UNIVERSITY OF DORTMUND

REIHE COMPUTATIONAL INTELLIGENCE

COLLABORATIVE RESEARCH CENTER 531

Design and Management of Complex Technical Processes
and Systems by means of Computational Intelligence Methods

Population Size vs. Runtime of a Simple Evolutionary Algorithm

Carsten Witt

No. CI-156/03

Population Size vs. Runtime of a Simple Evolutionary Algorithm

Carsten Witt[*]
FB Informatik, LS 2
Univ. Dortmund
44221 Dortmund, Germany
carsten.witt@uni-dortmund.de

October 8, 2003

Abstract

Evolutionary algorithms (EAs) find numerous applications, and practical knowledge on EAs is immense. In practice, sophisticated population-based EAs employing selection, mutation and crossover are applied. In contrast, theoretical analysis of EAs often concentrates on very simple algorithms like the (1+1) EA, where the population size equals 1. In this paper, the question is addressed whether the use of a population by itself can be advantageous. A population-based EA that does neither make use of crossover nor any diversity-maintaining operator is investigated on an example function. It is shown that an increase of the population size by a constant factor decreases the expected runtime from exponential to polynomial. Thereby, the so far best known gap is improved from super-polynomial vs. polynomial to exponential vs. polynomial. Moreover, it is proved that the exponential and polynomial runtime bounds occur with a probability exponentially close to one if the population size is a constant resp. a small polynomial. Finally, a second example function, where only a small population leads to a polynomial runtime, and a hierarchy result on the appropriate population size are presented.

1 Introduction

Evolutionary algorithms (EAs) are successfully applied in many areas, and experimental knowledge on EAs is immense. Yet, the theoretical knowledge on the design and analysis of EAs is far behind practical knowledge. A common approach to learn how evolutionary algorithms work is to analyze simple EAs for the maximization of functions. Here, we restrict ourselves to the case of pseudo-Boolean (fitness) functions $f: \{0,1\}^n \to \mathbb{R}$.

The simplest EA investigated so far is probably the well-known (1+1) EA (see, e.g., Droste, Jansen, and Wegener (2002); Garnier, Kallel, and Schoenauer (1999); He and Yao (2003); Wegener and Witt (2003)). It incorporates

---

a population of size 1 and, yet, is surprisingly efficient for many problems. We address the question when true populations, i.e., population sizes larger than 1, are helpful in EAs. Early attempts to find functions where a genetic algorithm (GA) with a true population and a crossover operator outperforms the (1+1) EA (in terms of the expected runtime) were unsuccessful (Mitchell, Forrest, and Holland (1992)). The same authors showed for their so-called royal road functions that even simple hillclimbers outperform the GA (Mitchell, Holland, and Forrest (1994)). Functions demonstrating the use of a crossover operator have been presented lately (Jansen and Wegener (2001b); Storch and Wegener (2003); see also Dietzfelbinger, Naudts, van Hoyweghen, and Wegener (2003)). However, this does not explain why a true population by itself can be advantageous. Therefore, we investigate EAs in which mutation is the only search operator and which do not employ diversity-maintaining mechanisms. For such EAs, we try to estimate the appropriate population size for example functions.

A similar approach has been described recently for $(1+\lambda)$ strategies by Jansen and De Jong (2002), who have shown that $\lambda = 1$ is an optimal choice for some well-studied functions. On the other hand, the authors have presented an example function where a $(1+\lambda)$ strategy with $\lambda > 1$ outperforms the (1+1) EA drastically. Yet, only offspring population size is considered in their paper. The influence of population size in an EA in which also a true parent population is maintained and which does not make use of crossover has been studied by Jansen and Wegener (2001b). The authors have proved rigorously that their EA outperforms the (1+1) EA on a specific function and have shown a runtime gap that is superpolynomial vs. polynomial. Further studies of the impact of population size on runtime have been presented by He and Yao (2002, 2003). However, the results obtained by the authors are based on EAs that employ mechanisms for maintaining diversity, single-individual EAs that use only a local search operator, or EAs that employ crossover. None of this is needed in the population-based EA that we will consider here. Instead, we widen the result by Jansen and Wegener (2001b) and exhibit a function where an EA with a large population size outperforms *the same* EA with a small population size (and also the (1+1) EA) and where the runtime gap is even exponential vs. polynomial. Moreover, the exponential and polynomial runtimes are proved to occur with probability exponentially close to 1, and the population size where the expected runtime changes from exponential to polynomial is asymptotically tight.

The paper is structured as follows. In Section 2, we define the EA and the function that we study in the forthcoming sections. Moreover, we supply some basic technical lemmas. In Section 3, we show an exponential lower bound on the expected runtime for population sizes of order $O(1)$. The probability of a polynomial runtime is exponentially small, and the expected runtime remains exponential if the population size is bounded by $cn^{1/2}$ for some small enough constant $c > 0$. In Section 4, we prove an $O(\mu n^{3/2} \log n)$ bound on the expected runtime for the same function when the population size μ is at least $3n^{1/2}$. Here, the probability of a runtime of $\omega(\mu n^{3/2} \log n)$ is exponentially small. The proof in Section 4 utilizes a theory on random trees, which is of interest on its own. To

2

make the paper self-contained, the required analysis of random trees is included in Appendix A.

In Section 5, we depict a reverse example, where a small population leads to a polynomial runtime whereas a large population leads to an exponential runtime. Finally, in Section 6, we develop a hierarchy result for a class of functions where the expected runtime of the GA changes from exponential to polynomial if its population size is of order $\Theta(n^{k-1/2})$ for an adjustable integer k. We finish with some conclusions.

2 Definitions

The population-based evolutionary algorithm that we study is quite simple. It can be considered as an elitist steady-state GA with population size μ employing fitness-proportional selection for reproduction, inversely proportional selection for deletion (meaning that worse individuals are more likely to be deleted) and standard bitwise mutation with probability $1/n$ like in the well-known (1+1) EA. Since these settings are common for genetic algorithms, we call our EA a Steady-State GA (or briefly *the GA*) even if it does not incorporate a crossover operator.

Definition 1 (The Steady-State GA)

For $i := 1$ To μ Do
  Choose $x_i \in \{0,1\}^n$ uniformly at random.
Repeat infinitely
  Choose $y \in \{x_1, \ldots, x_\mu\}$ such that
$$\text{Prob}(y = x_i) \;=\; f(x_i)/\textstyle\sum_{j=1}^{\mu} f(x_j).$$
  Create $x_{\mu+1}$ by flipping each bit of y
      independently with probability $1/n$.
  Rearrange $\{x_1, \ldots, x_{\mu+1}\}$ such that
      $f(x_1)$ is maximal and $f(x_{\mu+1})$ minimal.
  Choose $y \in \{x_2, \ldots, x_{\mu+1}\}$ such that
    $\text{Prob}(y = x_i) = \frac{f(x_1)+f(x_{\mu+1})-f(x_i)}{\sum_{j=2}^{\mu+1}(f(x_1)+f(x_{\mu+1})-f(x_j))}$ and delete it.

Since, in the deletion step, x_1 is always kept, the GA is elitist. We do not need to specify which of the best individuals (if there is more than one) is retained by the elitist strategy since our analyses will work for any choice of a best individual. If we set $\mu = 1$ and demand in the deletion step that the newly created individual is kept if its f-value equals the one of its father, we obtain the well-known (1+1) EA out of our GA. Moreover, our GA resembles the one considered by Jansen and Wegener (2001a). Yet, we do not introduce techniques to avoid duplicates or other mechanisms for maintaining diversity. We remark here that the theorems presented in this paper can also be proven for variants of our GA. For instance, all results remain valid if we replace the selection for replacement with a deterministic plus-selection. We stick to the fitness-proportional selection for deletion since we consider it to be more common for genetic algorithms.

The current runtime of the GA is measured by the sum of the initialization cost μ and the total number of iterations (also called *steps*) of the infinite loop up to now. Since only one evaluation of the objective function per iteration is necessary, the runtime corresponds also to the number of f-evaluations. This is a common approach in black-box optimization (see Droste, Jansen, Tinnefeld, and Wegener (2003)). However, one might not want to neglect the computational effort spent within an iteration. In this case, one may multiply the number of f-evaluations by the population size μ. Since only polynomial values of μ are considered, this does not influence the qualitative result of a polynomial-vs.-exponential gap.

One may object that the GA has been stated without a stopping criterion. Finding a reasonable stopping criterion is a problem that is relevant in practice. For our theoretical investigations, we consider the random time X_f until the current population of the GA contains at least one optimal individual for the first time, i. e., some $x \in \{0,1\}^n$ such that $f(x)$ is maximal. Then we say briefly that the GA has *reached the optimum*. We call X_f the runtime of the GA and study its expectation $\mathrm{E}(X_f)$ and the success probability $\mathrm{Prob}(X_f \leq t)$ for $t \geq 0$.

To prove the main result of this paper, we investigate the GA on a specially designed fitness function that exemplifies the use of populations and the influence of population size. The function is almost separable and can always be written as the sum of two functions, defined on disjoint subsets X_1 and X_2 of the whole variable set, respectively. These variables from X_1 and X_2 each have optimal assignments, which we call goals. Normally, after random initialization, the GA is able to search for the goals of X_1 and X_2 in parallel. However, as soon as search points that are optimal in one of the variable sets are found, the fitness landscape changes. We design the changes such that the goals play the following role.

If the second goal is reached before the first one, the GA is likely to get stuck in a local optimum. If the first one is reached before the second one, we have found a globally optimal search point. For small populations, the second goal is usually reached before the first one, leading to a large optimization time with high probability. If we increase the population size, the average progress to the second goal in a step becomes smaller whereas the progress towards the first goal is virtually independent of the population size. This holds since the progress to the first goal influences the f-value much stronger than the progress to the second one and since the GA chooses fitness-proportionally. If the population is large enough, the first goal is usually reached before the second one.

Let us make these ideas precise. In order to be able to show exponential bounds, we artificially scale the influence of one of the two variable sets exponentially. In practice, one may hope that the population size in GAs is relevant for functions that are in some respect similar to our function. Similar means that the function is separable with respect to two goals, one of which influences the function value stronger than the other one.

