Algorithmica (2021) 83:1096–1137
https://doi.org/10.1007/s00453-020-00778-4

The Complex Parameter Landscape of the Compact Genetic Algorithm

Johannes Lengler[1] · **Dirk Sudholt**[2] · **Carsten Witt**[3]

Received: 14 December 2018 / Accepted: 16 October 2020 / Published online: 4 November 2020

Abstract

The compact Genetic Algorithm (cGA) evolves a probability distribution favoring optimal solutions in the underlying search space by repeatedly sampling from the distribution and updating it according to promising samples. We study the intricate dynamics of the cGA on the test function OneMax, and how its performance depends on the hypothetical population size K, which determines how quickly decisions about promising bit values are fixated in the probabilistic model. It is known that the cGA and the Univariate Marginal Distribution Algorithm (UMDA), a related algorithm whose population size is called λ, run in expected time $O(n \log n)$ when the population size is just large enough ($K = \Theta(\sqrt{n} \log n)$ and $\lambda = \Theta(\sqrt{n} \log n)$, respectively) to avoid wrong decisions being fixated. The UMDA also shows the same performance in a very different regime ($\lambda = \Theta(\log n)$, equivalent to $K = \Theta(\log n)$ in the cGA) with much smaller population size, but for very different reasons: many wrong decisions are fixated initially, but then reverted efficiently. If the population size is even smaller ($o(\log n)$), the time is exponential. We show that population sizes in between the two optimal regimes are worse as they yield larger runtimes: we prove a lower bound of $\Omega(K^{1/3}n + n \log n)$ for the cGA on OneMax for $K = O(\sqrt{n}/\log^2 n)$. For $K = \Omega(\log^3 n)$ the runtime increases with growing K before dropping again to $O(K\sqrt{n} + n \log n)$ for $K = \Omega(\sqrt{n} \log n)$. This suggests that the expected runtime for the cGA is a bimodal function in K with two very different optimal regions and worse performance in between.

Keywords Estimation-of-distribution algorithms · Compact genetic algorithm · Evolutionary algorithms · Runtime analysis · Theory

An extended abstract of this article with parts of the results was presented at GECCO '18 [15].

✉ Johannes Lengler
johannes.lengler@inf.ethz.ch

[1] Department of Computer Science, ETH Zürich, Zürich, Switzerland

[2] University of Sheffield, Sheffield S1 4DP, United Kingdom

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

1 Introduction

Estimation-of-distribution algorithms (EDAs) are general metaheuristics for black-box optimisation that represent a more recent alternative to classical approaches like evolutionary algorithms (EAs). EDAs typically do not directly evolve populations of search points but build probabilistic models of promising solutions by repeatedly sampling and selecting points from the underlying search space. Hence, information about the search can be stored in a relatively compact way, which can make EDAs space-efficient.

Recently, there has been significant progress in the theoretical understanding of EDAs, which supports their use as an alternative to evolutionary algorithms. It has been shown that EDAs are robust to noise [6] and that they have at least comparable runtime behaviour to EAs. Different EDAs like the cGA [22], EDA-like ant colony optimisers (ACO) [17, 22], and the UMDA [2, 12, 14, 24] have been investigated from this perspective.

In this paper, we pick up recent research about the runtime behaviour of the compact Genetic Algorithm (cGA) [9]. The behaviour on the theoretical benchmark function OneMax $(x) := \sum_{i=1}^{n} x_i$ is of particular interest since this function tests basic hill-climbing properties and serves as a basis for the analysis on more complicated functions. Already early analyses of GAs [8] similar to the cGA indicate for OneMax that a population size of $\Omega(\sqrt{n})$ is necessary to prevent premature convergence of the system; together with convergence time analyses of such systems [16] this suggests a runtime that grows not much slower than linear in this case ($O(n \log n)$ for ideal parameter settings). These analyses rely on simplified models of GAs, and they yield good predictions for the behaviour of cGA for some regimes, but behave very differently from the cGA in other regimes. In particular, the simplified models do not resemble the cGA in the regime of medium population sizes that we consider in this paper, and so the performance of the cGA in this regime remained unknown. See also the survey [13] for further results from the theory of EDAs in the last 25 years.

Droste [3] was the first to prove rigorously that the cGA is efficient on OneMax by providing a bound of $O(n^{1+\varepsilon})$ on the runtime. Recently, this bound was refined to $O(n \log n)$ by Sudholt and Witt [21, 22]. However, this bound only applies to a very specific setting of the hypothetical population size K, which is an algorithm-specific parameter of the cGA. Parameters equivalent to K exist in other EDAs, including the UMDA mentioned above.

The choice of the parameter K is crucial for EDAs. It governs the speed at which the probabilistic model is adjusted towards the structure of recently sampled good solutions; more precisely, at hypothetical population size K the algorithm makes steps of size $1/K$. If this step size is too large, the adjustment is too greedy, it is too likely to adapt to incorrect parts of sampled solutions and the system behaves chaotically. If it is too small, adaptation takes very long. However, the dependency of the runtime of the cGA and the UMDA on the population size is very subtle[1]. For both

[1] Unfortunately, our understanding of these algorithms is somewhat fragmented, since some results are proven only for the cGA and some are proven only for the UMDA. However, despite their different appearances, the cGA and the UMDA have been shown to be closely related, and where results for both

the cGA and the UMDA, it is possible to pick some small step size that leads to optimal performance where with high probability all decisions are made correctly, but still as fast as possible. For the UMDA it was shown that there is another, much bigger step size (corresponding to smaller population size) that allows incorrect decisions to be reflected in the probabilistic model for a while, but this is compensated by faster updates.

More concretely, the results from [22] show that for $K \geq c\sqrt{n}\log n$, where c is an appropriate constant, the cGA and the UMDA (with K being replaced by the corresponding parameter λ) optimise OneMax efficiently since for all marginal probabilities of the model, the so-called *frequencies*, the probabilities of sampling a one increase smoothly towards their optimal value because of the small step size $1/K$. The same holds for the UMDA, leading to runtime bounds $O(\lambda n)$ and $O(\lambda\sqrt{n})$, respectively [2, 24]—where for some parameter ranges the results rely on the additional assumption $\lambda = (1 + \Theta(1))\mu$. In these regimes the dynamics of the algorithm can also be well described by gambler's ruin dynamics [8, 9]. At $K = c\sqrt{n}\log n$ (resp. $\lambda = c\sqrt{n}\log n$) both algorithms optimise OneMax in expected time $O(n \log n)$. For smaller step sizes (larger K), at least for the cGA it is known that the runtime increases as $\Omega(K\sqrt{n})$ [22].

On the other hand, it has been independently shown in [2, 14] and [23, 24] that the UMDA achieves the same runtime $O(n \log n)$ for $\lambda = c' \log n$ for a suitable constant c'. The analysis of these very large step sizes indicates that the search dynamics proceed very differently from the dynamics at small step sizes. Namely, for many frequencies the model first learns incorrectly that the optimal value is 0 and then efficiently corrects this decision. The results in [2] and [24] show a general runtime bound of $O(\lambda n)$ for all $\lambda \geq c' \log n$ and $\lambda = o(\sqrt{n}\log n)$ (if the additional assumption $\lambda = \Theta(\mu)$ is made for $\lambda = \Omega(\sqrt{\mu})$. We call this regime the *medium step size* regime, and it is separated from other regimes by two phase transitions: one for small step sizes, corresponding to $K > c\sqrt{n}\log n$ as discussed above, and one for even larger step sizes, corresponding to $K = o(\log n)$, where the system behaves so chaotically that correct decisions are regularly forgotten and the expected runtime on OneMax becomes exponential[2].

We also know that the runtime of the cGA is $\Omega(n \log n)$ for all K [22]. However, it remained an open question whether the runtime is $\Theta(n \log n)$ throughout the whole medium step size regime, or whether the runtime increases with K as suggested by the upper bound $O(\lambda n)$ for the UMDA.

Here we show that the runtime of the cGA does indeed increase, where we formally define runtime as the number of function evaluations until the optimum is sampled for the first time. To simplify the presentation, we assume throughout the paper

Footnote 1 (continued)
algorithms exist, they coincide. Thus we take results for the UMDA as strong indication for analogous behaviour of the cGA, and vice versa.

[2] We define the term "exponential" as $2^{n^{\Omega(1)}}$. This second phase transition has been made explicit in [17] with respect to an ACO algorithm that in fact represents a simple EDA, similar to the cGA. We present a rigorous and slightly stronger statement as part of Theorem 1.

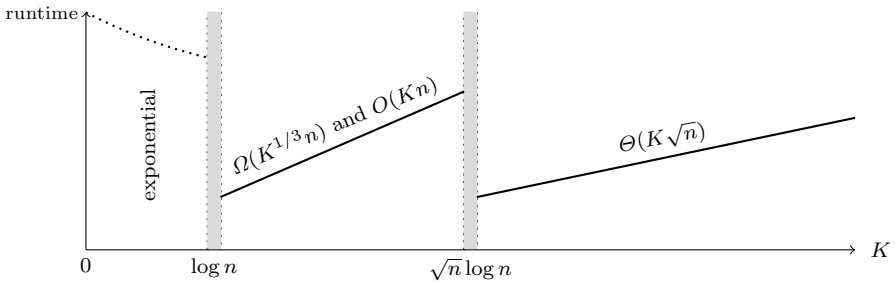

Fig. 1 The runtime landscape of the cGA on OneMax (simplified)

that K is in the set $\mathcal{K} := \{i(1/2 - 1/n) \mid i \in \mathbb{N}\}$ so that the state space of frequencies is restricted to $p_{i,t} \in \{1/n, 1/n + 1/K, \ldots, 1/2, \ldots, 1 - 1/n - 1/K, 1 - 1/n\}$. Then our main result is as follows.

Theorem 1 *Let $K \in \mathcal{K}$.*

If $K \leq 0.3 \log n$ then the runtime of the cGA on OneMax is exponential with overwhelming probability[3] *and in expectation.*

If $K = O(n^{1/2}/(\log(n) \log\log n))$ then the runtime is $\Omega(K^{1/3}n + n \log n)$ with probability $1 - o(1)$ and in expectation.

If $K = O(n^{1/2}/(\log(n) \log\log n))$ and $K = \Omega(\log^3 n)$ then for a suitable constant $\xi < 1$, even the time to create a solution with fitness at least ξn is $\Omega(K^{1/3}n)$ with probability $1 - o(1)$ and in expectation.

This result suggests that the runtime and the underlying search dynamics depend in an astonishingly complex way on the step size: as long as the step size is in the large regime ($K \leq 0.3 \log n$), the expected runtime is exponential. Assuming that the upper bound for the UMDA also holds for the cGA, it then decreases to $O(n \log n)$ at the point where the medium regime is entered. Then the runtime grows with K in the medium regime, where it grows up to $\Omega(n^{7/6}/\log n)$. Before entering the small step size regime ($K = c\sqrt{n} \log n$) the runtime drops again to $O(n \log n)$ [22]. For even smaller step sizes (larger K) the runtime increases again [22]. See Fig. 1 for a simplified illustration of Theorem 1, highlighting the different runtime regimes studied. Experiments conducted for different values of n and K in Sect. 6 confirm that the runtime indeed shows this complex bimodal behaviour.

In addition, the last statement in Theorem 1 shows that even finding a solution within a linear Hamming distance to the optimum takes time $\Omega(K^{1/3}n)$. This is remarkable as many other lower bounds, like the general $\Omega(n \log n)$ bound [22] rely on the fact that optimising the final few incorrect frequencies takes the claimed time (cf. the coupon collector's theorem).

[3] A probability p is called overwhelming if $1/(1 - p)$ is exponential.

The proof of our main theorem is technically demanding but insightful: we obtain insights into the probabilistic process governing the cGA through careful drift analysis. In very rough terms, we analyse the drift of a potential function that measures the distance of the current sampling distribution to the optimal distribution. However, the drift depends on the sampling variance, which is a random variable as well. This leads to a complex feedback system between sampling variance and drift of potential function that tends to self-balance. We are confident that the approach and the tools used here yield insights that will prove useful for analysing other stochastic processes where the drift is changing over time.

This paper is structured as follows. Section 2 defines the cGA and presents fundamental properties of its search dynamics. Section 3 elaborates on the intriguing search dynamics of the cGA in the medium parameter range, including a proof of the fact that many probabilities in the model initially are learnt incorrectly. Section 4 is the heart of our analysis and presents the so-called Stabilisation Lemma, proving that the sampling variance and, thereby, the drift of the potential approach a steady state during the optimisation. It starts with a general road map for the proof. Section 5 puts the whole machinery together to prove the main result. Finally, Sect. 6 contains experiments showing the average runtime across the whole parameter range for K.

2 The Compact Genetic Algorithm and Its Search Dynamics

The cGA, defined in Algorithm 1, uses marginal probabilities (which, as mentioned above, are also known as *frequencies*) $p_{i,t}$ that correspond to the probability of setting bit i to 1 in iteration t. In each iteration two solutions x and y are being created independently using the sampling distribution $p_{1,t}, \dots, p_{n,t}$. Then the fitter offspring amongst x and y is determined, and the frequencies are adjusted by a step size of $\pm 1/K$ in the direction of the better offspring for bits where both offspring differ. Here K determines the strength of the update of the probabilistic model.

The frequencies are always restricted to the interval $[1/n, 1 - 1/n]$ to avoid fixation at 0 or 1. This ensures that there is always a positive probability of reaching a global optimum. Throughout the paper, we refer to $1/n$ and $1 - 1/n$ as (lower and upper) borders. We call frequencies *off-border* if they do not take one of the two border values, i.e., they are not in $\{1/n, 1 - 1/n\}$.

Algorithm 1: Compact Genetic Algorithm (cGA)

$t \leftarrow 0$ and $p_{1,t} \leftarrow p_{2,t} \leftarrow \cdots \leftarrow p_{n,t} \leftarrow 1/2$
while *termination criterion not met* **do**
  for $i \in \{1, \ldots, n\}$ **do**
    $x_i \leftarrow 1$ with prob. $p_{i,t}$ and $x_i \leftarrow 0$ otherwise
    $y_i \leftarrow 1$ with prob. $p_{i,t}$ and $y_i \leftarrow 0$ otherwise
  if $f(x) < f(y)$ **then** swap x and y;
  for $i \in \{1, \ldots, n\}$ **do**
    if $x_i > y_i$ **then** $p'_{i,t+1} \leftarrow p_{i,t} + 1/K$;
    if $x_i < y_i$ **then** $p'_{i,t+1} \leftarrow p_{i,t} - 1/K$;
    if $x_i = y_i$ **then** $p'_{i,t+1} \leftarrow p_{i,t}$;
    $p_{i,t+1} \leftarrow \min\{\max\{1/n, p'_{i,t+1}\}, 1 - 1/n\}$
  $t \leftarrow t + 1$

Overall, we are interested in the cGA 's number of *function evaluations* until the optimum is sampled; this number is typically called *runtime* or *optimisation time*. Note that the runtime is twice the number of iterations until the optimum is sampled.

The behaviour of the cGA is governed by $V_t := \sum_{i=1}^{n} p_{i,t}(1 - p_{i,t})$, the sampling variance at time t. We know from previous work [17, 22] that V_t plays a crucial role in the drift of the frequencies. The following lemma makes this precise by stating transition probabilities and showing that the expected drift towards higher $p_{i,t}$ values is proportional to $1/\sqrt{V_t}$. Recall that all results in this paper tacitly assume $K \in \mathcal{K}$.

Lemma 2 *Consider the cGA on* OneMax. *Then* $p_{i,t+1} = \min\{\max\{1/n, p'_{i,t+1}\}, 1 - 1/n\}$ *where*

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

This implies

$$\mathrm{E}[p_{i,t+1} - p_{i,t} \mid p_{i,t}] = \Theta(1) \cdot \frac{p_{i,t}(1 - p_{i,t})}{K\sqrt{V_t}}$$

where the lower bound requires $p_{i,t} < 1 - 1/n$ *and the upper bound requires* $p_{i,t} > 1/n$.

Proof Note that $p'_{i,t+1} \neq p_{i,t}$ only if the offspring are sampled differently on bit i, which happens with probability $2p_{i,t}(1 - p_{i,t})$. This implies $p_{i,t+1} = p_{i,t}$ with probability $1 - 2p_{i,t}(1 - p_{i,t})$. We only need to bound the probability for $p'_{i,t+1} = p_{i,t} + 1/K$ as it implies the symmetric bound on the probability for $p'_{i,t+1} = p_{i,t} - 1/K$.

Consider the fitness difference $D_{i,t} = \sum_{j \neq i}(x_j - y_j)$ on all other bits. If $|D_{i,t}| \geq 2$ then bit i does not affect the decision whether to update with respect to x or y. Thus we have a conditional probability $\Pr\left(p'_{i,t+1} = p_{i,t} + 1/K \mid |D_{i,t}| \geq 2\right) = p_{i,t}(1 - p_{i,t})$

as $p_{i,t}$ increases if and only if bit i is set to 1 in the fitter individual and to 0 in the other. Such steps are called *random walk steps (rw-steps)* in [22].

If $D_{i,t} = 0$ then there is a higher probability for increasing $p_{i,t}$. In that case, and if $x_i \neq y_i$, then bit i does determine the decision whether to update with respect to x or y: the offspring with a bit value of 1 will be chosen for the update, leading to a conditional probability of $\Pr\left(p'_{i,t+1} = p_{i,t} + 1/K \mid D_{i,t} = 0\right) = 2p_{i,t}(1 - p_{i,t})$. In this scenario, selection between x and y yields a bias towards increasing $p_{i,t}$. Such steps are called *biased steps (b-steps)* in [22].

