



Deposited via The University of Sheffield.

White Rose Research Online URL for this paper:

<https://eprints.whiterose.ac.uk/id/eprint/120164/>

Version: Submitted Version

Proceedings Paper:

Corus, D., Dang, D.C., Ereemeev, A.V. et al. (2014) Level-Based Analysis of Genetic Algorithms and Other Search Processes. In: Parallel Problem Solving from Nature – PPSN XIII. The 13th PPSN: International Conference on Parallel Problem Solving from Nature, September 13-17, 2014, Ljubljana, Slovenia. Lecture Notes in Computer Science (8672). Springer, pp. 912-921. ISSN: 0302-9743.

https://doi.org/10.1007/978-3-319-10762-2_90

Reuse

Items deposited in White Rose Research Online are protected by copyright, with all rights reserved unless indicated otherwise. They may be downloaded and/or printed for private study, or other acts as permitted by national copyright laws. The publisher or other rights holders may allow further reproduction and re-use of the full text version. This is indicated by the licence information on the White Rose Research Online record for the item.

Takedown

If you consider content in White Rose Research Online to be in breach of UK law, please notify us by emailing eprints@whiterose.ac.uk including the URL of the record and the reason for the withdrawal request.

Level-Based Analysis of Genetic Algorithms and Other Search Processes

Dogan Corus^{*1}, Duc-Cuong Dang², Anton V. Eremeev³, and Per Kristian Lehre²

¹University of Sheffield, United Kingdom

²University of Nottingham, United Kingdom

³Omsk Branch of Sobolev Institute of Mathematics, Russia

October 28, 2016

Abstract

Understanding how the time-complexity of evolutionary algorithms (EAs) depend on their parameter settings and characteristics of fitness landscapes is a fundamental problem in evolutionary computation. Most rigorous results were derived using a handful of key analytic techniques, including drift analysis. However, since few of these techniques apply effortlessly to population-based EAs, most time-complexity results concern simplified EAs, such as the (1+1) EA.

This paper describes the *level-based theorem*, a new technique tailored to population-based processes. It applies to any non-elitist process where offspring are sampled independently from a distribution depending only on the current population. Given conditions on this distribution, our technique provides upper bounds on the expected time until the process reaches a target state.

We demonstrate the technique on several pseudo-Boolean functions, the sorting problem, and approximation of optimal solutions in combinatorial optimisation. The conditions of the theorem are often straightforward to verify, even for Genetic Algorithms and Estimation of Distribution Algorithms which were considered highly non-trivial to analyse. Finally, we prove that the theorem is nearly optimal for the processes considered. Given the information the theorem requires about the process, a much tighter bound cannot be proved.

1 Introduction

The theoretical understanding of Evolutionary Algorithms (EAs) has advanced significantly over the last decade. A contributing factor for this success may have been the strategy to analyse simple settings before proceeding to more complex scenarios, while at the same time developing appropriate analytic techniques. In

^{*}Dogan Corus' contributions to this paper was made while he was a PhD student at the University of Nottingham.

particular, much of the work assumed a population size of one, and no crossover operator. Current approaches to analysing evolutionary algorithms often rely on one or more of these simplifying assumptions.

This paper presents a general-purpose technique to analyse a large class of search heuristics involving non-overlapping populations. In our framework, each individual of the current population is independently sampled from the same distribution over the search space parametrised by the previous generation. A similar modelling of the search process first appeared in [56] to analyse Genetic Algorithms (GAs) however as far as we know, mainly results at the limit of infinite population were established. In this paper, we give the following general result for finite populations. Given some requirement on the upper tails of this distribution over an ordered partition of the search space and a minimum requirement on the population size, our method will guarantee an upper bound on the expected runtime to reach the last set of the partition.

Particularly, the partition of the search space is similar to the well-known fitness-level technique [58] to analyse *elitist* EAs, however at our general level of describing the search process, the traditional requirement on a *fitness-based* (this will be properly defined later on) partition is no longer required. Applications of the fitness-level technique itself are widely known in the literature for classical *elitist* EAs [58]. One of the first examples of using this technique in the analysis of non-elitist EAs is [24] where lower and upper bounds on the expected proportions of the population above certain fitness levels were found.

Related to our work, early research on analysing population-based EAs often ignored recombination operators. The family tree technique was introduced in [59] to analyse the $(\mu+1)$ EA. The performance of the $(\mu+\mu)$ EA for different settings of the population size was conducted in [33] using Markov chains to model the search processes, and in [5] using a similar argument to fitness-levels. The analysis of parallel EAs in [38] also made use of the fitness-levels argument. The inefficiency of standard fitness proportionate selection without scaling was shown in [46] and in [39] using drift analysis [30]. In the recently introduced switch analysis, the progress of the EA is analysed relative to an easier understand reference process [60]. When the method applies, bounds on the runtime of the reference process can be translated into bounds on the original process. In current applications of this method, the reference process is RLS^\square , a simple local search algorithm. It remains to be seen how such simple search heuristics can approximate the population dynamics of complex EAs.

Over the recent years, runtime analysis of EAs with recombination, often referred to as Genetic Algorithms, has been subject to increasing interest. Generalising the work in [46], [48, 49] showed that the Simple Genetic Algorithm [56] is inefficient on ONEMAX, even when crossover is used. A long sequence of work has attempted to show that enabling crossover can reduce the runtime. It has been shown that adding crossover to the $(\mu+1)$ EA can decrease the runtime on the JUMP problem, however only for small crossover probabilities [34, 36]. For realistic crossover probabilities, it was shown that $(\mu+1)$ GA can decrease the runtime by an exponential factor on instances of an FSM testing problem, however this result assumes a deterministic crowding diversity mechanism [41]. With the same setting on the standard ONEMAX function, crossover was shown to lead to a constant speedup in [54], however this result assumed a tailored selection mechanism. Seeking the construction of an efficient *unbiased* algorithms for ONEMAX, [20] introduced the $(1+(\lambda,\lambda))$ GA and showed a significant speed

up with the right choices of the offspring population size [18,19]. Another modified GA, but this time *non-elitist*, was introduced in [51], and its efficiency was proved on the noisy version of ONEMAX function. In [43], a runtime result is proposed for a class of convex search algorithms, including some non-elitist GAs with gene pool recombination and no mutation, on the so-called *quasi-concave fitness landscapes*. As a corollary, it has been shown that the convex search algorithm has $\mathcal{O}(n \log n)$ expected runtime on LEADINGONES. Those results gave the impression that adjustments or modifications to the standard setting of GAs, here elitist, are often required to illustrate the advantage of crossover. Until recently, it has been shown that the standard $(\mu+1)$ GA without too low crossover probability has a speed up of $\Omega(n/\log(n))$ on the JUMP problem compared to mutation-only algorithms [12].