Now we define $f: \{0,1\}^n \to \mathbb{R}$. Let $\ell := \lceil n^{1/2}/400 \rceil$ and $m := n - \ell$. Moreover, let $\ell^* := \lceil n^{1/2}/(800 \cdot \log n) \rceil$. We divide strings (individuals) $x = (x_1, \ldots, x_n) \in \{0,1\}^n$ into their prefix $(x_1, \ldots, x_m)$ of length m and their suffix $(x_{m+1}, \ldots, x_n)$ of length ℓ. By $\mathrm{PO}(x) := \sum_{i=1}^{m} x_i$, we denote the sum of the

prefix bits, i.e., the number of **p**refix **o**nes. We want the suffix to be of shape $1^i0^{\ell-i}$ and count i, the number of its leading ones, then. Thus, we define $\mathrm{LSO}(x) := \sum_{i=0}^{\ell-1}\prod_{j=0}^{i} x_{m+1+j}$ as the number of **l**eading **s**uffix **o**nes. If the suffix of a string x is of the desired form $1^i0^{\ell-i}$, we say that x is *shapely*, and *malformed* otherwise. W.l.o.g., m is divisible by 3. Now let

$$f(x) := \begin{cases} \mathrm{PO}(x) + n^{m(\mathrm{LSO}(x)+1)} & \text{if } x \text{ shapely and } \mathrm{PO}(x) \leq \frac{2m}{3}, \\ n^{m\ell+\mathrm{PO}(x)} + \mathrm{LSO}(x) & \text{if } x \text{ shapely and } \frac{2m}{3} < \mathrm{PO}(x) \leq \frac{2m}{3} + \ell^*, \\ n - \sum_{i=1}^{n} x_i & \text{otherwise.} \end{cases}$$

We have to discuss the structure of f. First, observe that $f(x)$ is always positive. In all three cases, the function is separable with respect to the prefix and suffix bits. The first two cases apply to strings x whose suffix is shapely. Consider the first case. With respect to the prefix bits, we have the well-known OneMax function. With respect to the suffix bits, we have to maximize the number of leading 1-bits. Apart from the exponential scaling, the latter function is also known as the LeadingOnes function (see Droste, Jansen, and Wegener (2002)). Maximizing the number of leading suffix ones (LSOs) can be viewed as the above-mentioned first (and more important) goal whereas setting the number of prefix ones (POs) to $2m/3$ is the second goal. If the number of LSOs equals ℓ and $\mathrm{PO}(x) = 2m/3$, we obtain the maximum f-value of $2m/3+n^{m(\ell+1)}$.

In the second case of f's definition, the f-value is the sum of two terms that depend exponentially on the number of POs and linearly on the number of LSOs, respectively. Any shapely string x where $2m/3 < \mathrm{PO}(x) \leq 2m/3+\ell^*$ has a larger f-value than each x' where $\mathrm{PO}(x') \leq 2m/3$ and $\mathrm{LSO}(x') < \ell$, i.e., than each x' from the first case having a non-optimal suffix. If $\mathrm{PO}(x) = 2m/3 + \ell^*$ and $\mathrm{LSO}(x) = \ell$, we obtain a locally maximal string with second-best f-value $n^{m\ell+2m/3+\ell^*} + \ell$. However, a globally maximal string has $2m/3$ POs. Hence, such a string has Hamming distance $\ell^* = \Omega(n^{1/2}/\log n)$ to a second-best one, wherefore escaping from the local optimum typically takes a long time.

The remaining points, which lead to the case "otherwise", have more than $2m/3 + \ell^*$ POs or a malformed suffix. Their f-value is very low and increases with respect to the number of zeros in both the prefix and suffix bits. This is meant to help the GA to find a shapely string after random initialization. The f-value of strings with more than $2m/3 + \ell^*$ POs is so low to guarantee that the Hamming distance of a second-best and an optimal string is at most ℓ^* in the prefix bits, which will be relevant for upper bounds (Section 4). Finally, we remark that evaluation of $f(x)$ is possible in polynomial time for each $x \in \{0,1\}^n$.

If our GA is employed to maximize f, it reaches with overwhelming probability a population such that all individuals contain at most $2m/3$ POs, all individuals are shapely and no individual is optimal (we call such a population an *ordinary population*). Therefore, it is important to study the selection probabilities of the Steady-State GA for ordinary populations and the function f.

Concerning ordinary populations, it is crucial to note that an individual's number of POs virtually does not influence the selection probabilities of the GA. We denote by the random variable M for a population $X = (x_1, \dots, x_\mu)$

the index in $\{1, \ldots, \mu\}$ of the individual chosen for mutation (regarding the fitness function f). Likewise, we denote by D for an enlarged population $Y = (y_1, \ldots, y_{\mu+1})$ the index in $\{1, \ldots, \mu + 1\}$ of the individual chosen for deletion. We write $\mu = poly(n)$ to denote that μ is bounded by some polynomial of n. Since populations are unordered multisets, the following statements of Lemma 1 and Lemma 2 hold for arbitrarily assumed orders of the considered populations.

Lemma 1 *Let $X = (x_1, \ldots, x_\mu)$ be an ordinary population and x_i and x_j two individuals in X such that $\mathrm{LSO}(x_i) = \mathrm{LSO}(x_j)$. Then*

$$1 - n^{-\Omega(m)} \leq \frac{\mathrm{Prob}(M = j)}{\mathrm{Prob}(M = i)} \leq 1 + n^{-\Omega(m)}.$$

Proof: According to the definition of the Steady-State GA (Definition 1),

$$\frac{\mathrm{Prob}(M = j)}{\mathrm{Prob}(M = i)} = \frac{f(x_j)}{f(x_i)}.$$

Since X is ordinary, the f-value is the sum of two terms that depend linearly on the number of POs and exponentially on the number of LSOs. More precisely, we have $f(x_i) = \mathrm{PO}(x_i) + n^{m(\mathrm{LSO}(x_i)+1)}$ and $f(x_j) = \mathrm{PO}(x_j) + n^{m(\mathrm{LSO}(x_j)+1)} = \mathrm{PO}(x_j) + n^{m(\mathrm{LSO}(x_i)+1)}$. Hence, $f(x_i)/f(x_j) = 1 \pm n^{-\Omega(m)}$. $\square$

Moreover, we show that it is very unlikely to select individuals for mutation whose LSO-value is not maximal or to select individuals for deletion whose LSO-value is maximal.

Lemma 2 *Let $X = (x_1, \ldots, x_\mu)$ be an ordinary and $Y = (y_1, \ldots, y_{\mu+1})$ be an enlarged ordinary population. Let ℓ_X and ℓ_Y be the maximum LSO-value for the individuals from X and Y, respectively. Moreover, let there be at least one y_i in Y such that $\mathrm{LSO}(y_i) < \ell_Y$. If $\mu = poly(n)$ then $\mathrm{Prob}(\mathrm{LSO}(x_M) = \ell_X) = 1 - n^{-\Omega(m)}$. Moreover, $\mathrm{Prob}(\mathrm{LSO}(y_D) = \ell_Y) = O(n^{-m})$ for arbitrary μ.*

Proof: For the first statement, we refer to the GA's selection probability for reproduction (Definition 1). Obviously, the probability of choosing some individual with maximal number of LSOs becomes minimal if $\mu - 1$ individuals x_i fulfill $\mathrm{LSO}(x_i) = \ell_X - 1$ and $\mathrm{PO}(x_i) = 2m/3$. Hence,

$$\mathrm{Prob}(\mathrm{LSO}(x_M) = \ell_X) \geq \frac{n^{m(\ell_X+1)}}{poly(n) \cdot (2m/3 + n^{m\ell_X}) + n^{m(\ell_X+1)}} = 1 - n^{-\Omega(m)}.$$

Similarly, $\mathrm{Prob}(\mathrm{LSO}(Y_D) = \ell_Y)$ becomes maximal if Y contains μ individuals with LSO-value ℓ_Y. Consider the probability for deletion. Then, by counting in the nominator only the term for the worst individual, which has an LSO-value of at most $\ell_Y - 1$,

$$\mathrm{Prob}(\mathrm{LSO}(y_D) = \ell_Y) \leq \frac{n^{m\ell_Y} + 4m/3}{n^{m(\ell_Y+1)}} = O(n^{-m}).$$

$\square$

3 An Exponential Lower Bound for Small Populations

In this section, we show that the GA is very inefficient on the function f if too small populations are used.

Theorem 1 *With probability at least $1 - 2^{-\Omega(n^{1/2})}$, the Steady-State GA with $\mu = O(1)$ requires at least $2^{\Omega(n^{1/2})}$ steps to optimize f.*

Proof: The proof idea is that with high probability, the GA reaches a population containing at least one individual with $2m/3 + \ell^*$ POs (hereinafter called *prefix-optimal individual (POI)*) without ever generating ℓ leading ones in the suffix of an individual. This can be shown since optimizing the prefix takes an expected number of $O(n)$ steps if $\mu = O(1)$ whereas optimizing the suffix takes $\Omega(n\ell)$ expected steps in any case. Since each globally optimal individual has Hamming distance $\Omega(\ell^*) = \Omega(n^{1/2}/\log n)$ to a POI, reaching the optimum from a POI requires an exponential waiting time with high probability.

The formal proof uses the proof idea of identifying a *typical run* of the GA. We show that the probability of not observing a typical run is exponentially small, more precisely $2^{-\Omega(n^{1/2})}$. The typical run is divided into three epochs. The first epoch lasts until the population contains at least one shapely individual with at most $2m/3 + \ell^*$ POs, the second one ends when the current population contains a POI for the first time, and the third one is finished when the optimum is reached. Typically, the third epoch takes the long waiting time of $2^{\Omega(n^{1/2})}$. The run is not typical if the optimum is reached before the third epoch.