In [17, proof of Lemma 1] it was shown that

$$\begin{aligned}\Pr\big(D_{i,t} = 0\big) &\geq \frac{1}{11\sqrt{\sum_{j\neq i} p_{j,t}(1-p_{j,t})}} \\ &\geq \frac{1}{11\sqrt{\sum_{j=1}^{n} p_{j,t}(1-p_{j,t})}} = \frac{1}{11\sqrt{V_t}}.\end{aligned}$$

In order to bound $\Pr\big(D_t = 0\big)$ from above, imagine that the cGA first creates all bits x_j for $j \neq i$, such that these bits are given and y_j for $j \neq i$ are random variables. Then $D_t = 0$ is equivalent to $\sum_{j\neq i} y_j = \sum_{j\neq i} x_j$. Note that $\sum_{j\neq i} y_j$ is a Poisson-Binomial distribution on $n-1$ bits. Using the general probability bound for such distributions from [1] (see Theorem 3.2 in [14]), for any fixed k,

$$\begin{aligned}\Pr\left(\sum_{j\neq i} y_j = k\right) &\leq O(1) \cdot \left(\sum_{j\neq i} p_{j,t}(1-p_{j,t})\right)^{-1/2} \\ &\leq O(1) \cdot \frac{1}{2}\left(\sum_{j=1}^{n} p_{j,t}(1-p_{j,t})\right)^{-1/2} = O(1/(\sqrt{V_t})),\end{aligned}$$

the second inequality following from $\sum_{j\neq i} p_{j,t}(1-p_{j,t}) \geq (n-1)\cdot 1/n \cdot (1-1/n) \geq 1/4$ and $p_{i,t}(1-p_{i,t}) \leq 1/2 \cdot 1/2 = 1/4$.

The remaining cases $D_{i,t} = -1$ and $D_{i,t} = +1$ fall in one of the above cases and can be handled in the same way. Together, $\Pr\left(p'_{i,t+1} = p_{i,t} + 1/K\right) = p_{i,t}(1-p_{i,t}) \cdot (1 + \Theta(1/\sqrt{V_t}))$, which proves the claimed probability bounds.

The statement on the expectation follows easily from the probability bounds and verifying the statement for boundary values, noting that $K \in \mathcal{K}$. □

Remark 1 A statement very similar to Lemma 2 also holds for the UMDA on OneMax, even though the latter algorithm uses a sampling and update procedure that is rather different from the cGA as it can in principle lead to large changes in a single iteration. However, the expected change of a frequency follows the same principle as for the cGA. Roughly speaking, the results from [12] and [23] together show that the UMDA 's frequencies evolve according to

$$\mathrm{E}[p_{i,t+1} - p_{i,t} \mid p_{i,t}] = \Theta(1) \cdot p_{i,t}(1 - p_{i,t})/\sqrt{V_t}$$

Note that this drift is by a factor of K larger than in the cGA. However, since each iteration of the UMDA entails λ fitness evaluations, where λ is a parameter that can be compared to K in the cGA, the overall runtime is the same for both algorithms.

The progress of the cGA can be measured by considering a natural potential function: the function $\varphi_t := \sum_{i=1}^{n}(1 - p_{i,t})$ measures the distance to the "ideal" distribution where all $p_{i,t}$ are 1. While the drift on individual frequencies is inversely proportional to the root of the sampling variance, $\sqrt{V_t}$, the following lemma shows that the drift of the potential is proportional to $\sqrt{V_t}$. It also provides a tail bound for the change of the potential.

Lemma 3 *Let* $\varphi_t := \sum_{i=1}^{n}(1 - p_{i,t})$, *then* $\mathrm{E}[\varphi_t - \varphi_{t+1} \mid \varphi_t] = O(\sqrt{V_t}/K)$. *Moreover, for all* t *such that* $V_t = O(K^2)$,

$$\Pr\left(|\varphi_t - \varphi_{t+1}| \geq \sqrt{V_t}\log n \mid \varphi_t\right) \leq n^{-\Omega(K \log\log n)}.$$

Proof Note that $\mathrm{E}[p_{i,t+1} - p_{i,t} \mid p_{i,t}] = O(1) \cdot \frac{p_{i,t}(1-p_{i,t})}{K\sqrt{V_t}}$ by Lemma 2 if $p_{i,t} > 1/n$. Otherwise, it is bounded by $O(1/(nK))$ as at least one offspring needs to be sampled at 1. In both cases, the drift for one frequency is bounded by $O(1) \cdot \left(\frac{1}{nK} + \frac{p_{i,t}(1-p_{i,t})}{K\sqrt{V_t}}\right)$, hence

$$\begin{aligned}\mathrm{E}[\varphi_t - \varphi_{t+1} \mid \varphi_t] &\leq \sum_{i=1}^{n} O(1) \cdot \left(\frac{1}{nK} + \frac{p_{i,t}(1-p_{i,t})}{K\sqrt{V_t}}\right) \\ &= O\left(\frac{1}{K} + \frac{V_t}{K\sqrt{V_t}}\right) = O\left(\frac{\sqrt{V_t}+1}{K}\right) = O\left(\frac{\sqrt{V_t}}{K}\right)\end{aligned}$$

where the last step used $V_t \geq 1 - 1/n$, hence $\sqrt{V_t} + 1 = O(\sqrt{V_t})$.

To bound the step size in one iteration, note that φ_t can only be changed by frequencies that are sampled differently in both offspring, and in that case φ_t is changed by at most $1/K$. Hence $|\varphi_t - \varphi_{t+1}|$ is stochastically dominated by the sum of indicator variables for each frequency i that take on value $1/K$ with probability $2p_{i,t}(1 - p_{i,t})$ and 0 otherwise. These variables are independent (not identically distributed) and their sum's expectation is $2V_t/K$.

We estimate the contribution of off-border frequencies to $|\varphi_t - \varphi_{t+1}|$ separately from the contribution of frequencies at a border, showing that both quantities are at most $(\sqrt{V_t}\log n)/2$ with the claimed probability. Let m denote the number of off-border frequencies at time t. Frequencies at a border only change with probability $2(1 - 1/n)/n$. The expected number of frequencies that change is $2(1 - 1/n)(n - m)/n \leq 2(1 - 1/n)$ and the probability that at least $(1 - 1/n)K/2 \cdot \log n$ frequencies at borders change is at most $(K \log n)^{-\Omega(K \log n)} \leq n^{-\Omega(K \log\log n)}$, which follows from the well-known Chernoff

bound $\Pr(X \geq (1+\delta)\mathrm{E}[X]) \leq \left(\frac{e^\delta}{(1+\delta)^{1+\delta}}\right)^{\mathrm{E}[X]}$ with $1+\delta := K(\log n)/4$. As $V_t \geq 1-1/n$ and $\sqrt{V_t} \geq 1-1/n$, with overwhelming probability fewer than $(1-1/n)K/2 \cdot \log n \leq (\sqrt{V_t}\log n)K/2$ frequencies at borders change. Since every change alters φ_t by $\pm 1/K$, the total contribution of frequencies at borders to $|\varphi_t - \varphi_{t+1}|$ is at most $(\sqrt{V_t}\log n)/2$.

For the m off-border frequencies we note that every such frequency contributes at least $1/K \cdot (1-1/K)$ to V_t, hence $V_t = \Omega(m/K)$. Recall from above that the sum of all variables leads to an expectation of $2V_t/K$, hence the expectation of just the off-border frequencies is at most $2V_t/K$. Using the assumption $V_t = O(K^2)$, which is equivalent to $\sqrt{V_t} = O(K)$, we have $2V_t/K = 2\sqrt{V_t} \cdot \sqrt{V_t}/K = O(\sqrt{V_t})$. We apply Chernoff-Hoeffding bounds (Lemma 23) with $\sum_i b_i^2 = m/K^2$ and a deviation from the expectation of $(\sqrt{V_t}\log n)/2 - O(\sqrt{V_t}) = \Omega(\sqrt{V_t}\log n)$. Denoting by X the contribution of frequencies that are off-border at time t to $|\varphi_t - \varphi_{t+1}|$, we have

$$\Pr\left(X \geq (\sqrt{V_t}\log n)/2 \mid \varphi_t\right) \leq 2e^{-\Omega((K^2 V_t \log^2 n)/m)} \leq 2e^{-\Omega(K\log^2 n)} = n^{-\Omega(K\log n)}.$$

Taking the union bound over all failure probabilities completes the proof. □

3 Dynamics with Medium Step Sizes

As described in the introduction, the cGA in the medium step size regime, corresponding to $K = o(\sqrt{n}\log n)$ and $K = \Omega(\log n)$, will behave less stable than in the small step size regime. In particular, many frequencies will be reinforced in the wrong way and will walk to the lower border before the optimum is found, resulting in an expected runtime of $\Omega(n\log n)$ [22]. With respect to the UMDA it is known [23] that such wrong decisions can be "unlearned" efficiently, more precisely the potential φ_t improves by an expected value of $\Omega(1)$ per iteration. This implies the upper bound $O(\lambda n)$ in the medium regime, which becomes minimal for $\lambda = \Theta(\log n)$. Even though formally we have no upper bounds on the runtime of the cGA on OneMax in the medium regime, we conjecture strongly that it exhibits the same behaviour as the UMDA and has expected runtime $O(Kn)$. We finally recall the first statement of Theorem 1: for extremely large step sizes, $K \leq 0.3\log n$, the runtime becomes exponential. This statement will be shown in Sect. 5; the main reason for the exponential time is that the system contains too few states to build a reliable probabilistic model.

The following lemma shows that a linear number of frequencies tends to reach the upper and lower borders in the initial phase of a run.

Lemma 4 *Consider the cGA with $K \leq \sqrt{n}$. Then with probability $1 - e^{-\Omega(\sqrt{n})}$ at least $\Omega(n)$ frequencies reach the lower border and at least $\Omega(n)$ frequencies reach the upper border within the first $O(K^2)$ iterations.*

A proof of Lemma 4 is given in the "appendix" as it repeats many arguments from the proof of Theorem 8 in [22], where calculations can be simplified because of the assumption on K.

Frequencies at any border tend to remain there for a long time. The following statement shows that in an epoch of length $r = o(n)$ the fraction of frequencies at a border only changes slightly.

Definition 1 Let $\gamma(t)$ denote the fraction of frequencies at the lower border at time t.

Lemma 5 *Consider the cGA with $K \leq \sqrt{n}$. For every $r = o(n)$ and every iteration $t \leq t' \leq t + r$ with probability $1 - e^{-\Omega(r)} - e^{-\Omega(\sqrt{n})}$ we have $\gamma(t') \geq \gamma(t) - O(r/n)$. With probability $1 - e^{-\Omega(r)} - e^{-\Omega(\sqrt{n})}$ there is a time $t_0 = O(K^2)$ such that $\gamma_0 := \gamma(t_0) = \Omega(1)$.*

Both statements also hold for the fraction of frequencies at the upper border.

Proof The first statement follows from the fact that a frequency at a border has to sample the opposite value in one offspring to leave its respective border. Taking a union bound over two created search points, the probability for leaving the border is at most $2/n$, hence the expected number of frequencies leaving a border during r steps is at most $2r$. The probability that at least $4r$ frequencies leave the border is $e^{-\Omega(r)}$ by Chernoff bounds. This implies the first inequality.

The statement regarding γ_0 follows from Lemma 4 and the first statement: $\Omega(n)$ frequencies will hit the lower and the upper border, respectively, within the first $t_0 = O(K^2)$ steps with probability $1 - 2^{-\Omega(n)}$ and, for frequencies hitting a border before time t_0 with probability $1 - e^{-\Omega(r)}$ less than $4r = o(n)$ frequencies will leave the border before time t_0. □

We now show that with high probability, every off-border frequency will hit one of the borders after a short number of iterations. The proof of the following lemma uses that the probability of increasing a frequency is always at least the probability of decreasing it. Hence, if every iteration was actually changing the probability, the time bound $O(K^2)$ would follow by standard arguments on the fair random walk on K states. However, the probability of changing the state is only $p_{i,t}(1 - p_{i,t})$ and the additional $\log K$-factor covers that the process has to travel through states with a low probability of movement before hitting a border.

Lemma 6 *Consider the frequency $p_{i,t}$ of a bit i of the cGA with $K = \omega(1)$ on OneMax. Let T be the first time when $p_{i,t} \in \{1/n, 1 - 1/n\}$. Then for every initial value $p_{i,0}$ and all $r \geq 8$, $\mathrm{E}[T \mid p_{i,0}] \leq 4K^2 \ln K$ and $\Pr(T \geq rK^2 \ln K \mid p_{i,0}) \leq 2^{-\lfloor r/8 \rfloor}$.*

Proof We consider the process X_t, $t \geq 0$, on the state space $\{q(0), q(1), \ldots, q(K')\}$ where $q(i) = 1/n + i/K$ and $K' = K(1 - 2/n)$; note that K' is an integer since $K \in \mathcal{K}$. Obviously, T equals the first hitting time of $q(0)$ or $q(K')$ for the X_t-process. To analyze T, we only use that X_t is stochastically at least as large as a fair random

walk with self-loop probability $1 - 2q(i)(1 - q(i))$ at state $q(i)$. More precisely, it holds for $X_t = q(i)$, where $i \in \{1, \dots, K' - 1\}$ that

$$X_{t+1} = \begin{cases} q(i), & \text{with probability } 1 - 2q(i)(1 - q(i)) \\ q(i+1), & \text{with probability } \geq q(i)(1 - q(i)) \\ q(i-1), & \text{with probability } \leq q(i)(1 - q(i)). \end{cases}$$

The aim is to show that state $q(0)$ or $q(K')$ is reached by random fluctuations due to the variance of the process even in the case that the transition probabilities are completely fair in both directions. Since we do not have a lower bound on the probability of going from i to $i - 1$ in the actual process, it may happen that the actual process is unable to hit state $q(0)$ whereas the general class of processes considered here may well be able to hit this state. Therefore, we make state $q(0)$ reflecting by defining $\Pr(X_{t+1} = q(1) \mid X_t = q(0)) := 1$. Then we estimate the first hitting time of state $q(K')$ for the process modified in this way. Since hitting state $q(0)$ is included in the stop event, we can only overestimate T by this modification.

We introduce the potential function g on $\{0, \dots, K'\}$ defined through $g(0) := 0$, $g(1) := 1$ and recursively $g(i+1) - g(i) := g(i) - g(i-1) + \frac{1}{q(i)(1-q(i))}$ for $i \geq 1$, and analyze the process $Y_t := g(K(X_t - 1/n))$ through drift analysis. To this end, we need an upper bound on $g(K')$ and a lower bound on the drift.

By expanding the recursive definition, we note that for $i \geq 1$

$$g(i+1) - g(i) = g(1) - g(0) + \sum_{k=1}^{i} \frac{1}{q(k)(1-q(k))} = g(1) + \sum_{k=1}^{i} \frac{1}{q(k)(1-q(k))}$$

and therefore, representing $g(K') - g(0)$ as a telescoping sum,

$$\begin{aligned} g(K') = g(K') - g(0) &= \sum_{j=0}^{K'-1} g(j+1) - g(j) \\ &= g(1) - g(0) + \sum_{j=1}^{K'-1} \left(g(1) + \sum_{k=1}^{j} \frac{1}{q(k)(1-q(k))} \right) \\ &\leq K' + \sum_{j=1}^{K'/2} \sum_{k=1}^{j} \frac{4K}{j} + \sum_{j=K'/2+1}^{K'-1} \sum_{k=1}^{j} \frac{4K}{K-j} \\ &\leq K' + \frac{K}{2}(4K(\ln(K/2) + 1)) + 4K \sum_{j=K/2+1}^{K-1} (\ln(K-1) + 1) \\ &\leq K + 2K^2(\ln(K/2) + 1) + 2K^2(\ln(K-1) + 1) \leq 4K^2 \ln(K), \end{aligned}$$

where the first inequality estimated $q(i) \geq i/(2K') \geq i/(2K)$ and $1 - q(i) \geq 1/2$ for $i \leq K'/2$ and symmetrically for $i > K'/2$. The last inequality holds if K is larger than some constant.

Using $\Pr(X_{t+1} = q(i+1) \mid X_t = q(i)) \geq \Pr(X_{t+1} = q(i-1) \mid X_t = q(i))$ and $\Pr(X_{t+1} \neq q(i) \mid X_t = q(i)) = 2q(i)(1 - q(i))$, we obtain for $i \geq 1$ that

$$\begin{aligned}
&\mathrm{E}[Y_{t+1} - Y_t \mid X_t = i] \\
&\qquad \geq q(i)(1-q(i))((g(i+1) - g(i)) - (g(i) - g(i-1))) \\
&\qquad = \; q(i)(1-q(i))\frac{1}{q(i)(1-q(i))} = 1
\end{aligned}$$

by definition of g. Moreover, clearly $\mathrm{E}[Y_{t+1} - Y_t \mid X_t = q(0)] = 1$. Altogether, by the additive drift theorem (Theorem 24), $\mathrm{E}[T \mid X_0] \leq 1 \cdot g(K') \leq 4K^2 \ln(K)$ as claimed.

For the tail bound, we note that the upper bound on $\mathrm{E}[T \mid X_0]$ holds for all starting points. Hence, we obtain from Markov's inequality that $T \leq 8K^2 \ln K$ with probability at least 1/2. The probability that the target is not reached within $m \geq 1$ phases of length $8K^2 \log K$ each is bounded by at most 2^{-m}. The claimed tail bound now follows for all $r \geq 8$. □

4 Stabilisation of the Sampling Variance

Now that we have collected the basic properties of the cGA, we can give a detailed road map of the proof. We want to use a drift argument for the potential φ_t (recall $\varphi_t := \sum_{i=1}^{n}(1 - p_{i,t})$). After a short initial phase, most of the frequencies are at the borders, but since a linear fraction is at the lower border we start with $\varphi_t = \Omega(n)$. As we have seen, the drift of φ_t is $O(\sqrt{V_t}/K)$, so the heart of the proof is to study how V_t evolves.