Significant progress in developing and understanding a formal model of canonical GA and its generalisations was made in [56] using dynamical systems. In particular it turned out that the behaviour of the dynamical systems model is closely related to the local optima structure of the problem in the case of binary search spaces [57]. Most of the findings in [56,57] apply to the infinite population case, so it is not clear how these results can be used in runtime analysis of EA.

A relatively new paradigm in Evolutionary Computation is Estimation of Distribution Algorithm (EDA) [37]. Unlike traditional EAs which use explicit genetic operators such as mutation, recombination and selection, an EDA builds a probabilistic model for sampling new search points so that the probability of creating an optimal solution via sampling eventually increases high. The algorithm often starts with a specific probabilistic model, which is gradually updated through selected solutions of intermediate samplings. Over the recent years, many variants of EDAs have been proposed, along with theoretical investigations on their convergence and scalability, e.g. [29,45,50,53,61]. However, rigorous runtime analysis results for this particular class of algorithms on discrete domain are still sparse. The first analysis of this kind was conducted in [21] for the compact Genetic Algorithm (cGA) [31] on linear functions. Further work showed that this algorithm can be resilient to noise [27].

Another simple EDA is the Univariate Marginal Distribution Algorithm (UMDA) which was proposed in [44], and analysed in a series of papers [6–9]. With the n -dimensional Hamming cube as search space, each generation of UMDA consists of first sampling a population of solutions based on a vector $(p_i)_{i \in [n]}$ of frequencies, i.e. assuming independence between bit positions, then summarising the selected solutions as the new sampling vector for the next generation. The initial result of [7] was provided for LEADINGONES and a harder function known as TRAPLEADINGONES under the so-called “no-random-error” assumption and with a sufficiently large population. The assumption was lifted due to the technique presented in [9]. Nevertheless, the analysis assumes an unrealistically large population size, leading in overall to a too high bound on the expected runtime. Note also that there are two versions of the algorithm based on whether or not margins are imposed to p_i , the difference between the two in terms of time complexity for various functions are discussed in [8]. More interestingly, [6] showed that UMDA without margins beats the $(1+1)$ EA on a particular function called SUBSTRING. However, it is not recommended to use UMDA without margins in practice, as the algorithm can always end up with a premature convergence.

In this paper, we show that all non-elitist EAs with or without crossover, and even UMDA can be cast and analysed in the same framework. A preliminary version of the paper was communicated in [10]. This followed the line of work dated back to the introduction of a fitness-level technique to analyse *non-elitist* EAs with linear ranking selection [42], later on generalised to many selection mechanisms and *unary variation operators* [39], with a refined result in [16]. The original fitness-level technique and its generalisation to the level-based technique have already found a number of applications, including analysis of EAs in uncertain environments, such as partial information [16], noisy fitness functions [14], and dynamic fitness functions [13]. It has also been applied to analyse the runtime of complex algorithms, such as GAs for shortest paths [11], EDAs [15], and self-adaptive EAs [17].

The present work improves the main result of [10] in many aspects. A more careful analysis of the population dynamics leads to a much tighter expression of the runtime bound compared to [10], immediately implying improved results in the previously mentioned applications. In particular, the leading term in the runtime is improved by a factor of $\Omega(\delta^{-3})$, where δ characterises how fast good individuals can populate the population. This significantly improves the results of [14] and [16] concerning noisy optimisation, for which δ is often very small (e.g. $1/n$). We also provide guideline how to use the theorem to analyse the runtime of non-elitist processes. Selected examples are given for the cases of GAs and UMDA in optimising standard pseudo-Boolean functions, a simple combinatorial problem, and in searching for local optima of NP-hard problems. Furthermore, we prove that the level-based theorem is close to optimal for the class of evolutionary processes it applies to.

The paper is structured as follows. In Section 2, we first present the general scheme of the algorithms covered by the main result of the paper and show how the GAs fit as special cases into this scheme. The main result of the paper is then presented along with a set of corollaries tailored to specific cases. Applications of the main result to different GAs are considered in Section 4. The section starts with runtime analysis of the Simple Genetic Algorithm on standard functions, followed by the results for combinatorial optimisation problems, finally, the main theorem is again applied to analyse an Estimation of Distribution Algorithm. Section 6 considers the tightness of the level-based theorem. Finally, concluding remarks are given in Section 7.

2 Main result

2.1 Abstract algorithmic scheme

We consider population-based algorithms at a very abstract level in which fitness evaluations, selection and variation operations, which depending on the current population P of size λ , are represented by a distribution $D(P)$ over a finite set \mathcal{X} . More precisely, the current population P is a vector $(P(1), \dots, P(\lambda))$ where $P(i) \in \mathcal{X}$ for each $i \in [\lambda]$. D is a mapping from \mathcal{X}^λ into the space of probability distributions over \mathcal{X} . The next generation is obtained by sampling each new individual independently from $D(P)$. This scheme is summarised in Algorithm 1. Here and below, for any positive integer n , we define $[n] := \{1, 2, \dots, n\}$.

Algorithm 1 Population-based algorithm.