Consider the first epoch, which starts with initialization and is finished when a shapely individual with at most $2m/3 + \ell^*$ POs (i. e., not leading to the case "otherwise" in the definition of f) is created or the optimum is reached. We call the creation of an individual as described a success. Clearly, the fraction of shapely individuals among all individuals is $(\ell + 1)/2^\ell$. Since the initial individuals are drawn uniformly at random, the probability of a success in the initialization is at most $\mu(\ell + 1)2^{-\ell} = 2^{-\Omega(n^{1/2})}$ since $\mu = O(1)$. Hence, we assume the GA to start with μ malformed individuals, and the run is not typical otherwise. To estimate the time t^* until a success occurs, we consider the maximum f-value in the population as a potential. Due to the elitist selection mechanism of the GA, the potential cannot decrease. Moreover, the time until a success is bounded above by the time until the potential becomes at least n. Due to the definition of f in the case "otherwise", we can apply results on the analysis of linear functions (see Droste, Jansen, and Wegener (2002)). The expected time until reaching at least potential n is $O(n \log n)$ if we count only the steps choosing an individual with maximum f-value. Since such a step occurs with probability at least $1/\mu$ according to Definition 1 and since the mutation operator is independent of the selection operator, the expected time until reaching at least potential n is $O(\mu n \log n)$. By Markov's inequality, the time is bounded by $c\mu n \log n$ for some constant c with probability at least

1/2, and, by repeating independent trials, potential at least n is reached with probability at least $1 - 2^{-\ell}$ within at most $\ell c \mu n \log n = O(\mu \ell n \log n)$ steps.

We prove that the first shapely individual with at most $2m/3 + \ell^*$ POs (created at the random time t^*) is likely to have not too many LSOs. Consider the set S consisting of all malformed strings and all strings with more than $2m/3 + \ell^*$ POs, i.e., those strings leading to the case "otherwise" in the definition of f. Each initial individual of the GA is drawn uniformly at random from S if it is from S. We introduce the subset $S' \leq S$ consisting of all strings from S having at least $\lceil 2\ell/3 \rceil$ (not necessarily leading) ones in the suffix. By elementary counting arguments, $|S'|/|S| = 2^{-\Omega(\ell)}$. Since $\mu = O(1)$, this implies that the probability of choosing an initial individual x from S' is $2^{-\Omega(\ell)}$ even if we know that x is from S. Consider the random string x' created by mutation of a string x that was drawn uniformly at random from S. If we know that x' is also from S, the symmetry of the mutation operator implies that x' is uniform over S as well. In particular, it is from S' with probability $2^{-\Omega(\ell)}$.

However, if the GA creates a sequence of individuals from S starting with a population from S, the elements of this sequence are not distributed uniformly over S. Instead, the selection operators prefer individuals with less ones since the f-value increases if the number of ones of a string from S is decreased. Since $|S'| \leq |S|/2$ and S' contains all the strings from S with i suffix ones for any $i \in \{\lceil 2\ell/3 \rceil, \ldots, \ell\}$, an inductive argument yields that the probability of hitting a string from S' even decreases in the considered sequence of steps. Hence, any string created before the first one outside S contains at most $2\ell/3$ suffix ones with probability $1 - 2^{-\Omega(\ell)}$. The probability of flipping at least $\ell/12$ bits in one step is bounded above by $\binom{n}{\ell/12} n^{-\ell/12} \leq 1/(\ell/12)! = 2^{-\Omega(\ell \log \ell)}$ according to Stirling's formula (see Feller (1971)). We have $t^* = O(\mu \ell n \log n)$ with probability $1 - 2^{-\Omega(\ell)}$, and, therefore, t^* is a polynomial with this probability. Since flipping at least $\ell/12$ bits has probability $2^{-\Omega(\ell \log \ell)}$ in a polynomial number of steps, the shapely individual with at most $2m/3 + \ell^*$ POs created at time t^* has less than $2\ell/3 + \ell/12 \leq 3\ell/4$ LSOs with probability $1 - 2^{-\Omega(\ell)}$. In the following, we assume this to have happened. This completes the analysis of the first epoch.

In the beginning of the second epoch, we have at least one shapely individual with an f-value of at least n^m. Hence, by Definition 1, the probability of choosing a malformed individual for mutation is bounded by $n/n^m = n^{-\Omega(m)}$ now. For the second and also for the third epoch, we assume no malformed individual to be ever chosen. Now we divide the second epoch into phases of length s (which value is left open for the moment), and want the GA to increase the number of POs of some individual to at least $2m/3 + \ell^*$ in a phase. To estimate the time until this event, we introduce again a potential P, defined as the number of POs of an individual with maximum f-value in the current population. If the number of POs of two individuals differs, so does their f-value. Hence, P is well defined. Now it is crucial to note that due to the definition of f and elitist selection, the P-value cannot decrease in a step where the mutation does not increase the number of LSOs. We call steps increasing the number of LSOs *bad* since only they can lead to a decrease of P. A phase

of s steps is called bad if it contains at least one bad step, and good otherwise. In good phases, we can control the potential P. Hence, we want to find a good phase in which the GA increases P to $2m/3 + \ell^*$, i.e., creates a POI, without generating an optimal individual. Afterwards, we have to bound the overall number of bad steps in preceding bad phases to exclude the event of reaching the optimum. Let us assume for the moment that we have a good phase that starts with a non-optimal population. We claim that by its end, we have, with a probability of at least $1 - 2^{-\Omega(n)}$, at least one POI in the population. (This may finish the phase prematurely, i.e., lead to an actual phase length of less than s.)

In the good phase, we investigate so-called *helpful* steps increasing the P-value by 1, and we only have to consider P-values of at most $2m/3 + \ell^*$. Hence, the probability of a helpful step is at least the probability of flipping exactly one prefix zero in an individual with maximum f-value, i.e., at least

$$\frac{1}{\mu}\left(\frac{m}{3} - \ell^*\right)\frac{1}{n}\left(1 - \frac{1}{n}\right)^{n-1} \geq c,$$

for some constant $c > 0$ (recall that $m = n - o(n)$ and $\mu = O(1)$). Within $\lceil m/c \rceil$ steps, we expect at least m helpful steps, and by Chernoff bounds (see Motwani and Raghavan (1995)), with probability at least $1 - 2^{-\Omega(n)}$, the number of helpful steps is bounded below by $2m/3 + \ell^*$ if the phase is not finished prematurely. Hence, we set $s := \lceil m/c \rceil$. As suggested, this implies the considered good phase to end with at least one POI with probability $1 - 2^{-\Omega(n)}$.

We still have to take into account the bad steps. The probability of a bad step is bounded by $1/n$ since it is necessary that the leftmost suffix zero flips. Moreover, $s = O(n)$. Hence, the probability of a bad phase is at most $1 - (1 - 1/n)^s \leq c'$, for some constant $c' < 1$. The event of some phase being bad is independent of the event of some other phase being bad. Therefore, the probability of observing at least $\ell c/8$ bad phases in a row is at most $(c')^{\ell c/8} = 2^{-\Omega(\ell)}$. Moreover, the expected number of bad steps in a total number of

$$\left(\frac{\ell c}{8} - 1\right) \cdot s \leq \left(\frac{\ell c}{8} - 1\right)\left(\frac{m}{c} + 1\right) \leq \frac{\ell n}{8} + \frac{\ell c}{8}$$

steps is at most $(\ell n/8 + \ell c/8)/n = \ell/8 + o(1)$. Altogether, the probability of at least $\ell/6$ bad steps within the phases is $2^{-\Omega(\ell)}$ as well (by Chernoff bounds). Hence, we can work under the condition that we observe a total number of at most $\ell/6$ bad steps within at most $\ell c/8$ bad phases, which condition does not increase the probability that a step within these phases is bad. At most $\ell/6$ bad steps flip a total number of at most $\ell/5$ bits with probability at least $1 - 2^{-\Omega(\ell)}$ according to Chernoff bounds. Since we consider only shapely individuals with at most $3\ell/4$ LSOs, this implies that the number of LSOs of any individual does not exceed $3\ell/4 + \ell/5 < \ell$ with probability $1 - 2^{-\Omega(\ell)}$ until the good phase. So far, the probability of a typical run has been bounded by $1 - 2^{-\Omega(n^{1/2})}$, i.e., the third epoch is entered with the mentioned probability.

Throughout the third epoch, we have at least one POI in the population. The epoch is finished when the optimum is reached. We consider events that

are necessary to reach the optimum if we exclude steps that flip many bits. The probability of flipping at least $\ell^*/2$ bits in one step is bounded by $1/(\ell^*/2)! = 2^{-\Omega(n^{1/2})}$. Working under the assumption that a step flips at most $\ell^*/2$ bits, it is necessary that the optimum is reached by mutating some individual with at most $2m/3+\ell^*/2$ POs. However, the probability of choosing such an individual for mutation is bounded above by

$$\mu \cdot \frac{\ell + n^{m\ell+2m/3+\ell^*/2}}{n^{m\ell+2m/3+\ell^*}} = n^{-\Omega(\ell^*)} = 2^{-\Omega(n^{1/2})}$$

since the population contains at least one POI. (The probability of choosing an individual with at most $2m/3$ POs and less than ℓ LSOs is even smaller.) Altogether, the probability of observing in a step of the third epoch an event leading to optimization is bounded above by $2^{-\Omega(n^{1/2})}$. Finally, the probability of such an event within $2^{\varepsilon n^{1/2}}$ steps of the third epoch is bounded by $2^{-\Omega(n^{1/2})}$ if the constant ε is small enough. This completes the analysis of the third epoch and, since the probability of a typical run has, altogether, been bounded by $1 - 2^{-\Omega(n^{1/2})}$, also the proof. $\qquad\square$

Theorem 1 applies also to the simple (1+1) EA since it is contained in Definition 1 if $\mu = 1$. Besides, it implies that even multistart variants of the (1+1) EA and the GA with $\mu = O(1)$ fail with probability exponentially close to 1 within polynomially many steps if the number of instances is bounded by any polynomial. Moreover, we conjecture that more complicated GAs (even with a crossover operator) are likely to require an exponential time on f if their population is small, e. g., a constant.

If the population is larger than a constant, more precisely $\mu \leq cn^{1/2}$ for some constant c, we can still prove an exponentially large expected optimization time for our GA, albeit we can prove only an exponentially small probability for exponential runtimes.

Corollary 1 *There is a constant $c > 0$ such that the Steady-State GA with population size $\mu \leq cn^{1/2}$ needs an expected runtime of $2^{\Omega(n^{1/2})}$ to optimize f.*

Proof: We prove that the GA requires at least $t = 2^{\Omega(n^{1/2})}$ steps with a probability of at least $p = 2^{-O(\mu)}$ if c is small enough. Hence, t can be bounded below by $2^{c'n^{1/2}}$ for some constant $c' > 0$, and p can be bounded below by $2^{-c''\mu} = 2^{-c''cn^{1/2}}$ for some large enough constant $c'' > 0$. If we choose c small enough, the product of these lower bounds is still bounded below by $2^{\Omega(n^{1/2})}$. This will imply the corollary.