However, the behaviour of V_t is complex. It is determined by the number and position of the frequencies in the off-border region (the other frequencies contribute only negligibly). By Lemma 2, each $p_{i,t}$ performs a random walk with (state-dependent) drift proportional to $1/\sqrt{V_t}$. Therefore, V_t affects itself in a complex feedback loop. For example, if V_t is large, then the drift of each $p_{i,t}$ is weak (not to be confused with the drift of φ_t, which is strong for large V_t). This has two opposing effects. Consider a frequency that leaves the lower border. On the one hand, the frequency has a large probability to be re-absorbed by this border quickly. On the other hand, if it does gain some distance from the lower border then it spends a long time in the off-border region, due to the weak drift. For small V_t and large drift, the situation is reversed. Frequencies that leave the lower border are less likely to be re-absorbed, but also need less time to reach the upper border. Thus the number and position of frequencies in the off-border region depends in a rather complex way on V_t.

To complicate things even more, the feedback loop from V_t to itself has a considerable lag. For example, imagine that V_t suddenly decreases, i.e. the drift of the $p_{i,t}$ increases. Then frequencies close to the lower border are less likely to return to the lower border, and this also affects frequencies which have already left the border earlier. On the other hand, the drift causes frequencies to cross the off-border region more quickly, but this takes time: frequencies that are initially in the off-border region will not jump to a border instantly. Thus the dynamics of V_t play a role. For instance, if a phase of small V_t (large drift of $p_{i,t}$) is followed by a phase of large V_t (small drift of $p_{i,t}$), then in the first phase many frequencies reach the off-border

region, and they all may spend a long time there in the second phase. This combination could not be caused by any static value of V_t.

Although the situation appears hopelessly complex, we overcome these obstacles using the following key idea: *the sampling variance V_t of all frequencies at time t can be estimated accurately by analysing the stochastic behaviour of one frequency i over a period of time.* More specifically, we split the run of the algorithm into epochs of length $K^2\beta(n) = o(n/\log\log n)$, with $\beta(n) = C\log^2 n$ for a sufficiently large constant C, long enough that the value of V_t may take effect on the distribution of the frequencies. We assume that in one such epoch we know bounds $V_{\min} \le V_t \le V_{\max}$, and we show that, by analysing the dynamics of a single frequency, (stronger) bounds $V'_{\min} \le V_t \le V'_{\max}$ hold for the next epoch. The following key lemma makes this precise.

Lemma 7 (Stabilisation Lemma) *Let $r := K^2\beta(n)$ with $C\log^3 n \le K \le n$ and with $\beta(n) = C\log^2 n$, for a sufficiently large constant $C > 0$. Let further $t_1 > 0$, $t_2 := t_1 + r$ and $t_3 := t_2 + r$. Assume $\gamma(t_1) = \Omega(1)$. There are $C', C'' > 0$ such that the following holds for all $V_{\min} \in [0, K^{2/3}/C']$ and $V_{\max} \in [C''K^{4/3}, \infty]$. Assume that $V_{\min} \le V_t \le V_{\max}$ for all $t \in [t_1, t_2]$. Then with probability $1 - q$ we have $V'_{\min} \le V_t \le V'_{\max}$ for all times $t \in [t_2, t_3]$, with the following parameters.*

(a) *If $V_{\min} = 0$, $V_{\max}$ arbitrary, then*

- $V'_{\min} = \Omega(\sqrt{K})$;
- $V'_{\max} = \infty$;
- $q = \exp(-\Omega(\sqrt{K}))$.

(b) *If $V_{\min} = \Omega(\sqrt{K})$, $V_{\max}$ arbitrary, then*

- $V'_{\min} = \Omega(\sqrt{K}V_{\min}^{1/4})$;
- $V'_{\max} = O(K\min\{K, \sqrt{V_{\max}}\}/\sqrt{V_{\min}})$;
- $q = \exp(-\Omega(\min\{\sqrt{V_{\min}}, \sqrt{K}/V_{\min}^{1/4}\}))$.

To understand where the values of $V'_{\min}$ and $V'_{\max}$ come from, we recall that $V_t = \sum_{i=1}^{n} p_{i,t}(1 - p_{i,t})$, and we regard the terms $p_{i,t}(1 - p_{i,t})$ from an orthogonal perspective. For a fixed frequency i that leaves the lower border at some time t_1, we consider the total lifetime contribution of this frequency to all V_t until it hits a border again at some time t_2, so we consider $P_i = \sum_{t=t_1}^{t_2} p_{i,t}(1 - p_{i,t})$. Note that V_t and P_i are conceptually very different quantities, as the first one adds up contributions of all frequencies for a fixed time, while the second quantifies the total contribution of a fixed frequency over its lifetime. Nevertheless, we show in Sect. 4.1 that their expectations are related, $\mathrm{E}[V_t] \approx 2\gamma(t)\mathrm{E}[P_i]$, where $2\gamma(t)$ is the expected number of frequencies that leave the

lower border in each round.[4] Crucially, $\mathrm{E}[P_i]$ is much easier to analyse: we link $\mathrm{E}[P_i]$ to the expected hitting time $\mathrm{E}[T]$ of a rescaled and loop-free version of the random walks that the frequencies perform. In Sect. 4.2 we then derive upper and lower bounds on $\mathrm{E}[T]$ that hold for all random walks with given bounds on the drift, which then lead to upper and lower bounds $V'_{\min} \leq \mathrm{E}[V_t] \leq V'_{\max}$.

To prove Lemma 7, it is not sufficient to know $\mathrm{E}[V_t]$, we also need concentration for V_t. Naturally V_t is a sum of random variables $p_{i,t}(1-p_{i,t})$, so we would like to use the Chernoff bound. Unfortunately, all the random walks of the frequencies are correlated, so the $p_{i,t}$ are not independent. However, we show by an elegant argument in Sect. 4.3 that we may still apply the Chernoff bound. We partition the set of frequencies into m batches, and show that the random walks of the frequencies in each batch do not substantially influence each other. This allows us to show that the contribution of each batch is concentrated with exponentially small error probabilities. The overall proof of Lemma 7 is then by induction. Given that we know bounds $V_{\min}$ and $V_{\max}$ for one epoch, we show by induction over all times t in the next epoch that V_t satisfies even stronger bounds $V'_{\min}$ and $V'_{\max}$.

In Sect. 5 we then apply Lemma 7 iteratively to show that the bounds $V_{\min}$ and $V_{\max}$ become stronger with each new epoch, until we reach $V_{\min} = \Omega(K^{2/3})$ and $V_{\max} = O(K^{4/3})$. At this point the approach reaches its limit, since then the new bounds $V'_{\min}$ and $V'_{\max}$ are no longer sharper than $V_{\min}$ and $V_{\max}$. Still, the argument shows that $V_t = O(K^{4/3})$ from this point onwards, which gives us an upper bound of $O(K^{-1/3})$ on the drift of φ_t and a lower bound of $\Omega(K^{1/3}n)$ on the runtime of the algorithm.

As the proof outline indicates, the key step is to prove Lemma 7, and the rest of the section is devoted to it.

4.1 Connecting V_t to the Lifetime of a Frequency

In this section we will lay the foundation to analyse $\mathrm{E}[V_t]$. We consider the situation of Lemma 7, i.e., we assume that we know bounds $V_{\min} \leq V_t \leq V_{\max}$ that hold for an epoch $[t_1, t_2]$ of length $t_2 - t_1 = r = K^2\beta(n)$. We want to compute $\mathrm{E}[V_t]$ for a fixed $t \in [t_2, t_3]$. Since $V_t = \sum_{i=1}^{n} p_{i,t}(1-p_{i,t})$, we call the term $p_{i,t}(1-p_{i,t})$ the *contribution* of the i-th frequency to V_t. The main result of this section (and one of the main insights of the paper) is that the contribution of the off-border frequency can be described by $\mathrm{E}[V_t] = \Theta(\gamma(t)\mathrm{E}[T])$, where T is the lifetime of a random variable that performs a rescaled and loop-free version of the random walk that each $p_{i,t}$ performs.

First we introduce the rescaled and loop-free random walk. It can be described as the random walk that $p_{i,t}$ performs for an individual frequency if we ignore self-loops, i.e., if we assume that in each step $p_{i,t}$ either increases or decreases

[4] The actual statement is a bit more subtle and involves lower and upper bounds on P_i, see Lemma 9.

by $1/K$. Moreover, it will be convenient to scale the random walk by roughly a factor of K so that the borders are 0 and K instead of $1/n$ and $1-1/n$. The exact scaling is given by the formula $X_{i,t} = (p_{i,t} - 1/n)/(K - 2/n)$. Formally, assume that X_t is a random walk on $\{0, \dots, K\}$ where the following bounds hold whenever $X_t \in \{1, \dots, K-1\}$.

$$X_{t+1} = \begin{cases} X_t + 1, \text{ with probability } \frac{1}{2} + d(t), \\ X_t - 1, \text{ with probability } \frac{1}{2} - d(t), \end{cases} \tag{2}$$
$$\text{where } d(t) = \Omega\big(1/\sqrt{V_{\max}}\big) \text{ and } d(t) = O\big(1/\sqrt{V_{\min}}\big).$$

Note that by Lemma 2, if we condition on $p_{i,t+1} \neq p_{i,t}$ then $p_{i,t}$ follows a random walk that increases with probability $1/2 + \Theta(1/\sqrt{V_t})$. Hence, if $V_{\min} \leq V_t \leq V_{\max}$ then this loop-free random walk of $p_{i,t}$ follows the description in (2) after scaling. Therefore, we will refer to the random walk defined by (2) as the *loop-free random walk* of a frequency. We remark that it is slight abuse of terminology to speak of *the* loop-free random walk, since (2) actually describes a class of random walks. Formally, when we prove upper and lower bounds on the hitting time of "the" loop-free random walk, we prove bounds on the hitting time of any random walk that follows (2).

To link $\mathrm{E}[V_t]$ and $\mathrm{E}[T]$, we need one more seemingly unrelated concept. Consider a frequency i that leaves the lower border at some time t_0, i.e., $p_{i,t_0-1} = 1/n$ and $p_{i,t_0} = 1/n + 1/K$, and let $t' > 0$ be the first point in time when $p_{i,t}$ hits a border, so $p_{i,t'} = 1/n$ or $p_{i,t'} = 1 - 1/n$. Then we call

$$P_i := \sum_{t=t_0}^{t'-1} p_{i,t}(1 - p_{i,t}), \quad \text{where } p_{i,t_0} = 1/n + 1/K \tag{3}$$

the *lifetime contribution* of the i-th frequency. Analogously, we denote by P'_i the lifetime contribution if frequency i leaves the upper border,

$$P'_i := \sum_{t=t_0}^{t'-1} p_{i,t}(1 - p_{i,t}), \quad \text{where } p_{i,t_0} = 1 - 1/n - 1/K. \tag{4}$$

Note that V_t and P_i are both sums over terms of the form $p_{i,t}(1 - p_{i,t})$. But while V_t sums over all i for fixed t, P_i sums over some values of t for a fixed i. Nevertheless, as announced in the proof outline, we will show that the expectations $\mathrm{E}[V_t]$ and $\mathrm{E}[P_i]$ are closely related, and this will be the link between $\mathrm{E}[V_t]$ and $\mathrm{E}[T]$. More precisely, we show the following lemma.

Lemma 8 *Consider the situation of Lemma 7. Let $t \in [t_2, t_3]$, and assume $V_{\min} \leq V_{t'} \leq V_{\max}$ for all $t' \in [t_1, t-1]$. Let S_{low} be the set of all frequencies i with $p_{i,t} \notin \{1/n, 1-1/n\}$, and such that their last visit of a border was in $[t_1, t]$, and it was at the lower border. Formally, we require that $t_0 := \max\{\tau \in [t_1, t] \mid p_{i,\tau} \in \{1/n, 1-1/n\}\}$ exists and that $p_{i,t_0} = 1/n$. Let S_{upp} be the analogous set, where the last visit was at the upper border. Then*

(a) $\mathrm{E}[\sum_{i\in S_{\text{low}}} p_{i,t}(1-p_{i,t})] = \Theta(\mathrm{E}[P_i])$.
(b) $\mathrm{E}[\sum_{i\in S_{\text{upp}}} p_{i,t}(1-p_{i,t})] = \Theta(\mathrm{E}[P'_i])$.
(c) $\mathrm{E}[\sum_{i\in\{1,\dots,n\}\setminus(S_{\text{low}}\cup S_{\text{upp}})} p_{i,t}(1-p_{i,t})] = O(1)$.

Proof *(a)* Recall that we assume $\gamma(t_1) = \Omega(1)$. With high probability $\gamma(t)$ is slowly changing by Lemma 5 and $\gamma(t) \le 1$ always holds trivially, so with high probability there is a constant $c > 0$ such that $c \le \gamma(t) \le 1$ for all $t \in [t_1, t_3]$. More precisely, since $t_3 - t_1 = \omega(\log n)$, the error probability in Lemma 5 is $n^{-\omega(1)}$. Since $p_{i,t}$ is polynomially lower bounded in n, such small error probabilities are negligible. In particular, we may assume that for every $t' \in [t_1, t_3]$, the expected number of frequencies $s(t)$ which leave the lower border at time t is $\mathrm{E}[s(t)] = \gamma(t)n \cdot \frac{2}{n}(1-\frac{1}{n}) = (2-o(1))\gamma(t) = \Theta(1)$.

Consider a frequency that leaves the lower border at time 0, and let $\rho_t := p_{i,t}(1-p_{i,t})$ if i has not hit a border in the interval $[1, t]$, and $\rho_t := 0$ otherwise. Hence, p_t is similar to the contribution of a frequency to V_t, but only up to the point where the frequency hits a border for the first time. We will show in the following that $\mathrm{E}[V_t]$ can nevertheless be related to $E_t := \mathrm{E}[\rho_t]$. First note that ρ_t is related to the lifetime contribution via $\mathrm{E}[P_i] = \mathrm{E}[\sum_{t=0}^{\infty} \rho_t] = \sum_{t=0}^{\infty} E_t$, since ρ_t is zero after the frequency hits a border. On the other hand, for a fixed $t \in [t_2, t_3]$ let us estimate $V_{t,\text{low}} := \sum_{i\in S_{\text{low}}} p_{i,t}(1-p_{i,t})$. Assume that frequency i leaves the border at some time $t - \tau \in [t_1, t]$. If it does not hit a border until time t, then it contributes ρ_τ to $V_{t,\text{low}}$. The same is true if it does hit a border, and doesn't leave the lower border again in the remainder of the epoch, since then $i \notin S_{\text{low}}$ and $\rho_\tau = 0$. For the remaining case, assume that i leaves the lower border several times $t - \tau_1, t - \tau_2, \dots, t - \tau_k$, with $\tau_1 > \tau_2 > \dots > \tau_k$. Then $\rho_{\tau_2} = \dots = \rho_{\tau_k} = 0$, and by the same argument as before, the contribution of i to $V_{t,\text{low}}$ is $\rho_{\tau_1} = \sum_{i=1}^{k} \rho_{\tau_k}$, where ρ_{τ_1} may or may not be zero. Therefore, we can compute $\mathrm{E}[V_{t,\text{low}}]$ by summing up a term E_τ for every frequency that leaves the lower border at time $t - \tau$, counting frequencies multiple times if they leave the lower border multiple times. Recall that the number of frequencies $s(t)$ that leave the lower border at time $t - \tau$ has expectation $\mathrm{E}[s(t)] = \Theta(1)$. Therefore,

$$\mathrm{E}[V_{t,\text{low}}] = E\Big[\sum_{\tau=0}^{t-t_1} s_{t-\tau} \cdot E_\tau\Big] = \Theta(1) \sum_{\tau=0}^{t-t_1} E_\tau. \tag{5}$$

The sum on the right hand side is almost $\mathrm{E}[P_i]$, except that the sum only goes to $t - t_1$ instead of ∞. Thus we need to argue that $\sum_{\tau=t-t_1+1}^{\infty} E_\tau$ is not too large. Since $t - t_1 \ge K'2\beta(n) \ge 8K \log K$, we may apply Lemma 6, and obtain that the probability that a frequency does not hit a border state in $\tau > t - t_1$ rounds is $e^{-\Omega(\tau/(K^2 \log K))}$. Hence, we may split the range $[t - t_1 + 1, \infty)$ into subintervals of the form $[i \cdot K^2 \log K, (i+1) \cdot K^2 \log K)$, then the i-th subinterval contributes $O((K^2 \log K)e^{-i})$. Therefore, setting $i_0 := \beta(n)/\log K \ge C \log n$, where we may assume $C > 3$, the missing part of the sum is at most

$$\sum_{\tau=r}^{\infty} e^{-\Omega(\tau/(K^2 \log K))} = O\Big(K^2 \log K \sum_{i=i_0}^{\infty} e^{-i}\Big) = o(1/K),$$

using $K \leq n$ in the last step. This is clearly smaller than the rest of the sum, since already $E_1 \geq 1/K \cdot (1 - 1/K)$. Hence $\mathrm{E}[V_{t,\mathrm{low}}] = \Theta(\mathrm{E}[P_i])$, as required.

For (b), the proof is the same as for (a), except that the number $s'(t)$ of frequencies that leave the upper border at time t is given by $(2 - o(1))\gamma'(t)$, where $\gamma'(t)n$ is the number of frequencies at the upper border at time t. Since $\gamma'(t) = \Theta(1)$, the same argument as in (a) applies.