Require:

- Finite state space \mathcal{X} , and population size $\lambda \in \mathbb{N}$,
 - Mapping D from \mathcal{X}^λ to the space of prob. dist. over \mathcal{X} .
 - Initial population $P_0 \in \mathcal{X}^\lambda$.
 - 1: **for** $t = 0, 1, 2, \dots$ until termination condition met **do**
 - 2: Sample $P_{t+1}(i) \sim D(P_t)$ independently
 for each $i \in [\lambda]$
 - 3: **end for**
-

A scheme similar to Algorithm 1 was studied in [56], where it was called *Random Heuristic Search with an admissible transition rule*. Some examples of such algorithms are Simulated Annealing (more generally any algorithm with the population composed of a single individual), Stochastic Beam Search [56], Estimation of Distribution Algorithms such as the Univariate Marginal Distribution Algorithm [6] and the Genetic Algorithm [28]. The previous studies of the framework were often limited to some restricted settings [47] or mainly focused on infinite populations [56]. In this paper, we are interested in finite populations and develop a general method to deduce the expected runtime of the search processes defined in terms of *number of produced search points*. This can be translated to the *number of evaluations* once a specific algorithm is instantiated and the optimisation scenario is specified (e.g. see [16]).

We illustrate the general scheme of Algorithm 1 on the example of GA, which is Algorithm 2. The term Genetic Algorithm is often applied to EAs that use

Algorithm 2 Genetic algorithm.

Require:

- Finite state space \mathcal{X} ,
 - Operators: Sel, Cross and Mut,
 - Population size $\lambda \in \mathbb{N}$ and recombination rate $p_c \in [0, 1]$.
 - 1: $P_0 \sim \text{Unif}(\mathcal{X}^\lambda)$
 - 2: **for** $t = 0, 1, 2, \dots$ until termination condition met **do**
 - 3: **for** $i = 1$ to λ **do**
 - 4: $u := P_t(\text{Sel}(P_t)), v := P_t(\text{Sel}(P_t))$.
 - 5: $x := \begin{cases} \text{Cross}(u, v) & \text{with prob. } p_c, \\ u & \text{with prob. } 1 - p_c. \end{cases}$
 - 6: $P_{t+1}(i) := \text{Mut}(x)$.
 - 7: **end for**
 - 8: **end for**
-

recombination operators with some a priori chosen probability $p_c > 0$. Here the standard operators of GA are formally represented by transition matrices:

- $p_{\text{sel}} : [\lambda] \times \mathcal{X}^\lambda \rightarrow [0, 1]$ represents selection operator Sel: $\mathcal{X}^\lambda \rightarrow [\lambda]$ which is randomised, where $p_{\text{sel}}(i|P_t)$ is the probability of selecting the i -th individual from population P_t . This probability can depend on the search point $P_t(i)$, its relationship to the other search points in P_t , and their mappings to fitness values by a function $f: \mathcal{X} \rightarrow \mathbb{R}$ that the algorithm aims

to optimise. Throughout the paper, we assume w.l.o.g. the maximisation of f .

- $p_{\text{mut}} : \mathcal{X} \times \mathcal{X} \rightarrow [0, 1]$, where $p_{\text{mut}}(y|x)$ is the probability of mutating $x \in \mathcal{X}$ into $y \in \mathcal{X}$ by a randomised mutation operator $\text{Mut} : \mathcal{X} \rightarrow \mathcal{X}$.
- $p_{\text{xor}} : \mathcal{X} \times \mathcal{X}^2 \rightarrow [0, 1]$, where $p_{\text{xor}}(x|u, v)$ is the probability of obtaining x as a result of randomised crossover operator (or recombination) between $u, v \in \mathcal{X}$. In what follows, crossover is denoted by $\text{Cross} : \mathcal{X} \times \mathcal{X} \rightarrow \mathcal{X}$.

Clearly, conditioned on the current population P_t , each individual $P_{t+1}(i)$ of the next generation is independently sampled from the same distribution which is parametrised by P_t . Thus, lines 4-6 of Algorithm 2 can be summarised as $P_{t+1}(i) \sim D(P_t)$ for some D induced by the genetic operators Sel , Cross and Mut . The algorithm fits perfectly in the scheme of Algorithm 1.

2.2 Level-based theorem

This section states the main result of the paper, a general technique for obtaining upper bounds on the expected runtime of any process that can be described in the form of Algorithm 1. We use the following notation. The natural logarithm is denoted by $\ln(\cdot)$. Suppose that for some m there is an ordered partition of \mathcal{X} into subsets (A_1, \dots, A_m) called *levels*, we define $A_{\geq j} := \cup_{i=j}^m A_i$, i.e. the union of all levels above level j . An example of a partition is the *canonical* partition, where each level regroups solutions having the same fitness value (see e.g. [39]). This partition is classified as *fitness-based* or *f-based*, if $f(x) < f(y)$ for all $x \in A_j, y \in A_{j+1}$ and all $j \in [m-1]$. As a result of the algorithmic abstraction, our main theorem is not limited to this particular type of partition. Let $P \in \mathcal{X}^\lambda$ be a population vector of a finite number $\lambda \in \mathbb{N}$ of individuals. Given any subset $A \subseteq \mathcal{X}$, we write $|P \cap A| := |\{i \mid P(i) \in A\}|$ to denote the number of individuals in population P that belong to the subset A .