The proof of the open claim follows the same structure as the proof of Theorem 1 such that we only describe the places where different arguments are needed. The analysis of the first epoch and third epoch can be carried out in the very same way since for the related estimations, $\mu = poly(n)$ is sufficient. For the second epoch, the probability of a helpful step is now bounded below by c^*/μ, for some constant $c^* > 0$. In a phase of length $s := \lceil m\mu/c^* \rceil$, we expect at least $2m/3 + \ell^*$ increases of the P-value with probability at least $1 - 2^{-\Omega(n)}$. However, the probability of this single phase being good is only bounded below

by $(1 - 1/n)^s \geq e^{-2\mu/c^*}$. This implies that the second epoch, now consisting of only a single phase, ends with probability at least $e^{-2\mu/c^*}$ at a population with at least one POI and no individual with at least $3\ell/4$ LSOs. The third epoch lasts $2^{\Omega(n^{1/2})}$ steps with probability $1 - 2^{-\Omega(n^{1/2})}$. Altogether, this implies that the number of steps is at least $2^{\Omega(n^{1/2})}$ with a probability of at least $2^{-O(\mu)}$. $\quad\square$

4 A Polynomial Upper Bound for Large Populations

In view of the proof of Theorem 1, it seems important for the GA to maximize the number of LSOs before increasing the number of POs beyond $2m/3$. In fact, the use of a large population leads to a better exploitation of the primary goal of maximizing the number of LSOs.

Theorem 2 *With probability at least $1 - 2^{-\Omega(n^{1/2})}$, the Steady-State GA with $\mu \geq 3n^{1/2}$ and $\mu = poly(n)$ optimizes f in $O(\mu \ell n \log n)$ steps. Its expected runtime is bounded by $O(\mu \ell n \log n)$.*

In conjunction with Corollary 1, this means that $\mu = \lceil 3n^{1/2} \rceil$ seems to be the best choice. At least for $\mu = \omega(n^2 \log n)$, the initialization cost of the GA becomes larger than the upper runtime bound $O(n^2 \log n)$ that we obtain for $\mu = \lceil 3n^{1/2} \rceil$.

Proof: Observe that the initialization cost μ is always covered by the claimed runtime bounds. For the first statement of the theorem, the proof idea is again to identify a typical run of the GA, described as follows. Since the population size is large, we expect mutations that increase the number of POs to distribute among many individuals (as opposed to Theorem 1). This is due to the fact that individuals with maximal LSO-value are likely to produce copies, leading quickly to populations where many individuals have the same LSO-value. Conversely, the number of mutations necessary to reach maximal LSO-value remains approximately the same compared to Theorem 1. If each individual receives few mutations increasing its PO-value, it is likely that the maximal LSO-value is reached before the PO-value exceeds $2m/3$.

Formally, we consider the following typical run. The arguments for the analysis of the first epoch in the proof of Theorem 1 tell us that the GA creates at least one shapely individual after $O(\mu \ell n \log n)$ steps with a probability at least $1 - 2^{-3\ell}$. (This non-asymptotic estimation will matter latter.) Let t^* denote the first point of time with a shapely individual in the population. The prefix of an individual is uniformly distributed over $\{0, 1\}^m$ in the initialization, the f-value before time t^* increases as the number of POs decreases, and the prefix is irrelevant for an individual to be shapely. Therefore, an inductive argument yields that each prefix bit of an individual at time t^* contains a 1-entry with probability at most $1/2$. This implies by Chernoff bounds that each individual at time t^* (and, especially, the shapely one) has at most $7m/12$ POs with probability $1 - \mu 2^{-\Omega(n)} = 1 - 2^{-\Omega(n)}$ since $\mu = poly(n)$. We assume this property to hold at time t^*.

11

Starting at time t^*, we consider a predefined phase of length at most $s := \lceil 4e\ell n \rceil$, i.e., $s = \Theta(n^{3/2})$. To avoid confusion, we call this the *s-phase*. Our aim is to prove that the run of the GA complies with the following two properties with probability at least $1 - 2^{-\Omega(n^{1/2})}$.

- Within the s-phase, no individual has more than $2m/3$ POs,

- there is at least one step in the s-phase where an individual with ℓ LSOs is generated. The s-phase is finished after such a step.

We will see that an optimal individual is easy to find after the s-phase has been finished according to the properties.

To analyze the probability of the intersection of the two events, we first study the second property assuming that the first property holds. We introduce for the current population the potential L, describing the maximal number of LSOs for the population's individuals. Note that we assume all populations to contain at least one shapely individual. Furthermore, since we assume no individual of the population to have more than $2m/3$ POs, elitist selection guarantees that the L-value does not decrease. We call an individual whose number of LSOs equals the current L-value *maximal*. If the phase contains at least ℓ so-called good steps where a maximal individual (MI) is chosen for mutation and where merely the leftmost suffix zero flips, an individual with ℓ LSOs is created.

According to Lemma 2, the probability of choosing an MI for mutation is at least $1 - n^{-\Omega(m)}$ since $\mu = poly(n)$. The probability of flipping the leftmost suffix zero is at least $(1/n)(1 - 1/n)^{n-1} \geq 1/(en)$, and the probability of a good step is still $(1 - o(1))/(en)$. The assumption that the first property holds does not influence these probabilities since prefix and suffix bits are treated independently. We can ignore the event that the s-phase is finished prematurely, i.e., after less than s steps. By Chernoff bounds, the probability of at least ℓ good steps within $\lceil 4e\ell n \rceil$ steps is bounded by $1 - e^{-9\ell/4 - o(\ell)} = 1 - 2^{-\Omega(n^{1/2})}$. This completes the proof for the second property of the s-phase.

For the first property, we consider an arbitrary but fixed individual x from the population at time t^*. If x is mutated, a descendant of x is produced. More generally, we can visualize the descendants of x and their descendants by a family tree $T_t(x)$ at time t as follows. $T_{t^*}(x)$ contains only x. $T_t(x)$ contains $T_{t-1}(x)$ and the additional edge $\{v, w\}$ if w is the result of a mutation of the individual v at time $t-1$ and v is contained in $T_{t-1}(x)$. Note that the tree $T_t(x)$ may contain individuals that have already been deleted from the population at time t. We consider x as the root of $T_t(x)$ and are interested in the depth of $T_t(x)$ and, especially, of $T_{t^*+s}(x)$. For convenience, we say that x reaches *depth* j iff $T_t(x)$ has depth j at the considered time step t.

Now our goal is to show that no individual from the population at time t^* reaches a depth of more than $n/15$ within s steps with probability $1 - 2^{-\Omega(n)}$. This implies by Chernoff bounds that each descendant of any individual x at time t^* differs from x in at most $m/12$ bits with probability $1 - 2^{-\Omega(n)}$ since $n/15$ mutations flip at most $m/12$ bits with probability $1 - 2^{-\Omega(m)} = 1 - 2^{-\Omega(n)}$. Since each considered x contains at most $7m/12$ POs, flipping $m/12$ bits does not

suffice to increase the number of x's POs to more than $2m/3$ with probability $1 - 2^{-\Omega(n)}$. This will imply the first property.

To show the claim that depth $n/15$ is not exceeded, we consider the growth of the family tree $T_t(x)$ for any fixed x more carefully. In all following considerations, we can assume that we have ordinary populations since otherwise the first property is violated or we are done. According to Lemma 2, the probability of at least one deletion of an MI within s steps is bounded by $O(sn^{-m}) = 2^{-\Omega(n \log n)}$ if there is still a non-MI in the population, and the probability of not choosing an MI for mutation is bounded by $sn^{-\Omega(m)} = 2^{-\Omega(n \log n)}$. We assume these events not to occur, which implies that w.l.o.g., x is an MI.

Consider the epoch starting at t^* until the step that L increases for the first time. In this epoch, we ignore some unwanted steps, namely the steps choosing an MI for mutation and creating a non-MI by this mutation. By our assumptions, the resulting individual will never be chosen again. We call the corresponding node in a family tree a *dead node*. No other nodes are called dead even if they correspond to individuals that have already been deleted from the population. If we ignore only the dead nodes, we underestimate the depth of x's final family tree at most by 1. In the following, we will analyze the probability of choosing any other node in the current family tree. Let $u := 3n^{1/2}$, i.e., u is our lower bound on μ from the theorem.

According to the current population in the epoch, we distinguish two cases (which do not need to occur both). The first case is that the population contains at least one non-MI. Then it can happen that the probability of choosing a node from x's family tree is very large, e.g., even $1 - n^{-\Omega(m)}$. Consider a step in the first case and let v be the current size of the tree (ignoring dead nodes). Clearly, $v \leq s = poly(n)$. If the step does not choose a descendant of x then it neither deletes descendants of x from the population nor alters the family tree of x because the population contains a non-MI. If a descendant of x is chosen, the corresponding node is chosen from the tree almost uniformly with probability at most $(1 + n^{-\Omega(m)})/v$ due to Lemma 1 and the property $v = poly(n)$, and a new son is appended to it. All nodes of the resulting tree are MIs by our assumptions. If still the first case holds in the next step affecting the tree, we are in the analogous situation for $v + 1$ nodes. We conclude for the steps of the first case that the process growing x's family tree is very similar to a *random recursive tree*, see Definition 2 in Appendix A. The only exception is exponentially small deviations from uniformity of the nodes' selection probabilities.

Consider the second case of the epoch, when the population contains MIs only. By our assumptions, this case, once entered, is not left until the end of the epoch. Now descendants of x can be deleted from the population, and the number of nodes in x's family tree that correspond to individuals in the current population can even become zero. Nevertheless, we know that each node from the current tree is chosen with probability at most $(1 + n^{-\Omega(m)})/u \leq 2/u$ (for large enough n) since Lemma 1 can be applied to a population of size at least u and consisting of MIs only. Altogether, in the whole epoch, the probability of choosing a node from the current tree is bounded by $(1 + n^{-\Omega(m)})/v \leq 2/v$ (for large enough n), when v is the current size of the tree without dead nodes, or it is at most $2/u$. Once the latter case is valid, it holds until the end. We have given

the whole process growing the family tree a new name, namely $2/u$-*constrained RRT* (Definition 3), and Lemma 7 provides upper bounds on its depth. Note that the model of a $2/u$-constrained RRT counts only the relevant steps in the epoch, namely the steps adding non-dead nodes to the family tree.