For (c), a frequency $i \in \{1, \dots, n\} \setminus (S_{\mathrm{low}} \cup S_{\mathrm{upp}})$ is either at the border at time t, or it is never at a border throughout the whole epoch. The former frequencies, which are at the border at time t, contribute $1/n \cdot (1 - 1/n)$ each, which sums to less than 1. For the other frequencies, similar as before, by Lemma 6 the probability that a frequency does not hit a border in $t - t_1 \geq K^2\beta(n)$ rounds is $e^{-\Omega(\beta(n)/\log K)} = o(1/n)$ since $\beta = C \log^2 n$ for a sufficiently large constant C. Therefore, the expected number of such frequencies is $o(1)$, and their expected contribution is $o(1)$. This proves (c). □

The next lemma links the lifetime contribution P_i and P'_i to the hitting time T of the loop-free random walk.

Lemma 9 *Consider the situation of Lemma 7. Assume for $a = 1$ or $a = K - 1$ that $T_{a,\min}$ and $T_{a,\max}$ are a lower and upper bound, respectively, on the expected hitting time of $\{0, K\}$ of every random walk as in (2) with $X_0 = a$. Then the lifetime contributions P_i and P'_i defined in (3) and (4) satisfy*

$$2T_{1,\min} \leq \mathrm{E}[P_i] \leq 2T_{1,\max}.$$
$$2T_{K-1,\min} \leq \mathrm{E}[P'_i] \leq 2T_{K-1,\max}.$$

We say that $\mathrm{E}[P_i] = \Theta(\mathrm{E}[T])$, where T is the hitting time of $\{0, K\}$ for the loop-free random walk starting at 1, and similarly for $\mathrm{E}[P'_i]$.

Proof Any frequency i contributes $p_{i,t}(1 - p_{i,t})$ to P_i. On the other hand, the expected time until the i-th frequency makes a non-stationary step (i.e., it changes by a non-zero amount) is $1/(2p_{i,t}(1 - p_{i,t}))$ (cf. Lemma 2). Therefore, the summed contribution to P_i until the frequency makes one non-zero step is in expectation exactly 1/2. Therefore, by Wald's equation, $\mathrm{E}[P_i] = 1/2 \cdot \mathrm{E}[\#\text{non-stationary steps}]$. However, the loop-free random walk is precisely defined to capture the random walk that the i-th frequency performs with its non-stationary steps, so $2\mathrm{E}[P_i]$ equals the hitting time of $\{0, K\}$ of a random walk as in (2) starting at $X_0 = 1$. This proves the first equation, and the second equation follows analogously. □

Lemmas 8 and 9 together yield the following corollary.

Corollary 10 *Consider the situation of Lemma 7, and let $T_{a,\min}$ and $T_{a,\max}$ be lower and upper bounds, respectively, on the expected hitting time of $\{0, K\}$ of every random walk as in (2) with $X_0 = a$. Assume $T_{1,\min} = \omega(1)$. Then for all $t \in [t_2, t_3]$,*

$$\mathrm{E}[V_t] = O(T_{1,\max} + T_{K-1,\max}) \quad \text{and} \quad \mathrm{E}[V_t] = \Omega(T_{1,\min} + T_{K-1,\min}).$$

By Corollary 10, in order to understand $\mathrm{E}[V_t]$ it suffices to analyse the expected hitting time $\mathrm{E}[T]$ of the loop-free random walk.

4.2 Bounds on the Lifetime of a Frequency

We now give upper and lower bounds on the expected lifetime of every loop-free random walk, assuming that we only have lower and upper bounds $\Delta_{\min}$ and $\Delta_{\max}$ on the drift that hold the whole time. We start with the upper bound.

Lemma 11 *Consider a stochastic process* $\{X_t\}_{t\geq 0}$ *on* $\{0, 1, \dots, K\}$, *variables* Δ_t *that may depend on* $X_0, \dots, X_t$ *and* $\Delta_{\min} > 0$, $1/(2K) \leq \Delta_{\max} \leq 1/2$ *such that for* $\Delta_{\min} \leq \Delta_t \leq \Delta_{\max}$,

$$\Pr\big(X_{t+1} = X_t + 1 \mid X_t\big) = \frac{1}{2}\big(1 + \Delta_t\big) \text{ for all } X_t < K \text{ and}$$
$$\Pr\big(X_{t+1} = X_t - 1 \mid X_t\big) = \frac{1}{2}\big(1 - \Delta_t\big) \text{ for all } X_t > 0.$$

Let T be the hitting time of states 0 or K, then regardless of the choice of the Δ_t,

$$\mathrm{E}[T \mid X_0 = 1] = O(\min\{K^2\Delta_{\max}, K\Delta_{\max}/\Delta_{\min}\}) \text{ and}$$
$$\mathrm{E}[T \mid X_0 = K-1] = O(\min\{K, 1/\Delta_{\min}\}).$$

Remark 2 The most important term for us is $\mathrm{E}[T \mid X_0 = 1] = O(K\Delta_{\max}/\Delta_{\min})$. This is tight, i.e., there is a scheme for choosing Δ_t that yields a time of $\Omega(K\Delta_{\max}/\Delta_{\min})$ if $\Delta_{\min} = \Omega(1/K)$.

Consider $\Delta_t = \Delta_{\max}$ for states $i \leq K/2$ and $\Delta_t = \Delta_{\min}$ for states $i > K/2$. Then with probability $\Theta(\Delta_{\max})$ the random walk never reaches 0. Once it reaches $K/2$, it can be shown that the expected time to reach K or 0 from there is $\Omega(K/\Delta_{\min})$ for $\Delta_{\min} = \Omega(1/K)$. (The latter condition is needed since, if $\Delta_{\min} = o(1/K)$, the random walk would be nearly unbiased and reach a border in expected time $O(K^2) = o(K/\Delta_{\min})$, which contradicts the claimed lower bound of $\Omega(K/\Delta_{\min})$.) We omit the details.

Proof We first give a brief overview over the proof. For $X_0 = 1$ we fix an intermediate state $k_0 = \Theta(1/\Delta_{\max})$ and show, using martingale theory and the upper bound $\Delta_{\max}$ on the drift, that (1) the time to reach either state 0 or state k_0 is $O(1/\Delta_{\max})$, and (2) the probability that k_0 is reached is $O(\Delta_{\max})$. In that case, using the lower bound $\Delta_{\min}$ on the drift, the remaining time to hit state 0 or state K is $O(K/\Delta_{\min})$ by additive drift. The time from k_0 is also bounded by $O(K^2)$ as it is dominated by the expected time a fair random walk would take if state 0 was made reflecting. The statement for $X_0 = K - 1$ is proved using similar arguments, starting from $K - 1$ instead of k_0.

We first show the upper bound for $X_0 = 1$. Let $k_0 = 1/(2\Delta_{\max})$ and note that $k_0 \leq K$ since $\Delta_{\max} \geq 1/(2K)$. Let τ be the first point in time when we either hit 0

or k_0, and let p_ℓ and p_h be the probability to hit 0 or k_0, respectively, at time τ. Now consider $Y_t := X_t^2$ during the time before we hit k_0. Then Y_t has a positive drift, more precisely

$$\begin{aligned}\mathrm{E}[Y_{t+1} - Y_t \mid X_t] &= -X_t^2 + \Pr(X_{t+1} - X_t = 1 \mid X_t) \cdot (X_t^2 + 2X_t + 1) \\ &\quad + \Pr(X_{t+1} - X_t = -1 \mid X_t) \cdot (X_t^2 - 2X_t + 1) \\ &= 1 + 2 \cdot \underbrace{X_t}_{\geq 0} \underbrace{\mathrm{E}[X_{t+1} - X_t \mid X_t]}_{\geq 0} \geq 1.\end{aligned}$$

Therefore, $Z_t := Y_t - t$ is a submartingale (has non-negative drift). By the optional stopping theorem [7, page 502], at time τ we have

$$1 = Z_0 \leq \mathrm{E}[Z_\tau] = p_\ell \cdot 0 + p_h \cdot k_0^2 - \mathrm{E}[\tau]. \tag{6}$$

On the other hand, since $X_t - t \cdot \Delta_{\max}$ is a supermartingale (has non-positive drift), we can do the same calculation and obtain

$$1 = X_0 \geq \mathrm{E}[X_\tau] - \mathrm{E}[\tau]\Delta_{\max} = p_\ell \cdot 0 + p_h \cdot k_0 - \mathrm{E}[\tau]\Delta_{\max}.$$

Solving for $\mathrm{E}[\tau]$ in both equations, we get

$$p_h k_0/\Delta_{\max} - 1/\Delta_{\max} \leq \mathrm{E}[\tau] \leq p_h k_0^2 - 1. \tag{7}$$

Now we ignore the term in the middle and sort for p_h:

$$p_h k_0(1/\Delta_{\max} - k_0) \leq 1/\Delta_{\max} - 1.$$

This is equivalent to

$$p_h \leq \frac{1 - \Delta_{\max}}{k_0(1 - k_0\Delta_{\max})} = \frac{2 - 2\Delta_{\max}}{k_0} \leq \frac{2}{k_0} = 4\Delta_{\max},$$

and plugging this into (7) yields $\mathrm{E}[\tau] = O(1/\Delta_{\max})$.

If state k_0 is reached, we use that the drift is always at least $\Delta_{\min}$. Then a distance of $K - k_0 \leq K$ has to be bridged, and by additive drift (Theorem 24) the expected remaining time until state K or state 0 is reached is $O(K/\Delta_{\min})$.

It is also bounded by $O(K^2)$ as the first hitting time of either state 0 or state K is stochastically dominated by the first hitting time of state K for a fair random walk starting in k_0 when state 0 is made reflecting. This is equivalent to a fair gambler's ruin game with $2K$ dollars (imagine a state space of $\{-K, \dots, 0, \dots, K\}$ where $-K$ and $+K$ are both ruin states), and the game starts with $K - k_0$ dollars. The expected duration of the game is $(K - k_0) \cdot (K + k_0) = O(K^2)$.

Together, we obtain an upper bound of

$$O(1/\Delta_{\max} + p_h \min\{K/\Delta_{\min}, K^2\}) = O(\min\{K^2\Delta_{\max}, K\Delta_{\max}/\Delta_{\min}\})$$

where $1/\Delta_{\max}$ can be absorbed since $1/\Delta_{\max} = O(K) = O(K^2\Delta_{\max})$.

For $X_0 = K - 1$ an upper bound $O(1/\Delta_{\min})$ follows from additive drift as only a distance of 1 has to be bridged, and the drift is at least $\Delta_{\min}$. To show an upper bound of $O(K)$, we again use that the aforementioned fair gambler's ruin game stochastically dominates the sought hitting time. As $X_0 = K - 1$, the game starts with 1 dollar and the expected duration of the game is $1 \cdot (2K - 1) = O(K)$. □

The following lemma gives a lower bound on the lifetime of every loop-free random walk.

Lemma 12 *Consider a stochastic process $\{X_t\}_{t\geq 0}$ on $\{0, 1, \dots, K\}$, variables Δ_t that may depend on $X_0, \dots, X_t$ and $\Delta_{\min} \geq 0$, $\Delta_{\max} \geq (4 \ln K)/K$ such that for $\Delta_{\min} \leq \Delta_t \leq \Delta_{\max}$,*

$$\Pr\big(X_{t+1} = X_t + 1 \mid X_t\big) = \frac{1}{2}\big(1 + \Delta_t\big) \text{ for all } X_t < K \text{ and}$$
$$\Pr\big(X_{t+1} = X_t - 1 \mid X_t\big) = \frac{1}{2}\big(1 - \Delta_t\big) \text{ for all } X_t > 0.$$

Let T be the hitting time of states 0 or K, then regardless of the choice of the Δ_t,

$$\Pr\Big(T > \tfrac{1}{2}K/\Delta_{\max} \mid X_0 = 1\Big) = \Omega(\sqrt{\Delta_{\max}/K} + \Delta_{\min})$$

and

$$\mathrm{E}[T \mid X_0 = 1] = \Omega(\sqrt{K/\Delta_{\max}} + K\Delta_{\min}/\Delta_{\max}).$$

Remark 3 There is a scheme for choosing Δ_t such that the bound on the expectation from Lemma 12 is asymptotically tight.

The scheme uses minimum drift $\Delta_{\min}$ until state 0 or state $\sqrt{K/\Delta_{\max}}$ is reached for the first time. In the latter case we switch to a maximum drift $\Delta_{\max}$. By gambler's ruin, the probability of reaching state $\sqrt{K/\Delta_{\max}}$ can be shown to be at most $1/\sqrt{K/\Delta_{\max}} + 4\Delta_{\min}$, and in this case the remaining time to reach state 0 or K is $O(K/\Delta_{\max})$ by additive drift. We omit the details.

Proof The lower bound on the expectation follows immediately from the lower bounds on the probabilities. We first give an overview of the proof of the lower bound on the expectation. We couple the process with two processes $X_t^{\min}$ and $X_t^{\max}$ that always use the minimum and maximum drift $\Delta_{\min}$ and $\Delta_{\max}$, respectively. The coupling ensures that $X_t^{\min} \leq X_t \leq X_t^{\max}$, hence as long as $X_t^{\min} > 0$ and $X_t^{\max} < K$, the process cannot have reached a border state. We show for both coupled processes that the probability of reaching their respective borders in time $\frac{1}{2}K/\Delta_{\max}$ is small, and then apply a union bound. For the $X_t^{\max}$ process a negligibly small failure probability follows from additive drift with tail bounds [11] and the condition $\Delta_{\max} \geq (4 \ln K)/K$. For the $X_t^{\min}$ process we show that the fair random walk on the integers, starting in state 1, does not reach state 0 in time $\frac{1}{2}K/\Delta_{\max}$ with probability $\Omega(\sqrt{\Delta_{\max}/K})$. In addition, the $X_t^{\min}$ process on the integers never reaches state 0

with probability $\Omega(\Delta_{\min})$ [5, page 351], which yields the second term in the claimed probability.

More specifically, we show that *all* schemes for choosing the Δ_t lead to the claimed probability bound. We couple the random walk with two processes: $X_t^{\min}$ is a random walk on $\{0, 1, \dots\}$ (i. e. with the border K removed) with the minimum drift, i. e. using the minimum possible values for Δ_t: $\Delta_t^{\min} := \Delta_{\min}$ for all t. Moreover, $X_t^{\max}$ is a process on $\{K, K-1, \dots\}$ (i. e. with the border 0 removed) with the maximum drift, $\Delta_t^{\max} := \Delta_{\max}$ for all t. The coupling works as follows: draw a uniform random variable r from [0, 1]. If $r \le 1/2 \cdot (1 - \Delta_t)$, X_t decreases its current state, and the same applies to $X_t^{\min}$ if $r \le 1/2 \cdot (1 - \Delta_t^{\min})$ and $X_t^{\max}$ if $r \le 1/2 \cdot (1 - \Delta_t^{\max})$. Otherwise, the random walks increase their current state. This coupling and $\Delta_t^{\min} \le \Delta_t \le \Delta_t^{\max}$ ensures that for every time step t, we have $X_t^{\min} \le X_t \le X_t^{\max}$.

This implies in particular that, as long as $X_t^{\min} > 0$ and $X_t^{\max} < K$, X_t will not have hit any borders. Let $T_0^{\min}$ be the first hitting time of the $X_t^{\min}$ process hitting state 0 and $T_K^{\max}$ be the first hitting time of the $X_t^{\max}$ process hitting state K. Thus the first hitting time T of the X_t process hitting either state 0 or state K is bounded from below by $T \ge \min\{T_0^{\min}, T_K^{\max}\}$.

In particular, by the union bound we have

$$\Pr\big(T \le K/(2\Delta_{\max})\big) \le \Pr\big(T_0^{\min} \le K/(2\Delta_{\max})\big) + \Pr\big(T_K^{\max} \le K/(2\Delta_{\max})\big) \tag{8}$$

and we proceed by bounding the last two probabilities from above.

By additive drift, it is easy to show that $\mathrm{E}[T_K^{\max}] = \Theta(K/\Delta_{\max})$, and this time is highly concentrated. Using Theorem 25, we have

$$\Pr\big(T_K^{\max} \le K/(2\Delta_{\max})\big) \le e^{-K\Delta_{\max}/4} \le 1/K \tag{9}$$

as $K\Delta_{\max} \ge 4 \ln K$. It remains to analyse $T_0^{\min}$, that is, the time until a random walk with drift $\Delta_{\min}$ on the positive integers, starting at $X_0 = 1$, hits state 0. This time stochastically dominates the time until a fair random walk (with no drift) hits state 0.