Theorem 1. *Given a partition (A_1, \dots, A_m) of \mathcal{X} , define $T := \min\{t\lambda \mid |P_t \cap A_m| > 0\}$ to be the first point in time that elements of A_m appear in P_t of Algorithm 1. If there exist $z_1, \dots, z_{m-1}, \delta \in (0, 1]$, and $\gamma_0 \in (0, 1)$ such that for any population $P \in \mathcal{X}^\lambda$,*

(G1) *for each level $j \in [m-1]$, if $|P \cap A_{\geq j}| \geq \gamma_0 \lambda$ then*

$$\Pr_{y \sim D(P)}(y \in A_{\geq j+1}) \geq z_j,$$

(G2) *for each level $j \in [m-2]$, and all $\gamma \in (0, \gamma_0]$ if $|P \cap A_{\geq j}| \geq \gamma_0 \lambda$ and $|P \cap A_{\geq j+1}| \geq \gamma \lambda$ then*

$$\Pr_{y \sim D(P)}(y \in A_{\geq j+1}) \geq (1 + \delta)\gamma,$$

(G3) *and the population size $\lambda \in \mathbb{N}$ satisfies*

$$\lambda \geq \left(\frac{4}{\gamma_0 \delta^2} \right) \ln \left(\frac{128m}{z_* \delta^2} \right) \text{ where } z_* := \min_{j \in [m-1]} \{z_j\},$$

then

$$E[T] \leq \left(\frac{8}{\delta^2}\right) \sum_{j=1}^{m-1} \left(\lambda \ln\left(\frac{6\delta\lambda}{4+z_j\delta\lambda}\right) + \frac{1}{z_j}\right).$$

Informally, the two first conditions require a relationship between the current population P and the distribution $D(P)$ of the individuals in the next generation: Condition (G1) demands that the probability of creating an individual at level $j+1$ or higher is at least z_j when some fixed portion γ_0 of the population has reached level j or higher. Furthermore, if the number of individuals at level $j+1$ or higher is at least $\gamma\lambda > 0$, condition (G2) requires that their number tends to increase further, e.g. by a multiplicative factor of $1+\delta$. Finally, (G3) requires a sufficiently large population size. When all conditions are satisfied, an upper bound on the expected time for the algorithm to create an individual in A_m can be guaranteed.

We suggest to follow the five steps below when applying the level-based theorem.

1. Identify a partitioning of the search space which reflects the “typical” progress of the population towards the target set A_m .
2. Find parameter settings of the algorithm and corresponding parameters γ_0 and δ of the theorem, such that condition (G2) can be satisfied. It may be necessary to adjust the partitioning of the search space.
3. For each level $j \in [m-1]$, estimate lower bounds z_j such that condition (G1) holds.
4. Determine the lower bound on the population size λ in (G3) using the parameters obtained in the previous steps.
5. Once all conditions are satisfied, compute the bound on the expected time from the conclusion of the theorem. A simple way (not necessarily the only way) to evaluate the sum $\sum_{j=1}^{m-1} \ln\left(\frac{6\delta\lambda}{4+z_j\delta\lambda}\right)$ is to underestimate the denominator $4+z_j\delta\lambda$ in each term by either 4, or $z_j\delta\lambda$. This gives the bounds $m \ln(3\lambda/2)$, or $\sum_{j=1}^{m-1} \ln(6/z_j)$ for this sum.

Some iterations of the above steps may be required to find parameter settings that yield the best possible bound.

We now illustrate this methodology on a simple example.

Corollary 2. *For any $\lambda \geq 72(\ln(n)+9)$, the expected number of points created until the population of Algorithm 3 contains the point n is $\mathcal{O}(n\lambda)$.*

Algorithm 3 Example algorithm to illustrate Theorem 1.

Require: Finite state space $\mathcal{X} = \{1, \dots, n\}$ for some $n \in \mathbb{N}$.

- 1: $P_0 \sim \text{Unif}(\mathcal{X}^\lambda)$, i.e. initial population sampled u.a.r.
 - 2: **for** $t = 0, 1, 2, \dots$ until termination condition met **do**
 - 3: **for** $i = 1$ to λ **do**
 - 4: Sort the current population $P_t = (x_1, \dots, x_\lambda)$
 such that $x_1 \geq x_2 \geq \dots \geq x_\lambda$.
 - 5: $z := x_k$ where $k \sim \text{Unif}(\{1, \dots, \lambda/2\})$.
 - 6: $y := z + \text{Unif}(\{-c, 0, 1\})$ for any $c \in [n]$
 - 7: $P_{t+1}(i) := \max\{1, \min\{y, n\}\}$.
 - 8: **end for**
 - 9: **end for**
-

The purpose of Algorithm 3 is to illustrate the application of Theorem 1 on a very simple example. The search space \mathcal{X} is the set of natural numbers between 1 and n . A population is a vector of λ such numbers, and the implicit objective is to obtain a population containing the number n . Following the scheme of Algorithm 1, the operator D corresponds to lines 4-6. The new individual y is obtained by first selecting uniformly at random one of the best $\lambda/2$ individuals in the population (lines 4 and 5), and “mutating” this individual by adding 1, subtracting c , or do nothing, with equal probabilities. The value of c does not matter in our analysis. Note that for c being fixed to 1, one could ignore the selection steps and easily come up with a rough bound $\mathcal{O}(n^2\lambda)$. However, for other choices of c , e.g. equal to n or randomly picked, without the tool proposed in Theorem 1 it is much less obvious how such a process should be approached and analysed.

We now carry out the steps described previously.

Step 1: It seems natural to partition the search space into $m = n$ levels, where $A_j := \{j\}$ for all $j \in [m]$.

Step 2: Assume that the *current level* is $j < n - 1$. This means that in P_t , there are $\gamma_0\lambda$ individuals in $A_{\geq j}$, i.e. with fitness at least j , and at least $\gamma\lambda$ but less than $\gamma_0\lambda$ individuals in $A_{\geq j+1}$, i.e. with fitness at least $j + 1$. We need to estimate $\Pr_{y \sim D(P_t)}(y \in A_{\geq j+1})$, i.e., the probability of producing an individual with fitness at least $j + 1$. To this end, we say that a selection event is “good” if in step 5, the algorithm selects an individual in $A_{\geq j+1}$, i.e. with fitness at least $j + 1$. If $\gamma \leq 1/2$, then the probability of a good selection event is at least $\gamma\lambda/(\lambda/2) = 2\gamma$. And we say that a mutation event is “good” if in line 5, the algorithm does not subtract 1 from the selected search point. The probability of a good mutation event is $2/3$. Selection and mutation are independent events, hence we have shown for all $\gamma \in (0, 1/2]$ that

$$\Pr_{y \sim D(P_t)}(y \in A_{\geq j+1}) \geq (2\gamma)(2/3) = \gamma \left(1 + \frac{1}{3}\right).$$

Condition (G2) is therefore satisfied with $\delta = 1/3$ if we choose any positive constant $\gamma_0 \leq 1/2$.