Up to here, we have only modeled a single epoch until L is increased. All subsequent epochs between L-increasing steps can be modeled in the very same way if we consider the first individual with the new L-value as the new root of the family tree and ignore all nodes created before. This is possible since they will never be chosen again. (It is even possible that x's family tree cannot grow any more since the new L-value was achieved by a node belonging to a different family tree.) Altogether, the depth of $T_{t^*+s}(x)$ is bounded by the sum of the depths of the at most ℓ independent $2/u$-constrained RRTs created in the single epochs. Lemma 8 contains an estimate for this. We invoke it with a value of $t := s = \lceil 4e\ell n \rceil$ and $k := \ell = \lceil n^{1/2}/400 \rceil$ and obtain

$$\frac{7t}{u} + 6kH_u \leq \frac{7 \cdot \lceil 4e\ell n \rceil}{3n^{1/2}} + 6\ell(\ln u + 1) \leq \left(\frac{28e}{1200} + O(n^{-1/2}) \right) n,$$

which is smaller than $n/15 - 1$ (reserving 1 for the dead nodes) if n is large enough. Hence, using the notation of Lemma 8, $\mathrm{Prob}(D(k,t) \geq n/15 - 1) = 2^{-\Omega(t/u)+O(k \log t)} = 2^{-\Omega(n)}$, which completes our analysis of x's family tree. Since $\mu = poly(n)$, the probability of any family tree's depth exceeding $n/15$ is $2^{-\Omega(n)}$ as well. This completes the proof for the first property of the s-phase.

An individual with ℓ LSOs and at most $2m/3$ POs is nearly optimal. Elitist selection guarantees that the maximum number of POs in the population cannot decrease after the s-phase has been finished according to the described properties. Therefore, we consider the maximum f-value of the population again as a potential like in the first epoch. Given that we mutate an individual with maximum f-value, the probability of increasing the potential is bounded below by some constant since we have at least $m/3$ prefix zeros. Hence, a similar reasoning to the analysis of the good phase from the proof of Theorem 1 yields that the GA reaches the optimum after another $O(\mu n)$ steps with probability at least $1 - e^{-9\ell/4}$. For the ensuing considerations, we bound the sum of all considered failure probabilities more precisely by $2e^{-9\ell/4-o(\ell)} + 2^{-\Omega(n)} + 2^{-3\ell}$, which is at most $2^{-2\ell}$ if n is large enough. The total length of epochs and phases has been bounded by $O(\mu\ell n \log n)$. Altogether, we have proved the first statement of the theorem.

Finally, we have to prove the statement on the expected runtime of the GA. To this end, we have to consider any possible initial populations. If the initial population consists only of individuals covered by the third case of f's definition, we can apply again the results on the optimization of linear functions (cf. Theorem 1). Hence, whatever the initial population is like, we arrive at a population containing at least one shapely individual with at most $2m/3 + \ell^*$ POs after an expected number of $O(\mu n \log n)$ steps. Afterwards, we distinguish three cases with respect to the individual x^* with the maximum f-value for the current population. The first case occurs if $\mathrm{LSO}(x^*) = \ell$ and $\mathrm{PO}(x^*) \leq 2m/3$, the second case if $\mathrm{LSO}(x^*) < \ell$ and $\mathrm{PO}(x^*) \leq 2m/3$ and the third one if $\mathrm{PO}(x^*) > 2m/3$. In the first case, we apply the arguments from the preceding

paragraph to bound the expected time until reaching the optimum by $O(\mu n)$. In the second case, we use the analysis of the above-described second property of the s-phase and obtain that an expected time of $O(\ell n)$ is sufficient to leave the second case. Pessimistically, we assume to be in the third case afterwards. Then the f-value increases with respect to the number of LSOs and POs and is larger than the value of any individual of the second case. Hence, in the third case, elitist selection guarantees that we arrive at a population containing at least one locally optimal individual with $2m/3 + \ell^*$ POs and ℓ LSOs after $O(\mu n + \mu \ell n) = O(\mu \ell n)$ expected steps if the first case (including the optimum) is not reached before.

We have proved that, after $O(\mu \ell n)$ expected steps, the current population contains at least one locally optimal individual if we are not done before, and it is sufficient to create an optimal individual by flipping ℓ^* POs in a locally optimal individual. The probability of choosing a locally optimal individual for mutation and flipping ℓ^* POs is at least

$$\frac{1}{\mu} \cdot \left(\frac{1}{n}\right)^{\ell^*} \cdot \left(1 - \frac{1}{n}\right)^{n-\ell^*} = 2^{-(\log n) \cdot \ell^* - O(\log \mu)} \geq 2^{-\ell}$$

for n large enough since $\mu = poly(n)$. The expected time until such a mutation happens is, therefore, at most 2^ℓ if n is large enough. Hence, the expected time until creating an optimal individual is bounded above (for n large enough) by $2^\ell + O(\mu \ell n)$. As proved on the preceding pages, the runtime is $O(\mu \ell n \log n)$ with probability at least $1 - 2^{-2\ell}$ otherwise. The product of the failure probability's bound $2^{-2\ell}$ and the runtime bound $2^\ell + O(\mu \ell n)$ is $o(1)$. $\qquad \square$

We have seen that the proof of Theorem 2 makes heavy use of the properties of the selection for reproduction. It is crucial that individuals with larger f-value get a better chance of being mutated since, otherwise, the exploitation of the primary goal of maximizing the LSO-value would not benefit from the population. The properties of the selection for replacement are not that important. We only have to make sure that MIs are very unlikely to be deleted if there are still non-MIs and that the operator is elitist. Hence, the results from the previous sections would even hold for a deterministic plus-selection.

5 An Example with Opposite Results

In contrast to the results from the last sections, we are interested in an example where the use of a population is harmful, i.e., leads to an exponential runtime, whereas the (1+1) EA and GA with $\mu = O(1)$ are efficient. This can be proven for a function where the role of local and global optima has been exchanged compared to the function f. We reuse the notations from Section 2 here. For any constant $c \geq 1$, define

$$g_c(x) := \begin{cases} \text{PO}(x) + n^{m(\text{LSO}(x)+1)} & \text{if } x \text{ shapely and } \text{PO}(x) \leq \frac{2m}{3}, \\ n^{m\ell+\text{PO}(x)} + \text{LSO}(x) & \text{if } x \text{ shapely and } \frac{2m}{3} < \text{PO}(x) < \frac{2m}{3} + \frac{\ell^*}{c}, \\ n^{m(\ell+2)} & \text{if } x \text{ shapely and } \text{PO}(x) = \lceil \frac{2m}{3} + \frac{\ell^*}{c} \rceil, \\ n - \sum_{i=1}^n x_i & \text{otherwise.} \end{cases}$$

15

On malformed strings, f and g_c are identical. In the following, we only discuss shapely strings and assume $c = 1$ for convenience. The value of g_1 differs from the value of f only on strings x with $PO(x) = 2m/3 + \ell^*$. These yield the maximum g_1-value of $n^{m(\ell+2)}$. Strings x with $LSO(x) = \ell$ and $PO(x) \leq 2m/3$ lead to a g_1-value of $PO(x) + n^{m(\ell+1)}$, which is second-best if $PO(x) = 2m/3$. All remaining strings have some g_1-value that is by a factor of at least $\Omega(n^{m/3-\ell^*-1})$ smaller than $n^{m(\ell+1)}$.

If $c > 1$, the optimal number of POs changes from $2m/3 + \ell^*$ to $\lceil 2m/3 + \ell^*/c \rceil$, and the case of very bad strings with a g_c-value of at most n contains more members. This is the only difference compared to g_1. If we are in the first case of the definition of g_c, we obtain the same value as on f. Since c is assumed to be a constant, the preceding asymptotic statements from the case $c = 1$ remain valid. The Hamming distance of an optimal and a second-best string is at least $\ell^*/c = \Omega(n^{1/2}/\log n)$.

The properties of g and its similarities to f lead to the following theorems, whose proofs are based on those of Theorem 1 and Theorem 2.

Theorem 3 *Let $c \geq 1$ be a constant. With probability at least $1 - 2^{-\Omega(n^{1/2})}$, the Steady-State GA with $\mu \geq 3n^{1/2}$ and $\mu = poly(n)$ requires at least $2^{\Omega(n \log n)}$ steps to optimize g_c.*

Proof: We call individuals x where $LSO(x) = \ell$ and $PO(x) \leq 2m/3$ *suffix-optimal*. According to the proof of Theorem 2, the GA reaches a population containing at least one suffix-optimal individual before reaching the optimum (at $PO(x) = \lceil 2m/3 + \ell^*/c \rceil$) with probability at least $1 - 2^{-\Omega(n^{1/2})}$. Now we can apply an argument similar to the analysis of the second epoch from the proof of Theorem 1.

The probability of creating an optimal individual by the direct mutation of a suffix-optimal one is bounded by $1/(\ell^*)! = 2^{-\Omega(\ell)}$. The probability of choosing an individual that is worse than a suffix-optimal one for mutation is even bounded by

$$\mu \cdot \frac{\ell + n^{m\ell + 2m/3 + \ell^* - 1}}{n^{m\ell + m}} = n^{-\Omega(m)} = 2^{-\Omega(n \log n)},$$

implying a waiting time of $2^{\Omega(n \log n)}$ with probability at least $1 - 2^{-\Omega(n^{1/2})}$. $\square$

Theorem 4 *Let $c \geq 1$ be a constant. With probability at least $1 - 2^{-\Omega(n^{1/2})}$, the Steady-State GA with $\mu = O(1)$ optimizes g_c within $O(n^{3/2})$ steps. If c is large enough, its expected runtime is also bounded by $O(n^{3/2})$.*

Proof: We apply the following arguments from the proof of Theorem 1. With probability $1 - 2^{-\Omega(n^{1/2})}$, the first shapely individual created by the GA has at most $3\ell/4$ LSOs, and such a population is reached after an expected number of $O(\mu n \log n) = O(n \log n)$ steps. Afterwards, since $c \geq 1$, it arrives at a population containing at least one shapely individual with $\lceil 2m/3 + \ell^*/c \rceil$ POs after at most $O(\ell)$ phases of length $O(n)$, i.e., after $O(n^{3/2})$ steps, with probability $1 - 2^{-\Omega(n^{1/2})}$. This completes the proof of the first statement.