For the fair random walk, the probability that state 0 will be hit at time t is [5, III.7, Theorem 2]

$$\frac{1}{t}\binom{t}{\frac{t+1}{2}} \cdot 2^{-t}$$

where the binomial coefficient is 0 in case the second argument is non-integral. Hence

$$\Pr\big(T_0^{\min} > K/(2\Delta_{\max})\big) \ge \sum_{t \ge K/(2\Delta_{\max})} \frac{1}{t}\binom{t}{\frac{t+1}{2}} \cdot 2^{-t}$$

The binomial coefficient (for odd t) is at least $\Omega(2^t/\sqrt{t})$. Hence we get a lower bound of

$$\Pr\left(T_0^{\min} > K/(2\Delta_{\max})\right) > \sum_{t \geq K/\Delta_{\max}, t \text{ odd}} \frac{1}{t} \cdot \Omega\left(\frac{1}{\sqrt{t}}\right)$$
$$= \Omega(1) \cdot \sum_{t \geq K/\Delta_{\max}, \text{ t odd}} \frac{1}{t^{3/2}}.$$

Including terms for even t as $1/t^{3/2} \geq 1/2 \cdot 1/t^{3/2} + 1/2 \cdot 1/(t+1)^{3/2}$ and using $\sum_{t \geq x} \frac{1}{t^{3/2}} \geq \int_x^\infty \frac{1}{t^{3/2}}\,\mathrm{d}t = \frac{2}{\sqrt{x}}$ leads to a lower bound of

$$\Omega(1) \cdot \frac{1}{2} \sum_{t \geq K/\Delta_{\max}+1} \frac{1}{t^{3/2}} = \Omega\left(\frac{1}{\sqrt{K/\Delta_{\max}}}\right) = \Omega(\sqrt{\Delta_{\max}/K}).$$

Hence $\Pr\left(T_0^{\min} \leq K/(2\Delta_{\max})\right) \leq 1 - \Omega(\sqrt{\Delta_{\max}/K})$ and plugging this and (9) into (8) yields

$$\Pr\left(T \leq K/(2\Delta_{\max})\right) \leq 1 - \Omega(\sqrt{\Delta_{\max}/K}) + 1/K \leq 1 - \Omega(\sqrt{\Delta_{\max}/K})$$

as $\sqrt{\Delta_{\max}/K} = \omega(1/K)$. Thus $\mathrm{E}[T] = \Omega(\sqrt{\Delta_{\max}/K} \cdot K/\Delta_{\max}) = \Omega(\sqrt{K/\Delta_{\max}})$.

We only need to prove a lower probability bound of $\Omega(\Delta_{\min})$ in case $\Delta_{\min} = \omega(\sqrt{\Delta_{\max}/K}) = \omega(1/K)$. The sought probability bound then follows from observing that, according to [5, page 351], the $X_t^{\min}$ process never reaches 0 with probability

$$1 - \frac{1/2 \cdot (1 - \Delta_{\min})}{1/2 \cdot (1 + \Delta_{\min})} = \frac{\Delta_{\min}}{1/2 \cdot (1 + \Delta_{\min})} = \Omega(\Delta_{\min})$$

and in that case (8) and (9) yield

$$\Pr\left(T \leq K/(2\Delta_{\max})\right) \leq 1 - \Omega(\Delta_{\min}) + 1/K \leq 1 - \Omega(\Delta_{\min}).$$

□

4.3 Establishing Concentration

Our major tool for showing concentration will be using the Chernoff bound [4] and the Chernoff-Hoeffding bound [4].

The basic idea is that for fixed t, we define for each frequency i a random variable $X_i := p_{i,t}(1 - p_{i,t})$ to capture the contribution of the i-th frequency to $V_t = \sum_{i=1}^n X_i$. In the previous sections we have computed $\mathrm{E}[V_t]$ by studying the expected lifetime $\mathrm{E}[T]$. Concentration of V_t would follow immediately by the Chernoff bound if the random walks of the different frequencies were independent of each other. Unfortunately, this is not the case. However, for the initial case of the stabilisation lemma, Lemma 7 (a), we show that the random walks behave almost independent, which allows us to show the following lemma.

Lemma 13 *Assume the situation of Lemma 7 (a). Then* $V_t = \Omega(\sqrt{K})$ *holds with probability* $1 - e^{-\Omega(\sqrt{K})}$ *for all* $t \in [t_2, t_3]$.

Proof We use an inductive argument over $t \in [t_2, t_3]$. Note that in Lemma 7 the statement gets weaker with increasing C' and C'', so we may assume that they are as large as we want. We claim that if they are chosen appropriately, then for part (b) of the lemma we have $V'_{\min} \geq V_{\min}$ and $V'_{\max} \leq V_{\max}$. Therefore, by induction hypothesis we may assume that $V_{\min} \leq V'_{\min} \leq V_{t'} \leq V'_{\max} \leq V_{\max}$ also holds for $t' \in [t_2, t-1]$. To check the claim for $V'_{\min}$, we write the first statement from Lemma 7 (b) as $V'_{\min} \geq c\sqrt{K} V_{\min}^{1/4}$ for some $c > 0$, making the Ω-notation explicit. Then we use the condition $V_{\min} \leq K^{2/3}/C'$, or equivalently $\sqrt{K} \geq (C' V_{\min})^{3/4}$, and hence $V'_{\min} \geq c\sqrt{K} V_{\min}^{1/4} \geq cC'^{3/4} V_{\min}$, which is larger than $V_{\min}$ if $C' > c^{-4/3}$. This proves the claim for $V'_{\min}$. After fixing C', we inspect the second statement from Lemma 7 (b), which implies in particular $V'_{\max} \leq c' K \sqrt{V_{\max}} / \sqrt{V_{\min}}$ for some c', since replacing a minimum by one of its terms can only make it larger. We plug in the two conditions $V_{\min} \leq K^{2/3}/C'$ and $V_{\max} \geq C'' K^{4/3}$, the latter in the equivalent form $K^{2/3} \leq \sqrt{V_{\max}/C''}$, and obtain

$$V'_{\max} \leq c' K \sqrt{V_{\max}} / \sqrt{V_{\min}} \leq c' \sqrt{C'} K^{2/3} \sqrt{V_{\max}} \leq c' \sqrt{C'/C''} V_{\max} \leq V_{\max},$$

where the last steps holds for fixed C', if we choose C'' sufficiently large. Thus we may assume $V'_{\min} \geq V_{\min}$ and $V'_{\max} \leq V_{\max}$.

As mentioned above, we know that $\mathrm{E}[V_t] = \mathrm{E}[T] = \Omega(\sqrt{K})$ by Corollary 10 and Lemma 12 with trivial drift bounds $\Delta_{\min} = 0$ and $\Delta_{\max} = 1/2$, so it remains to show concentration. Fix $i \in \{1, \dots, n\}$, and consider the random walk that $p_{i,t}$ performs over time. More precisely, we consider one step of this random walk, from t to $t+1$. If the offspring x and y have the same i-th bit, then $p_{i,t+1} = p_{i,t}$, so assume that x and y differ in the i-th bit. We want to understand how the drift of $p_{i,t}$ changes if we condition on what the other frequencies do.

So assume that we have already drawn all bits of the two offspring x and y at time $t+1$ except for the i-th bit. Let $f'(x) := f(x) - x_i$ and $f'(y) := f(y) - y_i$ be the number of one-bits among the $n-1$ uncovered bits of x and y, respectively. Assume also that someone tells us which of x, y is the selected offspring. In some cases, for example if $f'(x) \geq f'(y) + 2$ and x is selected, the probability that $x_i = 1$ is exactly 1/2, since the one-bit is equally likely in x and y, and it does not have any influence on the selection process. In other cases, for example if $f'(x) = f'(y) + 1$ and y is selected, then $\Pr(y_i = 1) = 1$, because this is the only scenario in which y can be selected. However, in *all* cases the selected offspring has probability at least 1/2 to have a one-bit at position i, because the selection process can never decrease the probability that the selected offspring has a one-bit at position i. Therefore, even after conditioning on the steps that all other $p_{j,t}$ perform, we still have a non-negative drift for $p_{i,t}$, i.e., for any collection $(q_j)_{j \in \{1,\dots,n\} \setminus \{i\}}$ of frequencies,

$$\mathrm{E}[p_{i,t+1} - p_{i,t} \mid p_{i,t},\ x_i \neq y_i \wedge \forall j \neq i : p_{j,t+1} = q_j] \geq 0.$$

On the other hand we have the upper bound $p_{i,t+1} - p_{i,t} \leq 1/K$ by definition of the algorithm. Therefore, we can use the following uncovering procedure to force independence between the contributions of different i. In each step, we first uncover for all bits whether the offspring coincide for this bit or not, which is independent for all bits. Then we uncover for all $1 \leq i \leq n$ one after the other whether the value of $p_{i,t+1}$ increases or decreases (or stays the same). Crucially, even conditioned on all the prior information, $p_{i,t}$ still has non-negative drift. Therefore, the associated loop-free random walk still follows the description in Lemma 12 with $\Delta_{\min} = 0$ and $\Delta_{\max} = 1$. Hence, if we uncover the random walks one by one as described above, then still the contribution of the frequencies sum up to $\Omega(\sqrt{K})$ in expectation, and the contribution of the i-th frequency is independent of the contribution of the previous frequencies.[5] Therefore, for a fixed $t \in [t_2, t_3]$ we may apply the Chernoff bound (Lemma 22 with $\delta = 1/2$), and obtain that $V_t = \Omega(\sqrt{K})$ with probability $e^{-\Omega(\sqrt{K})}$. Then the claim follows by a union bound over all $t_2 - t_1 = K^{O(1)}$ values of t, since $K^{O(1)} e^{-\Omega(\sqrt{K})} = e^{-\Omega(\sqrt{K})}$. □

We would like to use a similar argument also in the cases with non-trivial $\Delta_{\min}$ and $\Delta_{\max}$. Unfortunately, it is no longer true that the drift remains lower bounded by $\Delta_{\min} > 0$ if we uncover the random walk steps of the other frequencies. However, the bound still remains true if we condition on *only a few* of the other frequencies. More precisely, if we consider a batch of r frequencies $b_1, \dots, b_r$ for a suitably chosen $r \in \mathbb{N}$, then even if we condition on the values that the two offspring have in the bits $b_1, \dots, b_{r-1}$ then frequency of b_r will still perform a random walk where the drift in each round is in $\Theta(1/(K\sqrt{V_t}))$. Hence, we can couple the random walks of $b_1, \dots, b_{r-1}$ to r *independent* random walks, and apply the Chernoff bound to show that the contribution of this batch is concentrated. Afterwards we use a union bound over all batches.

Formally, we show the following pseudo-independence lemma. Note that there are two types of error events in the lemma. One is the explicit event $\mathcal{E}$, the other is the event that $B \notin \mathfrak{B}$, i.e., that the other frequencies in the batch display an atypical distribution. However, both events are very unlikely if V_t is large, which we may assume after one application of Lemma 13.

Lemma 14 *Consider a vector of probabilities $\boldsymbol{p}_t$ with potential $V_t = \sum_{i=1}^{n} p_{i,t}(1 - p_{i,t})$.*

Let $m = m(n) \geq 3$. Let $S \subseteq \{1, \dots, n\}$ be a random set which contains each position independently with probability $1/m$. Then there is an error event $\mathcal{E}$ of probability $\Pr(\mathcal{E}) = e^{-\Omega(V_t/m)}$ such that, conditioned on $\neg\mathcal{E}$, the following holds for all $i_0 \in S$. Let b_i^1 and b_i^2 be the i-th bit in the first and second offspring, respectively, and let $B := (b_i^j)_{i \in S \setminus \{i_0\}, j \in \{1,2\}}$. There is a set $\mathfrak{B} \subseteq \{0,1\}^{2(m-1)}$ such that $\Pr(B \in \mathfrak{B}) = 1 - e^{-\Omega(\min\{m, V_t/m\})}$ and such that for all $B_0 \in \mathfrak{B}$,

[5] Or more formally: we can couple the contribution of the i-th frequency to a random variable which is independent of the previous contributions, and which gives a lower bound on the contribution of the i-th frequency.

$$\mathrm{E}[p_{i_0,t+1} - p_{i_0,t} \mid \boldsymbol{p}_t, B = B_0, \neg\mathcal{E}] = \Theta\left(\frac{p_{i_0,t}(1-p_{i_0,t})}{K\sqrt{V_t}}\right). \tag{10}$$

Before we prove the lemma, we remark briefly on how we apply it. Recall that our overall proof strategy is to show that V_t is between $V_{\min} = \Omega(K^{2/3})$ and $V_{\max} = O(K^{4/3})$, and then stays in this regime for the remaining runtime. For this regime, by choosing $m = \sqrt{V_{\min}}$, both error events (the event $\mathcal{E}$ and the event $B \notin \mathfrak{B}$) have probability $e^{-\Omega(K^{1/3})} = e^{-\Omega(C^{1/3}\log n)}$, where C is the constant from the assumption $K \geq C\log^3 n$. So for any $n^{O(1)}$ iterations, the error events will not happen if C is sufficiently large.

Proof of Lemma 14 The error event $\mathcal{E}$ is that the contribution of S to V_t deviates from its expectation V_t/m, more precisely, $\sum_{i\in S} p_{i,t}(1-p_{i,t}) \notin [\frac{1}{2}V_t/m, 2V_t/m]$. To estimate its probability note that the contribution of *all* frequencies sum to V_t, so the contribution of the frequencies in S sums to V_t/m in expectation. We apply the Chernoff bound to random variables where the i-th random variable takes value $p_{i,t}(1-p_{i,t})$ if $i \in S$ (with probability $1/m$), and value 0 otherwise. Hence we apply Lemma 22 with $b = 1$. We obtain that the probability that the contribution of the frequencies in S deviates from its expectation by more than $\frac{1}{2}V_t/m$ is at most $e^{-\Omega(V_t/m)}$, as required.

We uncover the offspring in three steps. First we uncover all bits in $S \setminus \{i_0\}$, then we uncover the bits in $\bar{S} := \{1,\dots,n\} \setminus S$, and finally we uncover i_0. We call d_1 and d_2 the difference of the fitnesses of the uncovered bits in the first and second uncovering steps, respectively. Assume first that $|d_1 + d_2| \geq 2$. Then the values of i_0 in the two offspring do not have an effect on the selection step, and by symmetry $p_{i_0,t}$ performs a (possibly stagnating) unbiased random walk step. On the other hand, assume that $d_1 + d_2 = 0$, and that the two i_0-bits in the offspring are different. Then the offspring which has a one-bit in i_0 will always be selected. (The case $d_1 + d_2 = \pm 1$ contributes similarly as the case of zero difference, but is not needed for the argument.)

For the upper bound on the drift, assume that $d_1 = k$ for some $k \in \mathbb{Z}$. Note that the frequencies in $\bar{S}$ contribute at least $V_t/2$ to V_t, with room to spare. In particular, by the general probability bound for Poisson-Binomial distributions [1], $\Pr(d_2 \in \{-k-1, -k, -k+1\}) = O(1/\sqrt{V_t})$. Since this holds for any value of k, analogously to Lemma 2 we obtain

$$\mathrm{E}[p_{i_0,t+1} - p_{i_0,t} \mid \boldsymbol{p}_t, B = B_0, \neg\mathcal{E}] = O\left(\frac{p_{i_0,t}(1-p_{i_0,t})}{K\sqrt{V_t}}\right).$$

For the lower bound, we use a similar argument, but we need to be more careful since $\Pr(d_2 = -k) = \Omega(1/\sqrt{V_t})$ holds only if $|k| \leq \eta\sqrt{V_t}$ for a sufficiently small constant $\eta > 0$ [23, Lemma 2.5]. Thus the claim will follow as before if we define $\mathfrak{B} := \{B_0 \in \{0,1\}^{2(m-1)} \mid |d_1(B_0)| \leq \eta\sqrt{V_t}\}$. It only remains to check that $\Pr(B \notin \mathfrak{B}) = e^{-\Omega(\min\{m, V_t/m\})}$. To this end, we proceed in two steps. First, let S' be

the set of all positions $i \in S$ such that the two offspring differ in the i-th bit. We claim that then $|S'| \leq 4V_t/m$ with probability $e^{-\Omega(V_t/m)}$. Indeed, this follows from the Chernoff bound by using $|S|$ indicator random variables X_i, where $X_i = 1$ if $i \in S'$, and $X_i = 0$ otherwise. In a second step, we use $|S'|$ random variables Y_i, where for $i \in S'$ we set $Y_i = +1$ if the first offspring has a one-bit in i and the second offspring has a zero-bit in i, and $Y_i = -1$ otherwise. (Recall that by definition of S', the offspring differ in the bits in S'.) By symmetry $\mathrm{E}[Y_i] = 0$ for all i, and $d_1(B) = \sum_{i \in S'} Y_i$. Now we apply the Chernoff-Hoeffding bound, Lemma 23 to the random variables $(Y_i + 1) \in \{0, 2\}$ with $t = \eta\sqrt{V_t}$ and $b = 4|S'|$ and obtain that $\Pr(|\sum_{i \in S'} Y_i| \geq \eta\sqrt{V_t}) \leq 2e^{-\eta^2 V_t/(2|S'|)} = e^{-\Omega(m)}$, as required. □

We note that from Lemma 14 we may derive the following corollary.

Corollary 15 *In the situation of Lemma 7 with $V_{\min} = \omega(\log^2 K)$ and $V_{\min} \leq K^2$, we may split the set of frequencies randomly into $m = \sqrt{V_{\min}}$ batches of size $\Theta(n/m)$, such that for every batch S there are independent random walks $(L_{i,t})_{i \in S, t \geq 0}$ and $(U_{i,t})_{i \in S, t \geq 0}$ which both satisfy the recurrence (1), and such that $L_{i,t} \leq p_{i,t} \leq U_{i,t}$ holds for all off-border frequencies $i \in \{1, \dots, n\}$ and all $t_1 \leq t \leq t_2$ with probability at least $1 - e^{-\Omega(\sqrt{V_{\min}})}$.*

Proof For each frequency we decide randomly (independently) to which batch it belongs. Then each batch satisfies the description of Lemma 14, and with sufficiently large probability all batches have size $\Theta(n/m)$ by the Chernoff bound (since $V_{\min} \leq V_t \leq n$ we have $m \leq \sqrt{n}$). The coupling is an immediate consequence of Lemma 14, which states that for any value of the other frequencies in the batch, the frequency i_0 still performs a random walk that satisfies the recurrence (10). It just remains to check the error probabilities.