Step 3: Assume that population P_t has at least $\gamma_0\lambda$ individuals in $A_{\geq j}$. In this case, the algorithm produces an individual in $A_{\geq j+1}$ if in line 5 it selects an individual in $A_{\geq j}$ and mutates the individual by adding 1 in line 6. If we now fix $\gamma_0 = 1/2$, the probability of selecting an individual in $A_{\geq j}$ is 1 by the

assumption. Furthermore, the probability of adding 1 to the selected individual is exactly $1/3$. Hence, we have shown

$$\Pr_{y \sim D(P_t)}(y \in A_{\geq j+1}) \geq 1(1/3),$$

and we can satisfy condition (G1) by defining $z_j := 1/3$ for all $j \in [m-1]$.

Step 4: For the parameters we have chosen, it is easy to see by numerical calculation that the population size $\lambda \geq 72(\ln(n) + 9)$ satisfies condition (G3).

Step 5: We use $\sum_{j=1}^{m-1} \ln\left(\frac{6}{z_j}\right)$ instead of $\sum_{j=1}^{m-1} \ln\left(\frac{6\delta\lambda}{4+\delta\lambda z_j}\right)$, thus the expected time until the population has found the point n is no more than

$$\mathcal{O}\left(\sum_{j=1}^{n-1} \frac{1}{3} + \lambda \sum_{j=1}^{n-1} \ln\left(\frac{6}{1/3}\right)\right) = \mathcal{O}(n\lambda).$$

2.3 Proof of the level-based theorem

Theorem 1 will be proved using drift analysis, which is a standard tool in theory of randomised search heuristics. Our distance function takes into account both the ‘‘current level’’ of the population, as well as the distribution of the population around the current level. In particular, let the current level Y_t be the highest level $j \in [m]$ such that there are at least $\gamma_0\lambda$ individuals at level j or higher. Furthermore, for any level $j \in [m]$, let $X_t^{(j)}$ be the number of individuals at level j or higher. Hence, we describe the dynamics of the population by $m+1$ stochastic processes $X_t^{(1)}, \dots, X_t^{(m)}, Y_t$. Assuming that these processes are adapted to a filtration \mathcal{F}_t , we write $E_t[X] := E[X | \mathcal{F}_t]$ and $\Pr_t(\mathcal{E}) := E[\mathbb{1}_{\mathcal{E}} | \mathcal{F}_t]$. Our approach is to measure the distance of the population at time t by a scalar $g(X_t^{(Y_t+1)}, Y_t)$, where g is a function that satisfies the conditions in Definition 3.

Definition 3. A function $g : \{0\} \cup [\lambda] \times [m] \rightarrow \mathbb{R}$ is called a level function if

1. $\forall x \in \{0\} \cup [\lambda], \forall y \in [m-1] \quad g(x, y) \geq g(x, y+1)$,
2. $\forall x \in \{0\} \cup [\lambda-1], \forall y \in [m] \quad g(x, y) \geq g(x+1, y)$, and
3. $\forall y \in [m-1] \quad g(\lambda, y) \geq g(0, y+1)$.

It is clear from the definition that the sum of two level functions is also a level function. In addition, the three conditions ensure that the distance $g(X_t^{(Y_t+1)}, Y_t)$ of the population decreases monotonically with the current level Y_t . As the following lemma shows, this monotonicity allows an upper bound on the distance in the next generation which is partly independent of the change in current level.

Lemma 4. If $Y_{t+1} \geq Y_t$, then for any level function g

$$g\left(X_{t+1}^{(Y_{t+1}+1)}, Y_{t+1}\right) \leq g\left(X_{t+1}^{(Y_t+1)}, Y_t\right).$$

Proof. The statement is trivially true when $Y_t = Y_{t+1}$. On the other hand, if $Y_{t+1} > Y_t$, then the conditions in Definition 3 imply

$$g\left(X_{t+1}^{(Y_{t+1}+1)}, Y_{t+1}\right) \leq g(0, Y_{t+1}) \leq g(0, Y_t + 1)$$

$$\leq g(\lambda, Y_t) \leq g\left(X_{t+1}^{(Y_t+1)}, Y_t\right). \quad \square$$

We can now give the formal proof of Theorem 1.

Proof of Theorem 1. We will prove the theorem using Lemma 22 (the additive drift theorem) with respect to the parameter $a = 0$ and a stochastic process

$$Z_t := g\left(X_t^{(Y_t+1)}, Y_t\right),$$

where g is a level-function to be defined, and $(Y_t)_{t \in \mathbb{N}}$ and $(X_t^{(j)})_{t \in \mathbb{N}}$ for $j \in [m]$ are stochastic processes to be defined. We consider the filtration $(\mathcal{F}_t)_{t \in \mathbb{N}}$ induced by the sequence of populations $(P_t)_{t \in \mathbb{N}}$.

We will assume w.l.o.g. that condition (G2) is also satisfied for $j = m - 1$, for the following reason. Given an Algorithm 1 with certain mapping D , consider Algorithm 1 with a different mapping $D'(P)$: If $|P \cap A_m| = 0$ then $D'(P) = D(P)$; otherwise $D'(P)$ assigns probability mass 1 to some element x of P that is in A_m , e.g. to the first one among such elements. Note that D' meets conditions (G1) and (G2). Moreover, (G2) holds for $j = m - 1$. For the sequence of populations P'_0, P'_1, \dots of Algorithm 1 with mapping D' we can put $T' := \lambda \cdot \min\{t \mid |P'_t \cap A_m| > 0\}$. Executions of the original algorithm and the modified one before generation T'/λ are identical. On generation T'/λ both algorithms place elements of A_m into the population for the first time. Thus, T' and T are equal in every realisation and their expected values is the same.

