A.2 Bounds on the Cumulative Distribution Function of the Standard Normal Distribution

To prove Lemmas 10 and 16, we need the following estimates for $\Phi(x)$. More precise formulas are available (and can be found by searching for bounds on the so-called error function), but are not required for our analysis.

Lemma 21 ([9], p. 175). *For any* $x > 0$

$$\left(\frac{1}{x} - \frac{1}{x^3}\right)\frac{1}{\sqrt{2\pi}}e^{-x^2/2} \leq 1 - \Phi(x) \leq \frac{1}{x}\frac{1}{\sqrt{2\pi}}e^{-x^2/2},$$

and for $x < 0$

$$\left(\frac{-1}{x} - \frac{-1}{x^3}\right)\frac{1}{\sqrt{2\pi}}e^{-x^2/2} \leq \Phi(x) \leq \frac{-1}{x}\frac{1}{\sqrt{2\pi}}e^{-x^2/2}.$$

A.3 A Bound for Poisson Binomial Distributions

Theorem 22 (Adapted from Theorem 2.1 in [1]). *Let* $S_n = X_1 + \cdots + X_n$ *denote a sum of independent Bernoulli trials where* $\mathrm{P}[X_i = 1] = p_i$. *Then for every* $0 \leq j \leq n$

$$\mathrm{P}[S_n = j] \leq \frac{1}{2\sqrt{\sum_{i=1}^{n} p_i(1-p_i)}}.$$

References

1. Baillon, J.-B., Cominetti, R., Vaisman, J.: A sharp uniform bound for the distribution of sums of Bernoulli trials. Comb. Probab. Comput. **25**, 352–361 (2016)
2. Chen, T., Lehre, P.K., Tang, K., Yao, X.: When is an estimation of distribution algorithm better than an evolutionary algorithm? In: Proceedings of the IEEE Congress on Evolutionary Computation. IEEE Press, pp. 1470–1477 (2009)
3. Dang, D., Lehre, P.K.: Simplified runtime analysis of estimation of distribution algorithms. In: Proceedings of the Genetic and Evolutionary Computation Conference, pp. 513–518 (2015)
4. Doerr, B.: Analyzing randomized search heuristics: tools from probability theory. In: Auger, A., Doerr, B. (eds.) Theory of Randomized Search Heuristics. World Scientific, Singapore (2011)
5. Doerr, B., Johannsen, D., Winzen, C.: Drift analysis and linear functions revisited. In: Proceedings of the IEEE Congress on Evolutionary Computation, pp. 1967–1974 (2010)
6. Doerr, C., Lengler, J.: OneMax in black-box models with several restrictions. In: Proceedings of the Genetic and Evolutionary Computation Conference. ACM Press, pp. 1431–1438 (2015)