For a single time step, in Lemma 14 we have $\Pr(B \notin \mathfrak{B}) = e^{-\Omega(\sqrt{V_{\min}})}$. By a union bound over all $t_2 - t_1 = K^{O(1)}$ time steps, the probability that there is *any* iteration with $B \notin \mathfrak{B}$ is at most $K^{O(1)} e^{-\Omega(\sqrt{V_{\min}})} = e^{-\Omega(\sqrt{V_{\min}})}$, where the last equality follows since $\sqrt{V_{\min}} = \omega(\log K)$. Similarly, for a single round and a single frequency the probability of the error event is $\Pr(\mathcal{E}) = e^{-\Omega(\sqrt{V_{\min}})}$. The number of rounds is $t_2 - t_1 = K^{O(1)}$, and by Lemma 6 with probability $1 - e^{-\Omega(K)} = 1 - e^{-\Omega(\sqrt{V_{\min}})}$ there are only $K^{O(1)}$ frequencies starting from the boundaries in this epoch. By a union bound over all rounds and all off-border frequencies, the probability that there is ever an error event is at most $K^{O(1)} e^{-\Omega(\sqrt{V_{\min}})} = e^{-\Omega(\sqrt{V_{\min}})}$, since $\sqrt{V_{\min}} = \omega(\log K)$. □

Corollary 15 allows us to partition the frequencies randomly into m batches, such that in each batch the frequencies perform random walks that can be coupled to *independent* random walks. In particular, we will be able to apply the Chernoff-Hoeffding bounds to each batch. This gives concentration of the V_t as follows.

Lemma 16 *Assume the situation of Lemma 7 (b), in particular* $V'_{\min} = \Omega(\sqrt{K}V_{\min}^{1/4})$ *and* $V'_{\max} = O(K \min\{K, \sqrt{V_{\max}}\}/\sqrt{V_{\min}})$ *where we may choose the hidden constants suitably. Then with probability* $1 - \exp(-\Omega(\min\{\sqrt{V_{\min}}, \sqrt{K}/V_{\min}^{1/4}\}))$, *for all* $t \in [t_2, t_3]$, *we have* $V'_{\min} \le V_t \le V'_{\max}$.

Proof Apart from the complication with the batches, the proof is analogous to the proof of Lemma 13.

For simplicity we shall abbreviate $q := \min\{\sqrt{V_{\min}}, \sqrt{K}/V_{\min}^{1/4}\}$, and note that $K^{O(1)}e^{-\Omega(q)} = e^{-\Omega(q)}$. Therefore, it suffices to show all statements for a single t, since we can afford a union bound over all $K^{O(1)}$ values of t. More precisely, as for Lemma 13 we use induction on $t \in [t_2, t_3]$, and we may choose the constants C', C'' in Lemma 7 such that $V'_{\min} \ge V_{\min}$ and $V'_{\max} \le V_{\max}$. Therefore, by induction hypothesis we may assume that $V_{\min} \le V'_{\min} \le V_{t'} \le V'_{\max} \le V_{\max}$ also holds for $t' \in [t_2, t-1]$.

For every $1 \le i \le n$, we define a random variable $X_i := p_{i,t}(1 - p_{i,t})$, and we are interested in $V_t = \sum_{i=1}^{n} X_i$. The frequencies perform a random walk with drift between $\Omega(p_{i,t}(1-p_{i,t})/(K\sqrt{V_{\max}}))$ and $O(p_{i,t}(1-p_{i,t})/(K\sqrt{V_{\min}}))$. Therefore, the loop-free random walk with state space $\{1, \dots, K\}$ has drift between $\Delta_{\min} := 1/\sqrt{V_{\max}}$ and $\Delta_{\max} := 1/\sqrt{V_{\min}}$. By Corollary 10, $\mathrm{E}[V_t] = \Theta(\mathrm{E}[T])$, where T is the lifetime of a random walk on $\{1, \dots, K\}$ with drift between $\Delta_{\min}$ and $\Delta_{\max}$. By Lemma 11,

$$\mathrm{E}[T] = O(K\Delta_{\max} \min\{K, 1/\Delta_{\min}\}) = O(K \min\{K, \sqrt{V_{\max}}\}/\sqrt{V_{\min}}), \tag{11}$$

and by Lemma 12 (where the precondition $\Delta_{\max} \ge (4 \ln K)/K$ follows from $V_{\min} = O(K^{2/3})$ with room to spare) we have

$$\mathrm{E}[T] = \Omega(\sqrt{K/\Delta_{\max}}) = \Omega(\sqrt{K}V_{\min}^{1/4}). \tag{12}$$

Now we split the set $\{1, \dots, n\}$ of frequencies into $m := \sqrt{V_{\min}}$ batches as in Corollary 15. Since each frequency enters the batch with probability $1/m$, the contribution $X_S := \sum_{i \in S} X_i$ of the frequencies in S is

$$\mathrm{E}[X_S] = \Theta(\mathrm{E}[V_t]/m) = \Theta(\mathrm{E}[T]/m).$$

Even after conditioning on the random walks of the other frequencies, by Corollary 15 the i-th frequency of the batch still performs a random walk with drift between $p_{i,t}(1-p_{i,t})/(K\sqrt{V_{\max}})$ and $p_{i,t}(1-p_{i,t})/(K\sqrt{V_{\min}})$, with an error probability of $e^{-\Omega(\sqrt{V_{\min}})}$. Thus its loop-free random walk still has drift between $\Delta_{\min}$ and $\Delta_{\max}$. Therefore, the contribution of the i-th frequency stays the same even after conditioning on the contribution of the other frequencies in the batch. Hence, we may apply the Chernoff bound, and the probability that X_S deviates from its expectation by more than a factor of 2 is at most $e^{-\Omega(\mathrm{E}[X_S])} = e^{-\Omega(\sqrt{K}V_{\min}^{1/4}/m)} = \exp(-\sqrt{K}/V_{\min}^{1/4})$.

By a union bound over all $K^{O(1)}$ batches, the contribution of every batch is within a factor of 2 from its expectation. Therefore, $\mathrm{E}[V_t]/2 \le V_t \le 2\mathrm{E}[V_t]$, and the lemma follows from (11) and (12). □

Altogether, we have proven the Stabilisation Lemma 7: part (a) is proven in Lemma 13, and part (b) is proven in Lemma 16.

5 Proof of the Main Result

With the Stabilisation Lemma in place, we now prove the three statements in our main result, Theorem 1. We first show the first statement in Theorem 1 about too large step sizes, which is implied by the following slightly more detailed theorem.

Theorem 17 *If $K \leq \varepsilon \log n$ for any constant $0 < \varepsilon < 1/\log(10) \approx 0.301$ the runtime of the cGA on OneMax with probability $1 - 2^{-\Omega(n^{\Omega(1)})}$ is at least $2^{cn^{\Omega(1)}}$ for a suitable constant $c > 0$.*

The condition $K \leq \varepsilon \log n$ makes sense as we suspect from closely related results on the UMDA [2, 24] that the cGA optimises OneMax in expected time $O(n \log n)$ if $K \geq c \log n$ for a sufficiently large constant $c > 0$.

The main idea behind the proof of Theorem 17 is that if the step size $1/K$ is too large then frequencies frequently hit the lower border due to the large variance in the stochastic behaviour of frequencies. To keep the paper streamlined and focused on the medium step size regime, a proof of Theorem 17 is placed in the "appendix".

The following lemma is used to prove the remaining two statements in Theorem 1.

Lemma 18 *With probability $1 - \exp(-\Omega(K^{1/4}))$, we have $V_{\min} = \Omega(K^{2/3})$ and $V_{\max} = O(K^{4/3})$ after $i^* = O(\log \log K)$ epochs of length $r = K^2\beta(n)$.*

Moreover, for any fixed $t \geq i^* r$, as long as $\gamma(\tau) = \Omega(1)$ for all $\tau \in [i^* r, t-1]$, $V_{\max}$ and $V_{\min}$ are bounded in the same way during $[i^* r, t]$, with a failure probability of at most $t/r \cdot \exp(-\Omega(K^{1/3}))$, and with probability $1 - tn\exp(-\Omega(\beta(n)/\log n))$ the number of off-border frequencies at any time $t \in [i^* r, t]$ is at most $4K^2\beta(n)$. In particular, if $t = n^2$, $\beta(n) = C\log^2 n$, and $K \geq C\log^3 n$ for a sufficiently large constant $C > 0$, then the error probability is $o(1)$.

Proof By Lemma 4, we know that the initial fraction of frequencies at the lower border is $\Omega(1)$, with probability $1 - e^{-\Omega(\sqrt{n})}$. We apply the first statement of the Stabilisation Lemma 7 (a) with respect to an initial epoch of length r and obtain that with probability $1 - e^{-\Omega(\sqrt{K})}$ we have $V_t = \Omega(K^{1/2})$ in a epoch $[t_2, t_3]$ of length at least r. Applying the statement again, now with respect to this epoch and with the assumption $V_{\min} = \Omega(K^{1/2})$, we obtain $V_{\min} = \Omega(K^{5/8})$ for the next epoch, with error probability $\exp(-\Omega(\min\{\sqrt{V_{\min}}, \sqrt{K}/V_{\min}^{1/4}\})) = \exp(-\Omega(K^{1/4}))$. Iterating this argument i times, we have $V_{\min} = \Omega(K^{2/3-(2/3)(1/4)^{i+1}})$ after i epochs of length r, and each error probability is at most $\exp(-\Omega(K^{1/4}))$. In particular, choosing $i^* = c \ln \ln K$ for a sufficiently large constant $c > 0$, we get $V_{\min} = \Omega(K^{2/3-1/\log K}) = \Omega(K^{2/3})$ after $i^*/2$ iterations, with error probability $\exp(-\Omega(K^{1/4}))$ in each step.

Applying the second statement of the Stabilisation Lemma 7 with respect to the i^*-th epoch, we obtain with error probability $\exp(-\Omega(K^{1/3}))$ that $V_{\max} = O(K^2)$ for the next epoch. We apply the statement again, and the next epoch will satisfy $V_{\max} = O(K\sqrt{K^2/K^{2/3}}) = O(K^{5/3})$. Iterating this argument using the new value of $V_{\max}$ and still $V_{\min} = \Omega(K^{2/3})$ for $O(\log\log K)$ epochs similarly as above, we arrive at $V_{\max} = O(K^{4/3})$, with an error probability of $i^*/2 \cdot \exp(-\Omega(K^{1/3})) = \exp(-\Omega(K^{1/3}))$.

For $t \geq i^* r$, we may apply the same argument again, getting an error probability of $\exp(-\Omega(K^{1/3}))$ for each epoch. The statement on $V_{\min}$ and $V_{\max}$ then follows from a union bound over all epochs. For the number of off-border frequencies, by Lemma 6 every frequency hits a border after at most $K^2\beta(n)$ rounds with probability $1 - \exp(-\Omega(\beta(n)/\log n))$. By a union bound over all frequencies and all rounds, the probability that there is ever a frequency that does not hit a border within $K^2\beta(n)$ rounds is at most $tn\exp(-\Omega(\beta(n)/\log n))$. Therefore, for every τ, the only off-border frequencies at time τ are frequencies that left the border in the last $K^2\beta(n)$ rounds. The expected number of such frequencies is at most $2K^2\beta(n)$, and by the Chernoff bound, Lemma 22, the number exceeds $4K^2\beta(n)$ with probability at most $\exp(-\Omega(K^2\beta(n)))$, which is negligible compared to $\exp(-\Omega(\beta(n)/\log^2 n))$. This proves the statement on the number of off-border frequencies.

Finally, the statement for $t = n^2$ follows since $n^2 e^{-\Omega(\log n)} = o(1)$ if the hidden constant is large enough. □

We are finally ready to prove our main result.

Proof of Theorem 1 As mentioned earlier, the first statement follows from Theorem 17.

Concerning the second statement, a lower bound of $\Omega(\sqrt{n}K + n\log n)$ was shown in [22]. Hence it suffices to show a lower bound of $\Omega(K^{1/3}n)$ for $K \geq C\log^3 n$, where we may choose the constant C to our liking. In the following, we assume that all events that occur with high probability do occur.

Recall that the potential $\varphi_t := \sum_{i=1}^{n}(1 - p_{i,t})$ is the total distance of all frequencies to the optimal value of 1. By Lemma 5, we have a $\gamma_0 = \Omega(1)$ fraction of frequencies at the lower border at some time within the first $O(K^2)$ iterations with probability $1 - e^{-\Omega(K^2\beta(n))} - e^{-\Omega(\sqrt{n})}$. In particular, this implies $\varphi_t \geq \gamma_0(n-1)$.

Let $\xi := 1 - \gamma_0/8$. We show that the time until either φ_t has decreased to $\gamma_0/4 \cdot (n-1)$ or a solution with fitness at least ξn is found is $\Omega(K^{1/3}n)$ with high probability. This implies the second and third statements since in an iteration where $\varphi_t > \gamma_0/4 \cdot (n-1)$ the expected fitness is at most $n - \gamma_0/4 \cdot (n-1)$ and the probability of sampling a solution with fitness at least ξn is $2^{-\Omega(n)}$ by Chernoff bounds. This still holds when considering a union bound over $O(K^{1/3}n)$ steps.

By Lemma 18, with probability $\exp(-\Omega(K^{1/4})) = o(1)$ we will have $V_t = O(K^{4/3})$ after $T = O(r\log\log K) = o(n)$ steps. By Lemma 5, with high probability we will still have at least $\gamma_0/2 \cdot (n-1)$ frequencies at the lower border.

Moreover, also by Lemma 18, if we can show $\gamma(t) = \Omega(1)$ then the bound $V_t = O(K^{4/3})$ remains true for the next $K^{1/3}n$ rounds, with probability $1 - o(1)$. So it remains to show $\gamma(t) = \Omega(1)$ for $t \in [T, \Omega(K^{1/3}n)]$. Note that the prerequisites

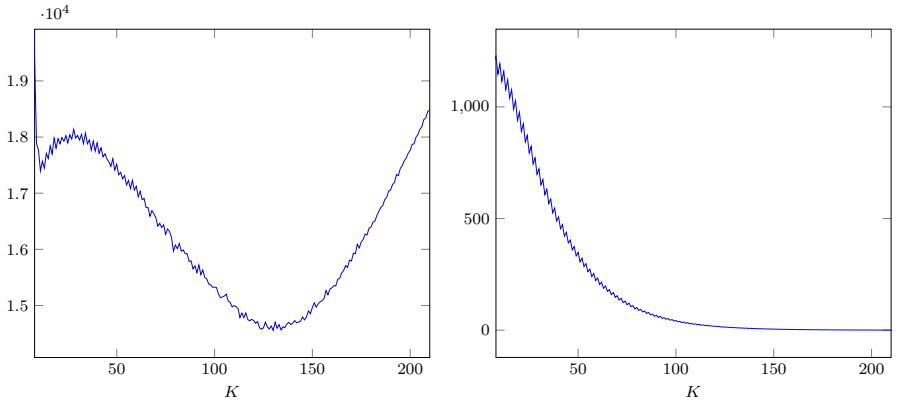

Fig. 2 Left-hand side: empirical runtime of the cGA on OneMax, right-hand side: number of hits of lower border; for $n = 1000$, $K \in \{8, 9, \dots, 210\}$, and averaged over 3000 runs

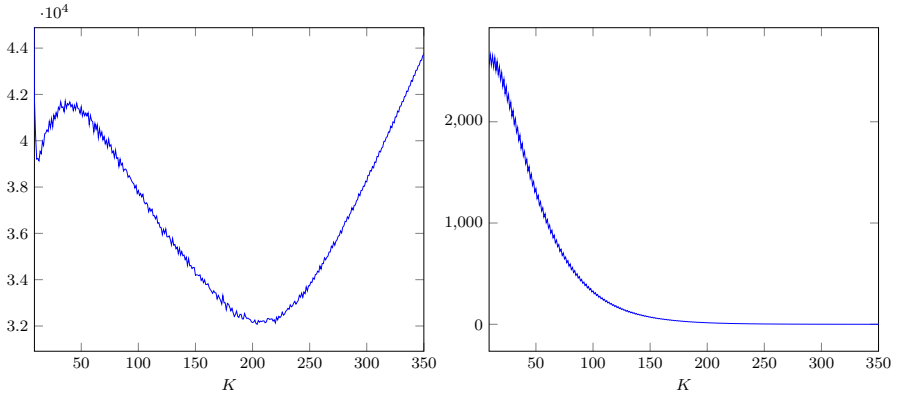

Fig. 3 Left-hand side: empirical runtime of the cGA on OneMax, right-hand side: number of hits of lower border; for $n = 2000$, $K \in \{9, 10, \dots, 350\}$, and averaged over 3000 runs

of Lemma 18 only concern times strictly before t, so we can use the statement of the lemma inductively to show that $\gamma(t) = \Omega(1)$. By Lemma 18, the number of off-border frequencies in each epoch is $O(K^2\beta(n))$, hence while $\varphi_t > \gamma_0/4 \cdot (n-1)$, we have $\gamma(t) \geq \gamma_0/4 - O(K^2\beta(n)/n) = \Omega(1)$ as off-border frequencies (and frequencies at the upper border) only contribute $O(K^2\beta(n)) = o(n)$ to φ_t. Hence Lemma 18 implies that with probability $1 - o(1)$, $V_t = O(K^{4/3})$ holds for all $t \in [T, n^2]$ such that $\varphi_t > \gamma_0/4 \cdot (n-1)$.