For any level $j \in [m]$ and time $t \geq 0$, let the random variable

$$X_t^{(j)} := |P_t \cap A_{\geq j}|$$

denote the number of individuals in levels $A_{\geq j}$ at time t . Because $A_{\geq j}$ is partitioned into disjoint sets A_j and $A_{\geq j+1}$, the definition implies

$$|P_t \cap A_j| = X_t^{(j)} - X_t^{(j+1)} \quad (1)$$

Algorithm 1 samples all individuals in generation $t + 1$ independently from distribution $D(P_t)$. Therefore, given the current population P_t , $X_{t+1}^{(j)}$ is binomially distributed

$$X_{t+1}^{(j)} \sim \text{Bin}(\lambda, p_{t+1}^{(j)})$$

where $p_{t+1}^{(j)} := \Pr_{t,y \sim D(P_t)}(y \in A_{\geq j})$ is the probability of sampling an individual in level j or higher.

The *current level* Y_t of the population at time $t \geq 0$ is defined as

$$Y_t := \max\left\{j \in [m] \mid X_t^{(j)} \geq \gamma_0 \lambda\right\}.$$

Note that $(X_t^{(j)})_{t \in \mathbb{N}}$ and $(Y_t)_{t \in \mathbb{N}}$ are adapted to the filtration $(\mathcal{F}_t)_{t \in \mathbb{N}}$ because they are defined in terms of the population process $(P_t)_{t \in \mathbb{N}}$.

When $Y_t < m$, there exists a unique $\gamma < \gamma_0$ such that

$$X_t^{(Y_t+1)} = |P_t \cap A_{\geq Y_t+1}| = \gamma \lambda \quad (2)$$

$$X_t^{(Y_t)} = |P_t \cap A_{\geq Y_t}| \geq \gamma_0 \lambda \quad (3)$$

$$X_t^{(Y_t-1)} = |P_t \cap A_{\geq Y_t-1}| \geq \gamma_0 \lambda \quad (4)$$

In the case of $X_t^{(Y_t+1)} = 0$, it follows from (1), (2) and (3) that $|P \cap A_j| = X_t^{(Y_t)} \geq \gamma_0 \lambda$. Condition (G1) for level $j = Y_t$ then gives

$$p_{t+1}^{(Y_t+1)} = \Pr_{y \sim D(P_t)}(y \in A_{\geq Y_t+1}) \geq z_{Y_t} \quad (5)$$

Otherwise if $X_t^{(Y_t+1)} \geq 1$, conditions (G1) and (G2) for level $j = Y_t$ with (2) and (3) imply

$$p_{t+1}^{(Y_t+1)} = \Pr_{y \sim D(P_t)}(y \in A_{\geq Y_t+1}) \quad (6)$$

$$\geq \max \left\{ (1 + \delta) \frac{X_t^{(Y_t+1)}}{\lambda}, z_j \right\}. \quad (7)$$

Condition (G2) for level $j = Y_t - 1$ along with (3) and (4) also gives

$$p_{t+1}^{(Y_t)} = \Pr_{y \sim D(P_t)}(y \in A_{\geq Y_t}) \geq (1 + \delta) \gamma_0. \quad (8)$$

We now define the process $(Z_t)_{t \in \mathbb{N}}$ as $Z_t := 0$ if $Y_t = m$, and otherwise, if $Y_t < m$, we let $Z_t := g(X_t^{(Y_t+1)}, Y_t)$, where for all k , and for all $1 \leq j < m$, $g(k, j) = g_1(k, j) + g_2(k, j)$ and

$$\begin{aligned} g_1(k, j) &:= \ln \left(\frac{1 + (\delta/2)\lambda}{1 + (\delta/2) \max\{k, z_j \lambda / (1 + \delta)\}} \right) \\ &\quad + \sum_{i=j+1}^{m-1} \ln \left(\frac{1 + (\delta/2)\lambda}{1 + (\delta/2)\lambda z_i / (1 + \delta)} \right) \\ g_2(k, j) &:= \frac{1}{q_j} \left(1 - \frac{\delta^2}{7} \right)^k + \sum_{i=j+1}^{m-1} \frac{1}{q_i} \end{aligned}$$

where $q_j := 1 - (1 - z_j)^\lambda$.

It follows from Lemma 21 that $g(k, j)$ is a level function. Furthermore, $g(k, j) \geq 0$ for all $k \in \{0\} \cup [\lambda]$ and all $j \in [m]$. Due to properties 1 and 2 of level functions and Lemma 18, the distance is always bounded from above by

$$\begin{aligned} g(0, 1) &\leq \sum_{i=1}^{m-1} \left(\ln \left(\frac{1 + (\delta/2)\lambda}{1 + (\delta/2) z_i \lambda / (1 + \delta)} \right) + \frac{1}{q_i} \right) \\ &< \sum_{i=1}^{m-1} \left(\ln \left(\frac{4 + 2\delta\lambda}{4 + \delta z_i \lambda} \right) + 1 + \frac{1}{\lambda z_i} \right) \end{aligned} \quad (9)$$

using that $z_i \leq 1$,

$$< \sum_{i=1}^{m-1} \left(\ln \left(\frac{4 + 2\delta\lambda}{z_i(4 + \delta\lambda)} \right) + 1 + \frac{1}{\lambda z_i} \right) \quad (10)$$

$$= \sum_{i=1}^{m-1} \left(\ln \left(\frac{1}{z_i} \left(1 + \frac{\delta \lambda}{4 + \delta \lambda} \right) \right) + 1 + \frac{1}{\lambda z_i} \right)$$

$\ln(x) \leq x - 1$ for all $x > 0$ so

$$< \sum_{i=1}^{m-1} \left(\frac{2}{z_i} + \frac{1}{\lambda z_i} \right)$$

and $z_i \geq z_*$ and $\lambda \geq \lceil 4 \ln(128) \rceil = 20$ from (G3) so

$$< \frac{m}{z_*} \left(2 + \frac{1}{\lambda} \right) \leq \frac{41m}{20z_*}. \quad (11)$$

Hence, we have $0 \leq Z_t < g(0, 1) < \infty$ for all $t \in \mathbb{N}$ which implies that $E[Z_t] < \infty$ for all $t \in \mathbb{N}$, and condition 2 of the drift theorem is satisfied.