For the statement on the expected runtime, note that a second-best individual with $2m/3$ POs and ℓ LSOs is created after an expected number of $O(n^{3/2})$ steps if no shapely individual with more POs is created before. This follows by taking into account the time until reaching a shapely individual and summing up the expected times of $O(\ell n)$ to maximize the number of LSOs and of $O(n)$ to increase the number of POs to $2m/3$ (cf. the proof of Theorem 2). Afterwards, it is sufficient to select an individual with maximum g_c-value, to flip $\lceil \ell^*/c \rceil$ prefix zeros and to leave the suffix unchanged in order to reach the optimum. The related probability is bounded below by

$$\frac{1}{\mu}\left(\frac{1}{n}\right)^{\ell^*/c+1}\left(1-\frac{1}{n}\right)^{n-\ell^*/c} = 2^{-\ell^*(\log n)/c - O(\log n)} = 2^{-n^{1/2}/(2c) - O(\log n)},$$

and the expected time till this event, therefore, by $2^{n^{1/2}/(2c)+O(\log n)}$. If we choose c large enough, the product of the failure probability's bound $2^{-\Omega(n^{1/2})}$ and the described upper bound is $o(1)$. $\qquad\square$

Remark: One may wonder if the expected runtime on g_c can remain a polynomial if μ is larger than a constant. However, we believe that this is not the case by having again a look at the proof of Theorem 2. The bad steps mentioned there affect individuals that are drawn nearly uniformly from the population. Moreover, since the influence of the number of POs is so small, the number of POs is close to uniform in those individuals whose number of POs differs from P. Hence, the probability of a bad step resetting P to a value around $m/2$ seems to converge to one if μ is greater than a constant. This would imply that the probability of reaching a second-best individual and getting caught in a local optimum would be become too large for the expected runtime to be a polynomial.

6 A Hierarchy Result

For the function f described in Section 2, we have shown that a population size $\mu \leq cn^{1/2}$, c some constant, leads to an exponential expected runtime whereas $\mu \geq 3n^{1/2}$ leads to a polynomial expected runtime. This means that we have determined the asymptotically exact threshold size $\Theta(n^{1/2})$ where the expected runtime switches from exponential to polynomial.

It is possible to extend the described result to a hierarchy result where the threshold size of the population is adjustable, namely $\Theta(n^{k-1/2})$ for an arbitrary constant integer $k \geq 1$. To accomplish this, we design a class of functions f_k, $k \geq 1$, that serve as a generalization of f. The main difference is that it takes on average $\Theta(n^k)$ steps until a success in the suffix bits occurs, which is is achieved by considering leading 1-blocks instead of leading ones there. We reuse the notations from Section 2 and assume $\ell = k \cdot \ell_k$ for the sake of simplicity. Now we divide the suffix of strings into ℓ_k consecutive blocks of length k each. For a string $x \in \{0,1\}^n$, we say that x is *shapely* if its suffix is of shape $1^{ik}0^{\ell-ik}$ for some $i \geq 0$, i.e., the first i blocks consist of ones only and the remaining entries of zeros, and it is called *malformed* otherwise. The

number of leading 1-blocks in the suffix (**l**eading **s**uffix **b**locks of ones) is given by $\mathrm{LSB}_k(x) := \sum_{i=1}^{\ell_k} \prod_{j=0}^{ik-1} x_{m+1+j}$. Note that for $k = 1$, the definition of shapeliness equals the one from Section 2, and $\mathrm{LSB}_1(x) = \mathrm{LSO}(x)$. Now let

$$f_k(x) := \begin{cases} \mathrm{PO}(x) + n^{m(\mathrm{LSB}_k(x)+1)} & \text{if } x \text{ shapely and } \mathrm{PO}(x) \le \frac{2m}{3}, \\ n^{m\ell_k + \mathrm{PO}(x)} + \mathrm{LSB}_k(x) & \text{if } x \text{ shapely and } \frac{2m}{3} < \mathrm{PO}(x) \le \frac{2m}{3} + \ell^*, \\ n - \sum_{i=1}^{n} x_i & \text{otherwise.} \end{cases}$$

Clearly, $f_1(x) = f(x)$. Consider the case $k > 1$ and assume that x is a shapely string with at most $2m/3$ POs. If $\mathrm{LSB}_k(x) < \ell_k$ then at least k suffix zeros have to be changed to increase the f_k-value. The structure of local and global optima is the same as for the function f. Therefore, we are able to show straightforward generalizations of Theorem 1, Corollary 1 and Theorem 2. Moreover, it is obvious that Lemma 1 and Lemma 2 carry over to f_k if we consider the LSB_k-value instead of the LSO-value.

Theorem 5 *Let $k \ge 1$ be a constant. With probability at least $1 - 2^{-\Omega(n^{1/2})}$, the Steady-State GA with $\mu = O(1)$ requires at least $2^{\Omega(n^{1/2})}$ steps to optimize f_k. There is a constant $c > 0$ such that the GA with $\mu \le cn^{k-1/2}$ needs an expected runtime of $2^{\Omega(n^{1/2})}$.*

Proof: For the first statement, the proof of Theorem 1 can be copied with the following changes. The fraction of shapely individuals among all individuals is only $(\ell_k + 1)/2^\ell \le (\ell + 1)/2^\ell$. The probability of a bad step is bounded by $1/n^k \le 1/n$ since the k leftmost suffix zeros have to flip. In the third epoch, the probability of choosing an individual with at most $2m/3 + \ell^*/2$ POs for mutation is bounded by

$$\mu \cdot \frac{\ell_k + n^{m\ell_k + 2m/3 + \ell^*/2}}{n^{m\ell_k + 2m/3 + \ell^*}} = n^{-\Omega(\ell^*)},$$

which is again $2^{-\Omega(n^{1/2})}$. This completes the proof of the first statement.

For the second statement, we apply an argument for a single phase of the second epoch as in the proof of Corollary 1. The probability of a helpful step is bounded below by c^*/μ for some constant $c^* > 0$. Since the probability of a bad step is at most $1/n^k$, the probability of a single phase of length $s := \lceil m\mu/c^* \rceil$ being good is bounded below by $(1 - 1/n^k)^s \ge \mathrm{e}^{-2\mu/(c^* n^{k-1})}$. Since $\mu \le cn^{k-1/2}$, this lower bound is at least $\mathrm{e}^{-2cn^{1/2}/c^*}$. If c is chosen small enough, the product of the last expression and the runtime of $2^{\Omega(n^{1/2})}$ that is necessary after the good phase with probability $1 - 2^{-\Omega(n^{1/2})}$ is still bounded below by $2^{\Omega(n^{1/2})}$. $\square$

Theorem 6 *Let $k \ge 1$ be a constant. With probability at least $1 - 2^{-\Omega(n^{1/2})}$, the Steady-State GA with $\mu \ge 3n^{k-1/2}$ and $\mu = poly(n)$ optimizes f_k within $O(\mu \ell n^k \log n)$ steps. Its expected runtime is bounded by $O(\mu \ell n^k \log n)$.*

Proof: The proof of Theorem 2 can be copied with the following changes. The length of the s-phase starting at time t^* is set to $s := \lceil 4e\ell n^k \rceil$. The potential L corresponds to the LSB_k-value rather than the LSO-value, and all applications

of Lemma 1 and Lemma 2 refer to the LSB_k-value. The probability of a step increasing the L-value is now bounded below by $(1/n^k)(1-1/n)^{n-k} \geq 1/(\mathrm{e}n^k)$, and the probability of a good step, therefore, at least $(1-o(1))/(\mathrm{e}n^k)$. We estimate the largest possible L-value by ℓ. By Chernoff bounds, an individual with maximal L-value is created within the s-phase with probability at least $1 - \mathrm{e}^{-9\ell/4-o(\ell)}$ again.

To show that depth $n/15$ of a family tree is not exceeded, we carry out the same arguments as in the proof of Theorem 2 and apply Lemma 8. The only new parameters are $t = s = \lceil 4\mathrm{e}\ell n^k \rceil$ and $u = 3n^{k-1/2}$. Hence, ignoring ceils and floors, $7t/u$ is upper bounded by the very same value as in the proof of Theorem 2, and $7t/u + 6kH_u$ is still at most $n/15 - 1$ (for n large enough) since u is a polynomial of n.

In the proof of the expected runtime, nearly no changes are required. The expected time until leaving the so-called second case, meaning that a best individual has less than ℓ LSOs and at most $2m/3$ POs, can now be bounded by $O(\ell n^k)$, and, using analogous arguments, the expected time until reaching a locally optimal individual by $O(\mu\ell n^k)$. Since k is a constant, the product of this upper bound and $2^{-2\ell}$ is still $o(1)$. □

7 Conclusions

We have shown that a population-based EA without recombination and without diversity-maintaining operators can outperform the (1+1) EA and the same population-based EA with a small population drastically. We have proved an exponential gap for the expected runtime on an explicitly defined function. The derived runtime bounds hold with probability exponentially close to 1, ruling out efficient optimization by multistart variants in the case of small populations, and the population size where the expected runtime changes from polynomial to exponential has been determined asymptotically tight. Moreover, we have presented a reverse result where only small populations allow efficient optimization. Finally, we have described a hierarchy result, where populations of size at most $cn^{k-1/2}$, $k \geq 1$ and c some small constant, lead to exponential runtimes whereas populations of size at least $3n^{k-1/2}$ allow polynomial runtimes. Our results help to understand the role of population size in EAs.

Some interesting questions have been left open. The results presented in this paper carry over to some variants of the studied GA. Regarding the selection for replacement, the results would even hold for a deterministic plus-selection. However, all results rely on the fitness-proportional selection-for-reproduction mechanism of the GA. The impact of different operators for the selection for reproduction, e.g., of a uniform selection such as in $(\mu+1)$ strategies, deserves further theoretical investigations.

A   Random Trees

Random recursive trees are a model of random trees studied extensively in probability theory (Pittel (1994); Smythe and Mahmoud (1995)). We obtain a random recursive tree of any desired size by means of the following stochastic process.