By Lemma 3, the drift of φ_t is at most $O\left(\sqrt{V_t}/K\right) = O(K^{-1/3})$ and the change of φ_t is bounded by $\sqrt{V_t}\log n = O(K^{2/3}\log n)$ with probability $1 - n^{-\Omega(K \log\log n)}$, even when taking a union bound over $O(K^{1/3}n)$ steps. Applying Theorem 1 in [11] with a

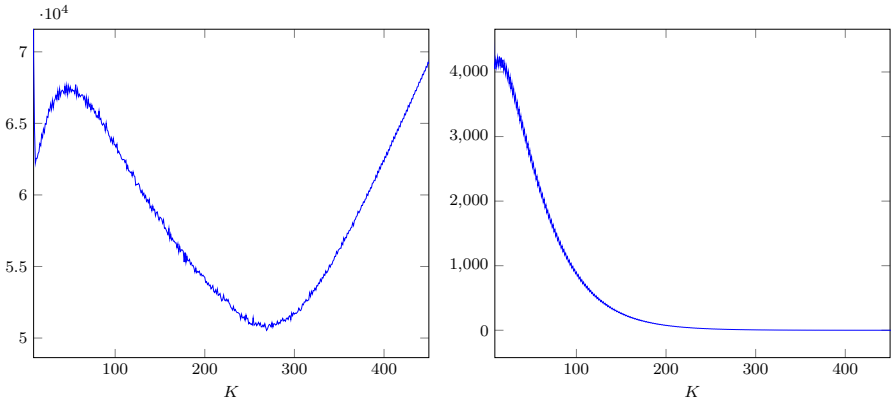

Fig. 4 Left-hand side: empirical runtime of the cGA on OneMax, right-hand side: number of hits of lower border; for $n = 3000$, $K \in \{9, 10, \dots, 400\}$, and averaged over 3000 runs

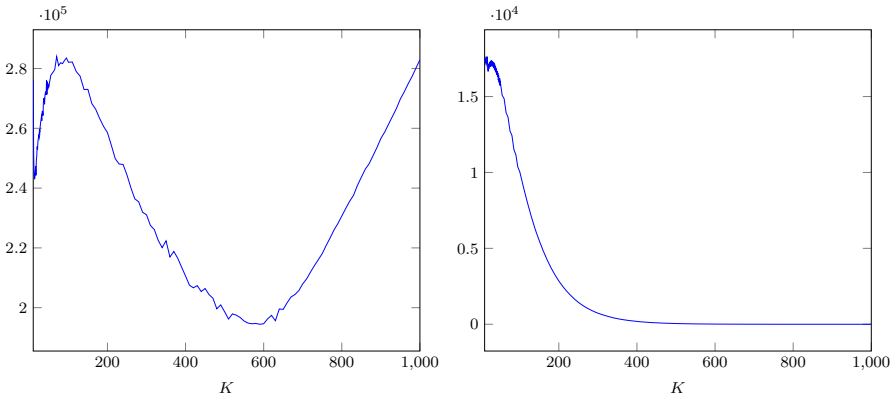

Fig. 5 Left-hand side: empirical runtime of the cGA on OneMax, right-hand side: number of hits of lower border; for $n = 10000$, $K \in \{10, 11, \dots, 50, 55, 60, \dots, 100, 110, 120, \dots, 1000\}$, and averaged over 500 runs

maximum step size of $O(K^{2/3} \log n)$, distance $\gamma_0/4 \cdot (n-1)$ and drift $O(K^{-1/3})$, the time until $\varphi_t \leq \gamma_0/4 \cdot (n-1)$ is at least $\Omega(\gamma_0/4 \cdot (n-1) \cdot K^{1/3}) = \Omega(K^{1/3} n)$ with probability $1 - e^{-\Omega\left(n \cdot K^{-1/3}/(K^{4/3} \log^2 n)\right)} = 1 - e^{-\Omega(n^{1/6}/\log^2 n)}$, where the last step uses $K = O(n^{1/2})$. Adding up failure probabilities completes the proof. □

6 Experiments

We have carried out experiments for the cGA on OneMax to gain some empirical insights into the relationship between K and the runtime. The algorithm was implemented in the C programming language using the WELL512a random number generator.

The experiments supplement our asymptotic analyses and confirm that the algorithm indeed exhibits a bimodal runtime behavior also for small problem sizes. We ran the cGA with $n = 1000$ (Fig. 2), $n = 2000$ (Fig. 3), $n = 3000$ (Fig. 4), all averaged over 3000 runs, and $n = 10000$ (Fig. 5), averaged over 500 runs, as detailed in the figures. In all four cases, we observe the same picture: the empirical runtime starts out from very high values, takes a minimum when K is around 10 and then increases again, e. g., up to $K = 30$ for $n = 1000$. Thereafter it falls again, e. g., up to $K \approx 130$ for $n = 1000$, and finally increases rather steeply for the rest of the range. The location of the first minimum does not change much in the three scenarios, but the second minimum clearly grows with K, roughly from 130 at $n = 1000$ via roughly 210 at $n = 2000$ to finally roughly 590 at $n = 10000$. As n grows, the relative difference between the maximum and second minimum increases as well, from roughly 23 % at $n = 1000$ to roughly 45 % at $n = 10000$. Close inspection of the left part of the plot also shows that the range left of the first minimum leads to very high runtimes. We could not plot even smaller values of K due to exploding runtimes. This is consistent with our exponential lower bounds for $K \leq 0.3 \log n$.

The right-hand sides of the pictures also illustrate that the number of times the lower frequency border is hit seems to decrease exponentially with K. The phase transition where the behavior of frequencies turns from chaotic into stable is empirically located somewhere around the value of K where the second minimum of the runtime is reached.

7 Conclusions

We have investigated the complex parameter landscape of the cGA, highlighting how performance depends on the step size $1/K$. In addition to an exponential lower bound for too large step sizes ($K \leq 0.3 \log n$), we presented a novel lower bound of $\Omega(K^{1/3}n + n \log n)$ for the cGA on OneMax that at its core has a very careful analysis of the dynamic behaviour of the sampling variance and how it stabilises in a complex feedback loop that exhibits a considerable lag. A key idea to handle this complexity was to show that the sampling variance V_t of all frequencies at time t can be estimated accurately by analysing the stochastic behaviour of one frequency i over a period of time.

Assuming that the cGA has the same upper bound as the UMDA for step sizes $K = \Theta(\log n)$, the expected runtime of the cGA is a bimodal function in K with worse performance in between its two minima.

We believe that our analysis can be extended towards an upper bound of $O(K^{2/3}n + n \log n)$, using that typically $V_t = \Omega(K^{2/3})$ after an initial phase, which implies a drift of $\Omega(\sqrt{V_t}/K) = \Omega(K^{-2/3})$ for φ_t. This would require additional arguments to deal with $\gamma(t)$ decreasing to sub-constant values where showing concentration becomes more difficult. Another avenue for future work would be to investigate whether the results and techniques carry over to the UMDA, where the frequencies can make larger steps.

Acknowledgements This paper was initiated at Dagstuhl seminar 17101 "Theory of Randomized Optimization Heuristics" and 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 & Technology).

Funding Open access funding provided by Swiss Federal Institute of Technology Zurich.

A Appendix

This appendix contains proofs omitted from the main part.

A.1 Proof of Lemma 4

The proof uses the following lemma from [22]. The notion of rw-steps and b-steps is briefly explained in the proof of Lemma 2; see [22] for further details.

Lemma 19 (Lemma 10 in [22]) *Consider a frequency of the cGA on OneMax and let p_t be its frequency at time t. Let $t_0, t_1, t_2, \ldots$ be the times where the 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 $-1 \leq s < 0$ constant, we have

$$\Pr\left(T_s \leq \alpha(sK)^2 \text{ or } p_t \text{ exceeds } 5/6 \text{ or reaches } 1/n \text{ before } t_{T_s}\right) \geq \left(\frac{1}{13\sqrt{1/(|s|\alpha)}} - \frac{1}{(13\sqrt{1/(|s|\alpha)})^3}\right)\frac{1}{\sqrt{2\pi}} e^{-\frac{169}{2|s|\alpha}} - O\left(\frac{1}{|s|\sqrt{\alpha K}}\right).$$

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

$$\Pr\left(T_s \geq \alpha(sK)^2 \text{ or a border is reached until time } t_{\alpha(sK)^2}\right) \geq 1 - e^{-1/(4\alpha)}.$$

Before we can apply this lemma, we follow [22] and bound the total effect of b-steps during the first $\Omega(K^2)$ steps. To this end, we need to show that a constant fraction of the frequencies stay in the range [1/6, 5/6]. The following lemma improves Lemma 15 in [22] towards much smaller failure probabilities that are independent of K. This improvement is vital for the upcoming proof of Theorem 17 in Sect. A.2 since Lemma 15 in [22] is trivial for $K = O(\log n)$.

Lemma 20 *Let $K \leq \sqrt{n}$ and $\kappa > 0$ be a small constant. There exists a constant ξ, depending on κ, such that the following properties hold regardless of the last $n/2$ frequencies throughout the first $T := \kappa K^2$ steps of cGA, with probability $1 - e^{-\Omega(\sqrt{n})}$:*

1. *The probability of a b-step at any frequency is always $O(1/\sqrt{n})$ during the first T steps, and*
2. *There is a subset S of ξn frequencies among the first $n/2$ frequencies such that*

(a) *the frequencies in S are always within [1/6, 5/6] during the first T steps,*
(b) *the total number of b-steps for each frequency in S is bounded by $K/6$, leading to a displacement of at most 1/6.*

Proof By Lemma 13 in [22], with probability $1 - e^{-\Omega(n)}$, for at least γn of these frequencies among the first $n/2$ frequencies, 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 of exactly γn of these frequencies. We set $\xi := \gamma/100$.

Now we will use an inductive argument. The inductive statement is that 1. and 2. hold for the first t rounds, and we let t run from 0 to T. More precisely, we will show the following two implications. Firstly, if 2. holds for $t-1$ rounds, then 1. holds for t rounds. Secondly, if 1. holds for t rounds, then 2. also holds for t rounds with probability $1 - e^{-\Omega(\sqrt{n})}$, where the hidden constant is uniform over all t. Note that the error probabilities accumulate to $Te^{-\Omega(\sqrt{n})} = e^{-\Omega(\sqrt{n})}$.

For the first implication, as long as there are at least $\gamma n/2$ frequencies in [1/6, 5/6], according to Lemma 12 in [22], for all frequencies the probability of a b-step in the next round is at most $c_2/\sqrt{n}$ for a positive constant c_2 that only depends on γ. This is exactly the first implication that we need for the inductive statement, so it remains to prove the second implication. So in the following we may assume that 1. holds for t rounds. We remark that for this step we will not use that 2. holds for $t-1$, although this would be a valid assumption. Rather, we show the existence of S from scratch.

As long as 1. holds, consider a fixed frequency in S'. The expected number of b-steps in $t \leq \kappa K^2$ steps is at most $\kappa \cdot c_2 K$. In fact, we will later use a slightly weaker bound of $\kappa \cdot c_3 K$ for some constant $c_3 > c_2$ that we will define later. Each b-step

changes the frequency by $1/K$. A necessary condition for increasing the frequency by a total of at least 1/6 is that we have at least $K/6$ b-steps among the first t steps. Choosing κ small enough to make $\kappa \cdot c_3 K \leq 1/2 \cdot K/6$, by Chernoff bounds the probability to get at least $K/6$ b-steps in t steps is at most $(e/4)^{K/12} \leq (e/4)^{1/12} < 0.97$.

So we conclude that each frequency in S' satisfies the condition in (b) with probability at least 0.03. Moreover, by choice of S', every such frequency automatically also satisfies the condition in (a). Thus the expected number of frequencies that satisfy the conditions in (a) and (b) is at least $0.03\gamma n$. It remains to show concentration, i.e., we show that the number of frequencies which have at most $K/6$ b-steps among the first t steps is concentrated.

The number of b-steps is not independent for different frequencies, so we cannot apply Chernoff bounds. However, we can use the same argument as in the proof of Corollary 15, which we repeat briefly. We split the set S' randomly into $\sqrt{n}$ batches, assigning each frequency independently to a batch. Then each batch satisfies the description of Lemma 14, and with probability $1 - e^{-\Omega(\sqrt{n})}$ all batches have size $\Theta(\sqrt{n})$ by the Chernoff bound. Now consider one fixed batch. By Lemma 14, the frequencies in this batch can be coupled to *independent* random walks with drift $\Theta(1/(K\sqrt{n}))$. Hence, there is $c_3 > 0$ such that the numbers of b-steps of the frequencies in the batch are dominated by *independent* binomial random variables with expectation at most $tc_3/\sqrt{n} \leq K/12$. Each binomial random variable exceeds $K/6$ with probability at most 0.97, so by the Chernoff bound at least 1/100 of the frequencies in the batch make at most $K/6$ b-steps [and thus satisfy the conditions in (a) and (b)], with probability $1 - e^{-\Omega(\sqrt{n})}$. By a union bound, the same is true for all batches simultaneously with error probability $\sqrt{n}e^{-\Omega(\sqrt{n})} = e^{-\Omega(\sqrt{n})}$, and in this case at least $\gamma n/100 = \xi n$ of the frequencies in S' satisfy the conditions in (a) and (b). This concludes the proof of the second implication, and of the lemma. □

Proof of Lemma 4 The proof follows closely arguments from the proof of Theorem 8 in [22], using our improved Lemma 20. For concentration we again need the batch argument as in Corollary 15. We will focus on proving that frequencies are likely to hit the lower border. Since the probability of a frequency hitting the upper border is no smaller than the probability of hitting the lower border, a symmetric statement also holds for frequencies hitting the upper border.

Let $T := \kappa K^2$ for a small enough constant $\kappa > 0$. We first fix one frequency, and we use Lemma 19 to show that some frequencies are likely to walk down to the lower border. Note that Lemma 19 applies for an arbitrary (even adversarial) mixture of rw-steps and b-steps over time. Lemma 20 states that there are $\Omega(n)$ frequencies whose displacement owing to b-steps during the first T steps is at most 1/6. We focus on these frequencies in the following and show that a constant fraction of them reach the lower border.

We shall fix such a frequency i and focus on the effect of its rw-steps during the first T steps. We will apply both statements of Lemma 19, to prove that p_i 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 when the frequency 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 19 now gives us a probability of at least $1 - e^{-1/(4\alpha)} = \Omega(1)$ (using $\alpha = O(1)$) for the event that the frequency does not exceed 5/6. In the following, we condition on this event.

We then revisit the same stochastic process and apply Lemma 19 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 19 again despite dependencies between the two applications.

We now apply the first statement of Lemma 19 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 frequency does not exceed 5/6.

We note that $\alpha = \Theta(1)$ by definition and that $1/\alpha = \Theta(1) = o(K)$ under our assumption $K = \omega(1)$. Now Lemma 19 states that the probability of the random walk reaching a total displacement of $-5/6$ (or hitting the lower border before) is at least

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

Since $K = \omega(1)$, $\alpha = \Theta(1)$ and $|s| = 1$, (13) is at least

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

Combining with the probability of not exceeding 5/6, which we have proved to be constant, the probability of the frequency hitting the lower border within T steps is $\Omega(1)$.

Therefore the expected number of frequencies which reach the lower border is $\Omega(n)$. To show the whp statement, we use the same trick as in the proofs of Corollary 15 and of Lemma 20, and split the set of frequencies into batches of size $\Theta(\sqrt{n})$. Then by Lemma 14, the frequencies in each batch can be coupled to *independent* random walks. This allows us to apply the Chernoff bound and to conclude that within each batch, with probability $1 - e^{-\Omega(\sqrt{n})}$ a constant fraction of the frequencies reaches the lower border. The statement of the lemma is then obtained by a union bound over all batches. We omit the details as they are analogous to Lemma 20. □

A.2 Proof of Theorem 17

Here we prove an exponential lower bound for too large step sizes as stated in Theorem 17.

The following lemma shows that with a good probability, a frequency will reach the lower border through a sequence of steps that are all decreasing. Such a sequence was called *landslide sequence* in the context of a simple ACO algorithm in [17].

Lemma 21 *Consider the cGA on OneMax at a point in time when the number of frequencies at the lower border is at most $n - \Omega(n)$. Then with probability at least $\Omega(10^{-K})$ within the following $O(K \log K)$ steps one of the remaining frequencies will reach the lower border.*

Proof We first estimate transition probabilities for a frequency i with $1/n < p_i < 1 - 1/n$ using arguments from the proof of Lemma 2, but providing bounds on the constants hidden in the $\Theta(\sqrt{V_t})$ terms.

For every $t \geq t^*$ we have

$$\Pr(p_{i,t+1} = p_{i,t} + 1/K \mid p_{i,t}) \leq 2p_{i,t}(1 - p_{i,t})$$

as x_i and y_i need to be sampled differently for $p_{i,t}$ to change. A sufficient condition for $p_{i,t}$ to decrease is that x_i is sampled at 0 (probability $1 - p_{i,t}$), y_i is sampled at 1 (probability $p_{i,t}$) and the fitness difference on all other bits, $D_{i,t} = \sum_{j \neq i}(x_j - y_j)$, to be at least 1. By symmetry, $\Pr(D_{i,t} \geq 1) = \Pr(D_{i,t} \leq -1) = 1/2 \cdot \Pr(D_{i,t} \neq 0)$. Using the general bound for Poisson binomial distributions from [1] (see Theorem 22 in [22]), for all p_j, $j \neq i$,

$$\Pr(D_{i,t} = 0) \leq \frac{1}{2\sqrt{\sum_{j \neq i} p_j(1 - p_j)}} \leq \frac{1}{2\sqrt{(n-1)/n \cdot (1 - 1/n)}} = \frac{n}{2n - 2} \leq \frac{1}{2} + \frac{1}{n}.$$

Together, we have

$$\Pr(p_{i,t+1} = p_{i,t} - 1/K \mid p_{i,t}) \geq p_{i,t}(1 - p_{i,t})\left(\frac{1}{4} - \frac{1}{2n}\right) \geq 2p_{i,t}(1 - p_{i,t})/9$$

for large enough n. Hence the conditional probability of $p_{i,t}$ decreasing, given that it changes, is at least $\frac{2/9 \cdot p_{i,t}(1-p_{i,t})}{(2+2/9)p_{i,t}(1-p_{i,t})} = \frac{1}{10}$.