The “drift” of the process is the random variable

$$\Delta_{t+1} := g \left(X_t^{(Y_t+1)}, Y_t \right) - g \left(X_{t+1}^{(Y_{t+1}+1)}, Y_{t+1} \right).$$

To compute the expected drift, we apply the law of total probability

$$\begin{aligned} E_t[\Delta_{t+1}] &= (1 - \Pr_t(Y_{t+1} < Y_t)) E_t[\Delta_{t+1} \mid Y_{t+1} \geq Y_t] \\ &\quad + \Pr_t(Y_{t+1} < Y_t) E_t[\Delta_{t+1} \mid Y_{t+1} < Y_t]. \end{aligned} \quad (12)$$

The event $Y_{t+1} < Y_t$ holds if and only if $X_{t+1}^{(Y_t)} < \gamma_0 \lambda$. Due to (8), we obtain the following by a Chernoff bound

$$\begin{aligned} \Pr_t(Y_{t+1} < Y_t) &= \Pr_t \left(X_{t+1}^{(Y_t)} < \left(1 - \frac{\delta}{1 + \delta} \right) (1 + \delta) \gamma_0 \lambda \right) \\ &\leq \exp \left(-\frac{\delta^2 \gamma_0 \lambda}{2(1 + \delta)} \right) \leq \frac{\delta^2 z_*}{128m}, \end{aligned} \quad (13)$$

where the second last inequality takes into account the population size required by condition (G3). Given the low probability of the event $Y_{t+1} < Y_t$, it suffices to use the pessimistic bound from (11)

$$E_t[\Delta_{t+1} \mid Y_{t+1} < Y_t] \geq -g(0, 1) \geq -\frac{41m}{20z_*}. \quad (14)$$

If $Y_{t+1} \geq Y_t$, we can apply Lemma 4

$$\begin{aligned} E_t[\Delta_{t+1} \mid Y_{t+1} \geq Y_t] &\geq E_t \left[g \left(X_t^{(Y_t+1)}, Y_t \right) - g \left(X_{t+1}^{(Y_t+1)}, Y_t \right) \mid Y_{t+1} \geq Y_t \right]. \end{aligned}$$

Note that event $Y_{t+1} \geq Y_t$ is equivalent to having $X_{t+1}^{(Y_t)} \geq \gamma_0 \lambda$, then due to Lemma 27, in the following we can skip the condition on the event when needed.

If $X_t^{(Y_t+1)} = 0$, then $X_t^{(Y_t+1)} \leq X_{t+1}^{(Y_t+1)}$ and

$$E_t \left[g_1 \left(X_t^{(Y_t+1)}, Y_t \right) - g_1 \left(X_{t+1}^{(Y_t+1)}, Y_t \right) \mid Y_{t+1} \geq Y_t \right] \geq 0,$$

because the function g_1 satisfies property 2 in Definition 3. Furthermore, we have the lower bound

$$\begin{aligned} E_t \left[g_2 \left(X_t^{(Y_t+1)}, Y_t \right) - g_2 \left(X_{t+1}^{(Y_t+1)}, Y_t \right) \mid Y_{t+1} \geq Y_t \right] \\ > \Pr_t \left(X_{t+1}^{(Y_t+1)} \geq 1 \right) (g_2(0, Y_t) - g_2(1, Y_t)) \geq \frac{\delta^2}{7}, \end{aligned}$$

where the last inequality follows because of (5) and $\Pr_t \left(X_{t+1}^{(Y_t+1)} \geq 1 \right) \geq 1 - \left(1 - p_{t+1}^{(Y_t+1)} \right)^\lambda \geq 1 - (1 - z_{Y_t})^\lambda = q_{Y_t}$, and $g_2(0, Y_t) - g_2(1, Y_t) = (1/q_{Y_t})(\delta^2/7)$.

In the other case, where $X_t^{(Y_t+1)} \geq 1$, we obtain

$$\begin{aligned} E_t \left[g_1 \left(X_t^{(Y_t+1)}, Y_t \right) - g_1 \left(X_{t+1}^{(Y_t+1)}, Y_t \right) \mid Y_{t+1} \geq Y_t \right] \\ \geq E_t \left[\ln \left(\frac{1 + \frac{\delta}{2} X_{t+1}^{(Y_t+1)}}{1 + \frac{\delta}{2} \max \left\{ X_t^{(Y_t+1)}, \frac{z_{Y_t} \lambda}{1+\delta} \right\}} \right) \right] \geq \frac{\delta^2}{7}. \end{aligned}$$

where the last inequality follows from (7) and Lemma 23. For function g_2 , we get

$$\begin{aligned} E_t \left[g_2 \left(X_t^{(Y_t+1)}, Y_t \right) - g_2 \left(X_{t+1}^{(Y_t+1)}, Y_t \right) \mid Y_{t+1} \geq Y_t \right] = \\ \frac{1}{q_{Y_t}} \left(\left(1 - \frac{\delta^2}{7} \right)^{X_t^{(Y_t+1)}} - E_t \left[\left(1 - \frac{\delta^2}{7} \right)^{X_{t+1}^{(Y_t+1)}} \right] \right) > 0 \end{aligned}$$

where the last inequality is due to Lemma 26, applied to

$$X_{t+1}^{(Y_t+1)} \sim \text{Bin}(\lambda, p_{t+1}^{(Y_t+1)})$$

with $p_{t+1}^{(Y_t+1)} \geq (1 + \delta)X_t^{(Y_t+1)}/\lambda$ (see (7)) and the parameter

$$\kappa = -\ln(1 - \delta^2/7) < \delta.$$

Taking into account all cases, we have

$$E_t [\Delta_{t+1} \mid Y_{t+1} \geq Y_t] \geq \frac{\delta^2}{7}. \quad (15)$$

We now have bounds for all the quantities in (12) with (13), (14) and (15), and we get

$$\begin{aligned} E_t [\Delta_{t+1}] &= (1 - \Pr_t(Y_{t+1} < Y_t)) E_t [\Delta_{t+1} \mid Y_{t+1} \geq Y_t] \\ &\quad + \Pr_t(Y_{t+1} < Y_t) E_t [\Delta_{t+1} \mid Y_{t+1} < Y_t] \\ &\geq \left(1 - \frac{\delta^2 z_*}{128m} \right) \frac{\delta^2}{7} - \left(\frac{\delta^2 z_*}{128m} \right) \left(\frac{41m}{20z_*} \right) > \frac{\delta^2}{8}. \end{aligned}$$