Definition 2 (Random Recursive Tree (RRT)) *An RRT at time 0 consists only of the root node. An RRT T_t at time $t \geq 1$ is obtained from an RRT T_{t-1} at time $t-1$ by choosing uniformly at random some node v from T_{t-1} and adding a new leaf as the son of v.*

Note that the RRT at time $t \geq 0$ consists of exactly $t + 1$ nodes. The RRT from Definition 2 is sometimes (and more precisely) called the *uniform random recursive tree.* We omit the adjective "uniform" here. Many (extremal) properties of RRTs are well known. For instance, Pittel (1994) shows the expected depth of an RRT with t nodes to equal $e \ln t$, and there are exact formulas for the expected number of leaves, expected outdegree of nodes etc.

RRTs can help to model the so-called family tree of an individual x in populations that contain more and more descendants of x. However, our populations are bounded by some maximum size $u \geq 1$. If we keep all individuals in the family tree of x forever, there must be nodes in the tree that correspond to individuals that have already been deleted from the population. If the selection mechanism of a population-based EA chooses all individuals present in the population uniformly with probability $1/u$, we can sometimes still conclude that each node from the family tree is chosen as the father of a new node with probability at most $1/u$. To allow for deviations from uniformity, we upper bound the actual probabilities by $2/(t + 1)$ and $2/u$. This leads to the definition of so-called $2/u$-constrained random recursive trees ($2/u$-constrained RRTs).

Definition 3 (p-RRT, $2/u$-constrained RRT) *Let $p := (p_t)_{t \geq 0}$ be a sequence of probability distributions on $\{0, \ldots, t\}$. A p-RRT at time 0 consists only of the root. A p-RRT T_t at time $t \geq 1$ is obtained from T_{t-1} by sampling, according to p_{t-1}, some $t^* \in \{0, \ldots, t-1\}$ and appending a new leaf to the node inserted at time t^*.*

A p-RRT is called a $2/u$-constrained RRT if there is an integer $u \geq 1$ such that p_t assigns to each element of $\{0, \ldots, t\}$ a probability of at most $2/(t+1)$ if $t \leq u - 1$ and at most $2/u$ otherwise.

Again, the tree at time t possesses $t + 1$ nodes. Obviously, more than one stochastic process leads to $2/u$-constrained RRTs. Up to time u, the process of a $2/u$-constrained RRT is very similar to the process of an ordinary RRT in terms of Definition 2. The only exception is that the father of the new node does not need to be chosen uniformly but with at most twice the corresponding probability. However, after time u, we allow much larger probabilities than in ordinary RRTs. Despite this, we are still able to provide useful estimates for the distribution of the depth of $2/u$-constrained RRTs. The corresponding analyses

rely on the following random variables, whose notation has been borrowed from Arya, Golin and Mehlhorn (1999).

Definition 4 *Let $L(t, d)$ be the number of nodes at depth d in a $2/u$-constrained RRT at time t. Let $E(t, d) := \mathrm{E}(L(t, d))$ denote its expectation.*

First of all, we consider the case $t \leq u$, i.e., the $2/u$-constrained RRT is similar to an ordinary RRT. The expected depth of ordinary RRTs can be determined exactly using analyses of Poisson processes (Pittel (1994)). However, to derive tail bounds on the depth for our variant of RRTs using estimates for $E(t, d)$, we adopt a combinatorial approach presented by Arya, Golin and Mehlhorn (1999), being more amenable to the reader familiar with asymptotic analysis. Note that the calculations in the proof of Lemma 4 mainly repeat the mentioned authors' analyses for the special case of a random circuit of fan-in 1.

Lemma 3 *Consider a $2/u$-constrained RRT. Then*

$$
\begin{aligned}
E(t, 0) &= 1 && \text{for } t \geq 1, \\
E(0, d) &= 0 && \text{for } d \geq 1, \\
E(t, d) &\leq E(t - 1, d) + \tfrac{2E(t-1, d-1)}{t} && \text{for } 1 \leq t \leq u,\ d \geq 1.
\end{aligned}
$$

Proof: The first and second relation hold by Definition 3. For the third relation, observe that $L(t, d)$ differs from $L(t - 1, d)$ if and only if a node from depth $d-1$ is chosen at time $t-1$ to be the father of the newly inserted node. If $L(t-1, d-1)$ equals i, this happens, according to Definition 3, with probability at most $2i/t$ since $t \leq u$. Hence, by the law of total probability,

$$
E(t, d) \leq E(t - 1, d) + \sum_{i=1}^{\infty} \mathrm{Prob}(L(t - 1, d - 1) = i) \cdot \frac{2i}{t}.
$$

Since the sum contains the definition of $E(t - 1, d - 1)$, the right-hand side equals $E(t - 1, d) + 2E(t - 1, d - 1)/t$ as suggested. $\qquad\square$

Using these relations, we are able to derive bounds on $E(t, d)$. Throughout the section, we write $H_t = \sum_{i=1}^{t} 1/i$ to denote the t-th Harmonic number.

Lemma 4 *For $0 \leq t \leq u$ and $d \geq 0$,*

$$
E(t, d) \leq \frac{(2H_t)^d}{d!}.
$$

Proof: We assume that the third relation from Lemma 3 is an equality and holds even for $t > u$. Upper bounds on this modified $E(t, d)$ will yield upper bounds on the original $E(t, d)$ for $t \leq u$. Now we introduce the generating function $E_d(x) = \sum_{t=0}^{\infty} E(t, d)x^t$ for the (modified) $E(t, d)$. Then

$$
E_0(x) = 1 + x + x^2 + \cdots = \frac{1}{1 - x},
$$

$$
E_d(x) = \frac{2}{1 - x} \int_0^x E_{d-1}(y)\, \mathrm{d}y \quad \text{for } d \geq 1. \tag{1}
$$

The first equality follows immediately by the identity $E(t, 0) = 1$ from Lemma 3. To prove the second equality (1), we use the second identity of Lemma 3 to show that $E_d(x) = \sum_{t=1}^{\infty} E(t, d)x^t$ if $d \geq 1$. Hence, for $d \geq 1$, the third relation (assumed to hold with equality) of Lemma 3 yields

$$
\begin{aligned}
E_d(x) &= \sum_{t=1}^{\infty} E(t, d)x^t = \sum_{t=1}^{\infty} \left(E(t-1, d) + \frac{2E(t-1, d-1)}{t} \right) \cdot x^t \\
&= x \cdot \sum_{t=0}^{\infty} E(t, d)x^t + 2 \sum_{t=0}^{\infty} E(t, d-1)\frac{x^{t+1}}{t+1} \\
&= x \cdot E_d(x) + 2 \sum_{t=0}^{\infty} E(t, d-1) \int_0^x y^t \, \mathrm{d}y = x \cdot E_d(x) + 2 \int_0^x E_{d-1}(y) \, \mathrm{d}y.
\end{aligned}
$$

The last equality follows by the Monotone Convergence Theorem (cf. Kingman and Taylor (1966)). Subtracting $x \cdot E_d(x)$ and dividing by $1-x$ finally proves (1).

Now we prove by induction on d that

$$
E_d(x) = \frac{(-2\ln(1-x))^d}{d!(1-x)}.
$$

The base case $d = 0$ is obvious. For the induction step, we use (1) and the induction hypothesis, writing

$$
\begin{aligned}
E_d(x) &= \frac{2}{1-x} \int_0^x E_{d-1}(y) \, \mathrm{d}y = \frac{2}{1-x} \cdot \int_0^x \frac{(-2\ln(1-y))^{d-1}}{(d-1)!(1-y)} \, \mathrm{d}y \\
&= \frac{2^d}{(d-1)!(1-x)} \cdot \left[\frac{(-\ln(1-y))^d}{d} \right]_0^x = \frac{(-2\ln(1-x))^d}{d!(1-x)}.
\end{aligned}
$$

Using this closed formula, we want to extract $E(t, d)$, the coefficient of x^t in the series expansion of $E_d(x)$. To do this, we introduce the Taylor expansion $-\ln(1-x) = \sum_{i=1}^{\infty} x^i/i$ and calculate

$$
(-\ln(1-x))^d = \left(\sum_{i=1}^{\infty} \frac{x^i}{i} \right)^d = \sum_{i=1}^{\infty} \left(\sum_{\substack{i_1,\ldots,i_d \geq 1 \\ i_1+\cdots+i_d=i}} \frac{1}{i_1 i_2 \cdots i_d} \right) x^i.
$$

The last equality follows by convolution of the series. This implies that the factor $(-\ln(1-x))^d/(1-x) = (-\ln(1-x))^d \cdot \sum_{i=0}^{\infty} x^i$ (for $|x| < 1$) from the above closed formula can be written as

$$
\left(\sum_{i=1}^{\infty} \left(\sum_{\substack{i_1,\ldots,i_d \geq 1 \\ i_1+\cdots+i_d=i}} \frac{1}{i_1 i_2 \cdots i_d} \right) x^i \right) \left(\sum_{j=0}^{\infty} x^j \right) = \sum_{i=1}^{\infty} \left(\sum_{j=0}^{i} \sum_{\substack{j_1,\ldots,j_d \geq 1 \\ j_1+\cdots+j_d=j}} \frac{1}{j_1 j_2 \cdots j_d} \right) x^i,
$$

where the last equality follows again by convolution. Therefore, $E(t, d)$, the coefficient of x^t in $E_d(x)$, is given by

$$
E(t, d) = \frac{2^d}{d!} \cdot \sum_{\substack{i_1,\ldots,i_d \geq 1 \\ i_1+\cdots+i_d \leq t}} \frac{1}{i_1 i_2 \cdots i_d} \leq \frac{2^d}{d!} \cdot \sum_{1 \leq i_1,\ldots,i_d \leq t} \frac{1}{i_1 i_2 \cdots i_d} = \frac{2^d}{d!} \left(\sum_{i=1}^{t} \frac{1}{i} \right)^d.
$$

The last expression equals $\frac{(2H_t)^d}{d!}$. Since our assumptions were valid for $t \le u$, so is the obtained upper bound. $\qquad \square$

Now we come to the times $t \ge u+1$, when the probability of choosing a node from a $2/u$-constrained RRT is only bounded by $2/u$.

Lemma 5 *Consider a $2/u$-constrained RRT. If $t \ge u+1$ and $d \ge 1$,*

$$E(t,d) \le E(t-1,d) + \frac{2E(t-1,d-1)}{u}.$$

Proof: This proof follows the same structure as the proof of the third relation of Lemma 3. Since $t \ge u+1$, the probability of choosing some father at time $t-1$ is now bounded by $2/u$ according to Definition 3. $\qquad \square$

Now we are able to combine the results obtained so far in order to provide an estimate for $E(t,d)$ that is valid for all t.