For the remainder we choose a frequency i with $1/n < p_i < 1 - 1/n$, if such a frequency exists. If no such frequency exists, there must be $\Omega(n)$ frequencies at the upper border and the probability that at least one such frequency detaches from the upper border in the next iteration is at least $\Omega(n) \cdot 1/n \cdot (1 - 1/n) \cdot 1/5 = \Omega(1)$, re-using arguments from above. We assume that this happens, keeping in mind a $\Omega(1)$ factor in the claimed probability (and absorbing the additional iteration in the time bound), and choose one such frequency i.

A sufficient condition for frequency i reaching the lower bound before returning to the upper bound is that $p_{i,t}$ always decreases if it is changed. This needs to happen

at most K times, yielding a probability of at least 10^{-K} as claimed. The expected time for this sequence of events to happen is at most

$$\sum_{j=1}^{K-1} \frac{1}{\Pr(p_{i,t+1} = p_{i,t} - 1/K \mid p_{i,t} = 1 - 1/n - j/K)}$$
$$\leq \sum_{j=1}^{K-1} \frac{5}{(1/n + j/K)(1 - 1/n - j/K)} \leq \sum_{j=1}^{K/2} \frac{10}{1/n + j/K} \leq 10K \sum_{j=1}^{K/2} \frac{1}{j} \leq 10K(\ln(K) + 1).$$

By Markov's inequality, the probability that the time is at most $20K(\ln(K) + 1)$ is at least 1/2. Absorbing this factor in the term $\Omega(10^{-K})$ completes the proof. □

Proof of Theorem 17 According to Lemma 4, with probability $1 - e^{-\Omega(\sqrt{n})}$ at least $\gamma_0 n$ frequencies reach their lower border within the first $t^* = O(K^2)$ iterations, for some constant $\gamma_0 > 0$. As argued in Lemma 5, frequencies that hit the lower border before time t^* have a chance to leave the border again. However, since the probability of a frequency detaching from the lower border is at most $2/n$ irrespective of other frequencies, the probability that at time t^* there will be at least $\gamma_0 n/2$ frequencies at the lower border is $1 - 2^{-\Omega(n)}$ by Chernoff bounds.

Let γ_t denote the number of frequencies at the lower border at iteration t. We consider periods of $T = O(K \log K)$ iterations, where the O-term is the one from Lemma 21, and how the number of frequencies at the lower border changes in expectation during such a period. By Lemma 21, if $\gamma_t \leq n - \Omega(n)$, the number increases by 1 with probability at least $p^+ = \Omega(10^{-K})$ and since every frequency at the lower border detaches only with probability at most $2/n$, we have

$$\mathrm{E}[\gamma_{t+T} - \gamma_t \mid \gamma_t, \gamma_t \leq n - \Omega(n)] \geq p^+ - \frac{2\gamma_t T}{n}.$$

Note that for every frequency i, every time t and all remaining frequencies, the probability that frequency i is at the lower border at time $t + 1$ is maximised if frequency i is already at the lower border at time t. More formally, the sought probability is at least $1 - 2/n$ if $p_{i,t} = 1/n$, it is at most $(1/n + 1/K)(1 - 1/n - 1/K) \ll 1 - 2/n$ if $p_{i,t} = 1/n + 1/K$ (by Lemma 2) and it is 0 otherwise. Hence we are being pessimistic if we underestimate the number of frequencies at the lower border.

We argue in the following that the number of frequencies at the lower border stochastically dominates that of a simpler Markov chain $Z_0, Z_1, Z_2, \ldots$ defined as follows. One step of the Z-process reflects a simplified view of T iterations of the cGA, with Z_t being defined so that it is stochastically dominated by the number of frequencies at the lower border after $t \cdot T$ iterations of the cGA. We will define the Z-process so that it is capped: $Z_t \in [0, b + 1]$ for a value $b \leq n - \Omega(n)$ chosen later. The value of Z_{t+1} is determined by starting with Z_t, subtracting $Z_t \cdot T$ independent Bernoulli variables with parameters $2/n$ and, if and only if $Z_t \leq b$, adding the outcome of a Bernoulli trial with parameter p^+.

The simpler process Z_t is stochastically dominated by γ_{t^*+tT} since $Z_t = \min\{\gamma_{t^*+tT}, b + 1\}$ (thus in particular $Z_0 \leq \gamma_{t^*}$) and all transition probabilities are estimated pessimistically: for all $d \geq 1$ and all $i \leq b + 1$ we have

$$\Pr(Z_{t+1} = Z_t + d \mid Z_t = i) \leq \Pr(\gamma_{t^*+tT+T} = \gamma_{t^*+tT} + d \mid \gamma_{t^*+tT} = i)$$

as the left-hand side is 0 for $d > 1$ or $i = b + 1$ and $p^+ \cdot (1 - 2/n)^{iT}$ otherwise, which is a lower bound for $\Pr(\gamma_{t^*+tT+T} = \gamma_{t^*+tT} + 1 \mid \gamma_{t^*+tT} = i)$ by Lemma 21 and the fact that all i frequencies at the lower border remain there for T steps with probability at least $(1 - 2/n)^{iT}$. Furthermore, for all $d \geq 1$,

$$\begin{aligned}\Pr(Z_{t+1} = Z_t - d \mid Z_t = i) &= (1 - p^+)(2/n)^d(1 - 2/n)^{iT-d} + p^+(2/n)^{d+1}(1 - 2/n)^{iT-d-1} \\ &\geq \Pr(\gamma_{t^*+tT+T} = \gamma_{t^*+tT} - d \mid \gamma_{t^*+tT} = i)\end{aligned}$$

where the last inequality follows from the same arguments as above. Note that state $b + 1$ is a reflective state, but this does not affect the drift estimates for states $Z_t \leq b$ as the process can only increase by 1 in each step.

We apply the negative drift theorem [18, 19] in the variant with self-loops [20], stated as Theorem 26 in Sect. B.2, to the process $Z_1, Z_2, \ldots$. The interval is chosen as $[a, b]$ with $a := b/2$ and $b + 1 := \min\{p^+ n/(4T), \gamma_0 n/2\}$, such that we start at a state at least b. This implies

$$\mathrm{E}[Z_{t+1} - Z_t \mid Z_t, Z_t \leq b] \geq p^+ - \frac{2(b+1)T}{n} \geq \frac{p^+}{2}$$

and also that the converse of the self-loop probability is $\Pr(Z_{t+1} \neq Z_t \mid Z_t) \leq 3p^+/2$ by a union bound over all Bernoulli trials. This establishes the first condition of the negative drift theorem with self-loops.

To establish the second condition, note that $\Pr(Z_{t+1} \neq Z_t \mid Z_t)$ is bounded from below by p^+ if $Z_t \leq b$ and $1 - (1 - 2/n)^{(b+1)T} = 1 - (1 - 2/n)^{p^+ n/4} = \Omega(p^+)$ if $Z_t = b + 1$ and $b + 1 = p^+ n/(4T)$; if $b + 1 = \gamma_0 n/2$ a lower bound of $\Omega(1) = \Omega(p^+)$ follows in the same way. Hence for all Z_t and all $d \geq 1$

$$\begin{aligned}\Pr(Z_{t+1} = Z_t - d \mid Z_t) &\leq \binom{Z_t T}{d}\left(\frac{2}{n}\right)^d \leq \left(\frac{2Z_t T}{n}\right)^d \\ &\leq \frac{4Z_t T}{n} \cdot 2^{-d} \leq p^+ \cdot 2^{-d} \leq r \cdot \Pr(Z_{t+1} \neq Z_t \mid Z_t) \cdot 2^{-d}\end{aligned}$$

when choosing $r = O(1)$ appropriately. Along with $\Pr(Z_{t+1} = Z_t + 1 \mid Z_t) \leq p^+ \leq r \cdot \Pr(Z_{t+1} \neq Z_t \mid Z_t) \cdot 2^{-1}$ and $\Pr(Z_{t+1} = Z_t + d \mid Z_t) = 0$ for $d > 1$, this establishes the second condition of the negative drift theorem. Invoking said theorem and noting that $(b - a)/r = \Omega(p^+ n/T) = \Omega((10)^{-K} n/(K \log K)) = \Omega(n^{1-\varepsilon \log(10)}/(\log(n) \log\log n))$ shows that with probability $1 - 2^{-\Omega(n^{1-\varepsilon \log(10)}/(\log(n) \log\log n))}$ the time to reduce the number of frequencies at the lower border below $a = \Omega(n^{1-\varepsilon \log(10)}/(\log(n) \log\log n))$ is at least $2^{cn^{1-\varepsilon \log(10)}/(\log(n) \log\log n)}$ for a suitable constant $c > 0$. Note that while $\gamma_t \geq a$ the probability of sampling the optimum in one iteration is at most $2n^{-a}$ since at least a frequencies at the lower border have to be sampled at 1 in one of the two search points. Taking a union bound over $2^{cn^{1-\varepsilon \log(10)}/(\log(n) \log\log n)}$ iterations still yields a failure probability that is absorbed in the term $1 - 2^{-\Omega(n^{1-\varepsilon \log(10)}/(\log(n) \log\log n))}$.

Noting that the exponent can be simplified using $\varepsilon < 1/\log(10)$ to $n^{1-\varepsilon\log(10)}/(\log(n)\log\log n) = n^{\Omega(1)}$ completes the proof. □

B Mathematical Tools

B.1 Chernoff Bounds

Lemma 22 (*Chernoff Bound* [4]) *Let $b > 0$. Let $X_1, \ldots, X_n$ be independent random variables (not necessarily i.i.d.) that take values in $[0, b]$. Let $S = \sum_{i=1}^{n} X_i$ and $\mu = \mathrm{E}[S]$. Then for all $0 \le \delta \le 1$,*

$$\Pr(S \le (1-\delta)\mu) \le e^{-\delta^2\mu/(2b)}$$

and

$$\Pr(S \ge (1+\delta)\mu) \le e^{-\delta^2\mu/(3b)}.$$

Lemma 23 (*Chernoff-Hoeffding Bound* [4]) *Let $X_1, \ldots, X_n$ be independent random variables, where X_i has values in $[0, b_i]$ (not necessarily i.i.d.). Let $S = \sum_{i=1}^{n} X_i$, and let $b := \sum_{i=1}^{n} b_i^2$. Then*

$$\Pr(S - \mathrm{E}[S] \ge t) \le e^{-2t^2/b}$$

and

$$\Pr(|S - \mathrm{E}[S]| \ge t) \le 2e^{-2t^2/b}.$$

B.2 Drift Theorems

Theorem 24 (*Additive Drift* [10]) *Let $(X_t)_{t\ge 0}$ be a sequence of non-negative random variables over a finite state space $S \subseteq \mathbb{R}$. Let T be the random variable that denotes the earliest point in time $t \ge 0$ such that $X_t = 0$. If there exists $c > 0$ such that, for all $t < T$,*

$$\mathrm{E}[X_{t+1} - X_t \mid X_t] \le c,$$

then

$$\mathrm{E}[T \mid X_0] \ge \frac{X_0}{c}.$$

Theorem 25 (*Concentration for Additive Drift* [11]) *Let $(X_t)_{t\ge 0}$ be a sequence of random variables over $\mathbb{R}$, each with finite expectation and let $n > 0$. With $T = \min\{t \ge 0 : X_t \ge n \mid X_0 \ge 0\}$ we denote the random variable describing the*

earliest point that the random process exceeds n, given a starting value of at least 0. *Suppose there are* $\varepsilon, c > 0$ *such that, for all* $t < T$,

1. $\mathrm{E}[X_t - X_{t+1} \mid X_0, \ldots, X_t] \leq \varepsilon$, and
2. $|X_t - X_{t+1}| < c$.

Then, for all $s \leq n/(2\varepsilon)$,

$$\Pr(T < s) \leq \exp\left(-\frac{n^2}{8c^2 s}\right).$$

The following theorem is an adaptation of the negative drift theorem [18, 19] for large self-loop probabilities [20]. The theorem uses transition probabilities $p_{i,j}$ and the notation "$p_{k,k\pm d} \leq x$" as a shorthand for "$p_{k,k+d} \leq x$ and $p_{k,k-d} \leq x$".

Theorem 26 (Negative drift with self-loops [20]) *Consider a Markov process* $X_0, X_1, \ldots$ *on* $\{0, \ldots, m\}$ *with transition probabilities* $p_{i,j}$ *and suppose there exist integers* a, b *with* $0 < a < b \leq m$ *and* $\varepsilon > 0$ *such that for all* $a \leq k \leq b$ *the drift towards 0 is*

$$E(k - X_{t+1} \mid X_t = k) < -\varepsilon \cdot (1 - p_{k,k}) \tag{14}$$

where $p_{k,k}$ *is the self-loop probability at state k. Further assume there exist constants* $r, \delta > 0$ *(i. e. they are independent of m) such that for all* $k \geq 1$ *and all* $d \geq 1$

$$p_{k,k\pm d} \leq \frac{r(1 - p_{k,k})}{(1+\delta)^d}. \tag{15}$$

Let T be the first hitting time of a state at most a, starting from $X_0 \geq b$. *Let* $\ell = b - a$. *Then there is a constant* $c > 0$ *such that*

$$\Pr T \leq 2^{c\ell/r} = 2^{-\Omega(\ell/r)}.$$

References

1. Baillon, J.-B., Cominetti, R., Vaisman, J.: A sharp uniform bound for the distribution of sums of bernoulli trials. Combinat. Probab. Comput. **25**, 352–361 (2016)
2. Dang, D.-C., Lehre, P.K., Nguyen, P.T.H.: Level-based analysis of the univariate marginal distribution algorithm. Algorithmica **81**, 668–702 (2019)
3. Droste, S.: A rigorous analysis of the compact genetic algorithm for linear functions. Nat. Comput. **5**(3), 257–283 (2006)
4. Dubhashi, D.P., Panconesi, A.: Concentration of measure for the analysis of randomized algorithms. Cambridge University Press, Cambridge (2009)
5. Feller, W.: An Introduction to Probability Theory and Its Applications, vol. 1. Wiley, New York (1968)
6. Friedrich, T., Kötzing, T., Krejca, M.S., Sutton, A.M.: The compact genetic algorithm is efficient under extreme gaussian noise. IEEE Trans. Evolut. Comput. **21**(3), 477–490 (2017)

7. Grimmett, G., Stirzaker, D.: Probab. Random Process. Oxford University Press, Oxford (2001)
8. Harik, G.R., Cantú-Paz, E., Goldberg, D.E., Miller, B.L.: The gambler's ruin problem, genetic algorithms, and the sizing of populations. Evolut. Comput. **7**(3), 231–253 (1999)
9. Harik, G.R., Lobo, F.G., Goldberg, D.E.: The compact genetic algorithm. IEEE Trans. Evolut. Comput. **3**(4), 287–297 (1999)
10. He, J., Yao, X.: A study of drift analysis for estimating computation time of evolutionary algorithms. Nat. Comput. **3**(1), 21–35 (2004)
11. Kötzing, T.: Concentration of first hitting times under additive drift. Algorithmica **75**, 490–506 (2016)
12. Krejca, M.S., Witt, C.: Lower bounds on the run time of the univariate marginal distribution algorithm on OneMax. In: Proc. of FOGA '17, pp. 65–79. ACM Press, (2017)
13. Krejca, M.S., Witt, C.: Theory of estimation-of-distribution algorithms. In: Doerr, B., Neumann, F. (eds.) Theory of Evolutionary Computation: Recent Developments in Discrete Optimization, pp. 406–442. Springer, Berlin (2019)
14. Lehre, P.K., Nguyen, P.T.H.: Tight bounds on runtime of the univariate marginal distribution algorithm via anti-concentration. In: Proc. of GECCO '17, pp. 1383–1390. ACM Press, (2017)
15. Lengler, J., Sudholt, D., Witt, C.: Medium step sizes are harmful for the compact genetic algorithm. In: Proc. of GECCO '18, pp. 1499–1506. ACM Press, (2018)
16. Mühlenbein, H., Schlierkamp-Voosen, D.: Predictive models for the breeder genetic algorithm, I: continuous parameter optimization. Evolut. Comput. **1**(1), 25–49 (1993)
17. Neumann, F., Sudholt, D., Witt, C.: A few ants are enough: ACO with iteration-best update. In: Proc. of GECCO '10, pp 63–70. ACM Press, (2010)
18. Oliveto, P.S., Witt, C.: Simplified drift analysis for proving lower bounds in evolutionary computation. Algorithmica **59**(3), 369–386 (2011)
19. Oliveto, P.S., Witt, C.: Erratum: Simplified Drift Analysis for Proving Lower Bounds in Evolutionary Computation. ArXiv e-prints, (2012)
20. Rowe, J.E., Sudholt, D.: The choice of the offspring population size in the $(1,\lambda)$ evolutionary algorithm. Theor. Comp. Sci. **545**, 20–38 (2014)
21. Sudholt, D., Witt, C.: Update strength in EDAs and ACO: How to avoid genetic drift. In: Proc. of GECCO '16, pp. 61–68. ACM Press, (2016)
22. Sudholt, D., Witt, C.: On the choice of the update strength in estimation-of-distribution algorithms and ant colony optimization. Algorithmica **81**(4), 1450–1489 (2019)
23. Witt, C.: Upper bounds on the runtime of the Univariate Marginal Distribution Algorithm on OneMax. In: Proc. of GECCO '17, pp. 1415–1422. ACM Press, (2017)
24. Witt, C.: Upper bounds on the running time of the univariate marginal distribution algorithm on OneMax. Algorithmica **81**, 632–667 (2019)