7. Droste, S.: A rigorous analysis of the compact genetic algorithm for linear functions. Nat. Comput. **5**(3), 257–283 (2006)
8. Droste, S., Jansen, T., Wegener, I.: Upper and lower bounds for randomized search heuristics in black-box optimization. Theory Comput. Syst. **39**, 525–544 (2006)
9. Feller, W.: An Introduction to Probability Theory and Its Applications, vol. 1. Wiley, New York (1968)
10. Feller, W.: An Introduction to Probability Theory and Its Applications, vol. 2. Wiley, New York (1971)
11. Friedrich, T., Kötzing, T., Krejca, M.S.: EDAs cannot be balanced and stable. In: Proceedings of GECCO'16, pp. 1139–1146 (2016)
12. Friedrich, T., Kötzing, T., Krejca, M.S., Sutton, A.M.: The benefit of recombination in noisy evolutionary search. In: Proceedings of the 26th International Symposium on Algorithms and Computation. Springer, pp. 140–150 (2015)
13. Friedrich, T., Ktzing, T., Krejca, M.S., Sutton, A.M.: The compact genetic algorithm is efficient under extreme gaussian noise. IEEE Trans. Evol. Comput. **21**(3), 477–490 (2017)
14. Gleser, L.J.: On the distribution of the number of successes in independent trials. Ann. Probab. **3**(1), 182–188 (1975)
15. Harik, G.R., Lobo, F.G., Goldberg, D.E.: The compact genetic algorithm. IEEE Trans. Evol. Comput. **3**(4), 287–297 (1999)
16. Hauschild, M., Pelikan, M.: An introduction and survey of estimation of distribution algorithms. Swarm Evol. Comput. **1**(3), 111–128 (2011)
17. Johannsen, D.: Random Combinatorial Structures and Randomized Search Heuristics. Ph.D. thesis, Universität des Saarlandes, Saarbrücken, Germany and the Max-Planck-Institut für Informatik (2010)
18. Krejca, M., Witt, C.: Lower bounds on the run time of the univariate marginal distribution algorithm on OneMax. Theor. Comput. Sci. (2018, to appear); preprint at https://doi.org/10.1016/j.tcs.2018.06.004
19. Krejca, M.S., Witt, C.: Lower bounds on the run time of the univariate marginal distribution algorithm on OneMax. In: Proceedings of FOGA 2017. ACM Press, pp. 65–79 (2017)
20. Lehre, P.K., Nguyen, P.T.H.: Improved runtime bounds for the univariate marginal distribution algorithm via anti-concentration. In: Proceedings of GECCO'17. ACM Press, pp. 414–434 (2017)
21. Lehre, P. K., Witt, C.: Concentrated hitting times of randomized search heuristics with variable drift. In: Proceedings of the 25th International Symposium on Algorithms and Computation, vol. 8889 of Lecture Notes in Computer Science. Springer, pp. 686–697 (2014). Extended version at arXiv:1307.2559
22. Lehre, P.K., Witt, C.: General drift analysis with tail bounds. ArXiv e-prints (2017). arXiv:1307.2559
23. Marshall, A.W., Olkin, I., Arnold, B.C.: Inequalities: Theory of Majorization and Its Applications, 2nd edn. Springer, Berlin (2011)
24. Motwani, R., Raghavan, P.: Randomized Algorithms. Cambridge University Press, Cambridge (1995)
25. Neumann, F., Sudholt, D., Witt, C.: Analysis of different MMAS ACO algorithms on unimodal functions and plateaus. Swarm Intell. **3**(1), 35–68 (2009)
26. Neumann, F., Sudholt, D., Witt, C.: A few ants are enough: ACO with iteration-best update. In: Proceedings of the Genetic and Evolutionary Computation Conference, pp. 63–70 (2010)
27. Oliveto, P.S., Witt, C.: Improved time complexity analysis of the simple genetic algorithm. Theor. Comput. Sci. **605**, 21–41 (2015)
28. Rowe, J.E., Sudholt, D.: The choice of the offspring population size in the $(1, \lambda)$ evolutionary algorithm. Theor. Comput. Sci. **545**, 20–38 (2014)
29. Stützle, T., Hoos, H.H.: MAX-MIN ant system. J.Future Gen. Comput. Syst. **16**, 889–914 (2000)
30. Sudholt, D.: A new method for lower bounds on the running time of evolutionary algorithms. IEEE Trans. Evol. Comput. **17**(3), 418–435 (2013)
31. Sudholt, D., Witt, C.: Update strength in EDAs and ACO: how to avoid genetic drift. In: Proceedings of the Genetic and Evolutionary Computation Conference, New York, NY, USA. ACM, pp. 61–68 (2016)
32. Witt, C.: Tight bounds on the optimization time of a randomized search heuristic on linear functions. Comb. Probab. Comput. **22**(2), 294–318 (2013)
33. Witt, C.: Upper bounds on the running time of the univariate marginal distribution algorithm on OneMax. Algorithmica (2018, to appear); preprint at https://doi.org/10.1007/s00453-018-0463-0

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.