Lemma 6 *Consider a $2/u$-constrained RRT. Then*

$$E(t,d) \le \tfrac{1}{d!} \cdot \left(2H_u + \tfrac{2t}{u}\right)^d \quad for \ t \ge 0, \ d \ge 0,$$

where $0! = 1$.

Proof: We prove the inequality by induction on d and t. First consider the case $d = 0$. It follows by Lemma 3 since the upper bound equals 1 in this case.

For the induction step from $d-1$ to $d \ge 1$, we have to consider all values of t and to carry out another induction on t. For $t \le u$, the inequality follows from Lemma 4. For the induction step from $t-1$ to t, we assume $t \ge u+1$ and use the recurrence from Lemma 5. We obtain, by the induction hypothesis for $t-1$,

$$
\begin{aligned}
E(t,d) &\le \frac{1}{d!}\left(2H_u + \frac{2t-2}{u}\right)^d + \frac{2}{u} \cdot \frac{1}{(d-1)!}\left(2H_u + \frac{2t-2}{u}\right)^{d-1} \\
&\le \frac{1}{d!}\left(\left(2H_u + \frac{2t-2}{u}\right)^d + \frac{2d}{u}\left(2H_u + \frac{2t-2}{u}\right)^{d-1}\right) \\
&\le \frac{1}{d!}\left(2H_u + \frac{2t}{u}\right)^d.
\end{aligned}
$$

The last inequality follows by means of the Binomial Theorem if we expand $2H_u + 2t/u$ as $(2H_u + (2t-2)/u) + 2/u$ and consider only the first two terms. This completes the induction step for t and, therefore, for d. $\qquad \square$

The result of Lemma 6 can easily be transferred into a tail bound on the distribution of the depth.

Lemma 7 *Let $D(t)$ be the depth of a $2/u$-constrained RRT at time t and let $d \ge 6t/u + 6H_u$. Then $\mathrm{Prob}(D(t) \ge d) \le \left(\frac{e}{3}\right)^d$.*

Proof: W. l. o. g., d is an integer. For $D(t) \geq d$ to occur, $L(t, d) \geq 1$ is necessary (and sufficient). Since, by Markov's inequality, $\text{Prob}(L(t, d) \geq 1) \leq E(t, d)$, the lemma follows if we establish the inequality $E(t, d) \leq (\frac{e}{3})^d$ for arbitrary t.

Application of Stirling's formula on the inequality of Lemma 6 yields

$$E(t, d) \leq \frac{e^d}{d^d} \left(2H_u + \frac{2t}{u} \right)^d.$$

According to our assumptions, this is upper bounded by

$$\frac{e^d}{(6H_u + 6t/u)^d} \cdot \left(2H_u + \frac{2t}{u} \right)^d \leq \left(\frac{e}{3} \right)^d$$

as suggested. □

In our applications, we are confronted with random tress that can be modeled as a concatenation of $2/u$-constrained RRTs. Therefore, we consider k independent $2/u$-constrained RRTs, where the i-th RRT, $i \in \{1, \ldots, k\}$, possesses t_i nodes, and introduce the random variables D_i for the height of the i-th tree. Let $D(k, t) := \sum_{i=1}^{k} D_i$ be the sum of their heights, and let $t := \sum_{i=1}^{k} t_i$ be the total number of nodes. Furthermore, assume that the assumption $t^k = 2^{o(t/u)}$ is met. Then it is not too difficult to show that $E(D(k, t)) = O(t/u + kH_u)$. However, we want to show that this upper bound holds with high probability. Since we can bound the expectation of each D_i, we could apply Hoeffding bounds to show that the bound on $D(k, t)$ holds with probability $1 - 2^{-\Omega((t/u)^{1/2})}$. But we can do better by taking into account the knowledge of the D_i's distribution from Lemma 7.

Lemma 8 *Let $D(k, t)$ be the sum of the heights of k independent $2/u$-constrained RRTs with a total number of t nodes. Then*

$$\text{Prob}(D(k, t) \geq 7t/u + 6kH_u) \leq \left(\frac{e}{3} \right)^{t/u - O(k \log t)} = 2^{-\Omega(t/u) + O(k \log t)}.$$

Proof: Denote the number of nodes of the i-th RRT, $i \in \{1, \ldots, k\}$, by t_i and consider its depth D_i. If we choose $d_i := 6(t_i - 1)/u + 6H_u$ (taking into account that t_i nodes are present at time $t_i - 1$), Lemma 7 yields for all $d \geq d_i$

$$\text{Prob}(D_i \geq d) \leq \left(\frac{e}{3} \right)^d.$$

For each fixed outcome of the D_i-values such that $D(k, t) \geq 7t/u + 6kH_u$, we will bound the probability of this outcome by $2^{-\Omega(t/u) + O(k \log t)}$. Afterwards, we will estimate the number of such outcomes to prove the lemma.

Let $(e_1, \ldots, e_k)$ be an outcome where $e_1 + \cdots + e_k \geq 7t/u + 6kH_u$. Let k^* be the number of e_i where $e_i \geq d_i$ and assume w. l. o. g. that $e_i \geq d_i$ for $i \in \{1, \ldots, k^*\}$. According to the definition of the d_i and e_i, we have $e_{k^*+1} + \cdots + e_k \leq 6t/u + 6kH_u$, implying $k^* \geq 1$ and $e_1 + \cdots + e_{k^*} \geq t/u$. Due to the independence of the RRTs, the probability of this event is at most

$$\left(\frac{e}{3} \right)^{e_1} \cdot \left(\frac{e}{3} \right)^{e_2} \cdots \left(\frac{e}{3} \right)^{e_{k^*}} = \left(\frac{e}{3} \right)^{e_1 + \cdots + e_{k^*}} \leq \left(\frac{e}{3} \right)^{t/u},$$

which bounds the probability of each considered outcome of the e_i-values.

It remains to estimate the number of outcomes. Obviously, $D(k, t)$ can take at most t values. A trivial estimation yields at most t^k different ways of representing a single value of $D(k, t)$ as a sum of k non-negative integers and, altogether, we have at most t^{k+1} outcomes. Hence, the total probability of the event $D(k, t) \geq 7t/u + 6kH_u$ is at most

$$t^{k+1} \cdot \left(\frac{e}{3}\right)^{t/u} = \left(\frac{e}{3}\right)^{t/u - O(k \log t)},$$

which is $2^{-\Omega(t/u) + O(k \log t)}$ since $e/3 < 1$. $\qquad\square$

We could improve the upper bound of Lemma 8 by a factor of $2^{-\Omega(k \log k)}$ if we estimated the number of outcomes in its proof more carefully. However, k is too small in our applications for this to make an asymptotic difference.

References

Arya, S., Golin, M. J., and Mehlhorn, K. (1999). On the expected depth of random circuits. *Combinatorics, Probability and Computing*, **8**, 209–228.

Dietzfelbinger, M., Naudts, B., van Hoyweghen, C., and Wegener, I. (2003). The analysis of a recombinative hill-climber on H-IFF. To appear in IEEE Transactions on Evolutionary Computation.

Droste, S., Jansen, T., Tinnefeld, K., and Wegener, I. (2003). A new framework for the valuation of algorithms for black-box optimization. In *Proc. of Foundations of Genetic Algorithms 7 (FOGA 2002)*, 253–270. Morgan Kaufmann.

Droste, S., Jansen, T., and Wegener, I. (2002). On the analysis of the (1+1) evolutionary algorithm. *Theoretical Computer Science*, **276**, 51–81.

Feller, W. (1971). *An Introduction to Probability Theory and its Applications*. Wiley, New York.

Garnier, J., Kallel, L., and Schoenauer, M. (1999). Rigorous hitting times for binary mutations. *Evolutionary Computation*, **7**(2), 173–203.

He, J. and Yao, X. (2002). From an individual to a population: An analysis of the first hitting time of population-based evolutionary algorithms. *IEEE Transactions on Evolutionary Computation*, **6**(5), 495–511.

He, J. and Yao, X. (2003). Towards an analytic framework for analysing the computation time of evolutionary algorithms. *Artificial Intelligence*, **145**(1–2), 59–97.

Jansen, T. and De Jong, K. (2002). An analysis of the role of offspring population size in EAs. In *Proc. of GECCO 2002, Genetic and Evolutionary Computation Conference*, 238–246.

Jansen, T. and Wegener, I. (2001a). On the utility of populations. In *Proc. of GECCO 2001, Genetic and Evolutionary Computation Conference*, 1034–1041.

Jansen, T. and Wegener, I. (2001b). Real royal road functions – where crossover provably is essential. In *Proc. of GECCO 2001, Genetic and Evolutionary Computation Conference*, 375–382.

Kingman, J. F. C. and Taylor, S. J. (1966). *Introduction to Measure and Probability.* Cambridge University Press.

Mitchell, M., Forrest, S., and Holland, J. H. (1992). The royal road for genetic algorithms: Fitness landscapes and GA performance. In Varela, F. J. and Bourgine, P. (eds.), *Towards a Practice of Autonomous Systems: Proceedings of the First European Conference on Artificial Life*, 245–254. MIT Press, Paris.

Mitchell, M., Holland, J. H., and Forrest, S. (1994). When will a genetic algorithm outperform hill climbing. In Cowan, J. D., Tesauro, G., and Alspector, J. (eds.), *Advances in Neural Information Processing Systems*, vol. 6, 51–58. Morgan Kaufmann.

Motwani, R. and Raghavan, P. (1995). *Randomized Algorithms.* Cambridge University Press.

Pittel, B. (1994). Note on the heights of random recursive trees and random m-ary search trees. *Random Structures and Algorithms*, **5**(2), 337–348.

Smythe, R. T. and Mahmoud, H. M. (1995). A survey of recursive trees. *Theory of Probability and Mathematical Statistics*, **51**, 1–27.

Storch, T. and Wegener, I. (2003). Real royal road functions for constant population size. In *Proc. of GECCO 2003, Genetic and Evolutionary Computation Conference*, no. 2724 in LNCS, 1406–1417.

Wegener, I. and Witt, C. (2003). On the optimization of monotone polynomials by the (1+1) EA and randomized local search. In *Proc. of GECCO 2003, Genetic and Evolutionary Computation Conference*, no. 2723 in LNCS, 622–633.