We now verify condition 3 of the drift theorem (Lemma 22), i.e., that T has finite expectation. Let $p_* := \min\{(1 + \delta)/\lambda, z_*\}$, and note by conditions (G1) and (G2) that the current level increases by at least one with probability

$\Pr_t(Y_{t+1} > Y_t) \geq (p_*)^{\gamma_0 \lambda}$. Furthermore, due to the definition of the modified process D' , if $Y_t = m$ then $Y_{t+1} = m$. Hence, the probability of reaching $Y_t = m$ is lower bounded by the probability of the event that the current level increases in all of at most m consecutive generations, i.e., $\Pr_t(Y_{t+m} = m) \geq (p_*)^{\gamma_0 \lambda m} > 0$. It follows that $E[T] < \infty$.

By Lemma 22 and the upper bound on $g(0, 1)$ in (9),

$$\begin{aligned} E[T] &\leq \lambda \cdot \frac{g(0, 1)}{\delta^2/8} \\ &< \left(\frac{8\lambda}{\delta^2}\right) \sum_{i=1}^{m-1} \left(\ln\left(\frac{4+2\delta\lambda}{4+z_i\delta\lambda}\right) + 1 + \frac{1}{\lambda z_i}\right) \end{aligned}$$

then using that $4 \leq \delta\lambda/5$ from (G3) and $(1/5 + 2)e < 6$

$$< \left(\frac{8\lambda}{\delta^2}\right) \sum_{i=1}^{m-1} \left(\ln\left(\frac{6\delta\lambda}{4+z_i\delta\lambda}\right) + \frac{1}{\lambda z_i}\right). \quad \square$$

3 Tools for Analysis of Genetic Algorithms

This section provides two corollaries of Theorem 1 tailored to Algorithm 2. After that, we give sufficient conditions for tunable parameters of many selection mechanisms allowing the applicability of the corollaries.

Since no explicit fitness function is defined in Algorithm 2, no assumption on a f -based partition will be required by the corollaries. Nevertheless, we have to generalise the *cumulative selection probability* function of Sel operator, denoted $\beta(\gamma, P)$ [16] which is defined relative to the fitness function f , to the one that is relative to the order of the partition (A_1, \dots, A_m) .

Recall that to define $\beta(\gamma, P)$ of Sel w.r.t. f for a population P of λ search points, we first assume (f_1, \dots, f_λ) to be the vector of sorted fitness values of P , i.e. $f_i \geq f_{i+1}$ for each $i \in [\lambda - 1]$. Then given $\gamma \in (0, 1]$, define

$$\beta(\gamma, P) := \sum_{i=1}^{\lambda} p_{\text{sel}}(i | P) \cdot [f(P(i)) \geq f_{\lceil \gamma \lambda \rceil}],$$

where $[\cdot]$ denotes the Iverson bracket.

Similarly, given a partition (A_1, \dots, A_m) , if we use $(\ell_1, \dots, \ell_\lambda)$ to denote the sorted levels of search points in P , i.e. $\ell_i \geq \ell_{i+1}$ for each $i \in [\lambda - 1]$, then the *cumulative selection probability* of function of Sel w.r.t. (A_1, \dots, A_m) is

$$\zeta(\gamma, P) := \sum_{i=1}^{\lambda} p_{\text{sel}}(i | P) \cdot [P(i) \in A_{\geq \ell_{\lceil \gamma \lambda \rceil}}].$$

Let (A_1, \dots, A_m) be a f -based partition of \mathcal{X} , it follows from the above definitions that

$$\forall P \in \mathcal{X}^\lambda, \forall \gamma \in (0, 1] \quad \zeta(\gamma, P) \geq \beta(\gamma, P) \quad (16)$$

for which the equality occurs when the partition is canonical.

3.1 Analysis of non-permanent use of crossover

We first derive from Theorem 1 a corollary that is adapted to Algorithm 2 with $p_c < 1$. This setting covers the case $p_c = 0$, i.e. only unary variation operators are used. This specific case is the main subject of [16], and to some extent our corollary shares many similarities with the main theorem of that paper. As we will see later on, stronger and more general results can be claimed with the corollary.

Corollary 5. *Let (A_1, \dots, A_m) be a partition of \mathcal{X} , define $T := \min\{t\lambda \mid |P \cap A_m| > 0\}$ to be the first point in time that Algorithm 2 with $p_c < 1$ obtains an element of A_m . If there exist $s_1, \dots, s_{m-1}, s_*, p_0, \delta \in (0, 1]$, and a constant $\gamma_0 \in (0, 1)$ such that*

(M1) *for each level $j \in [m - 1]$*

$$p_{\text{mut}}(y \in A_{\geq j+1} \mid x \in A_j) \geq s_j$$

(M2) *for each level $j \in [m - 1]$*

$$p_{\text{mut}}(y \in A_{\geq j} \mid x \in A_j) \geq p_0$$

(M3) *for any population $P \in (\mathcal{X} \setminus A_m)^\lambda$ and $\gamma \in (0, \gamma_0]$*

$$\zeta(\gamma, P) \geq \frac{(1 + \delta)\gamma}{p_0(1 - p_c)}$$

(M4) *the population size λ satisfies*

$$\lambda \geq \left(\frac{4}{\gamma_0 \delta^2}\right) \ln \left(\frac{128m}{\gamma_0 s_* \delta^2}\right) \text{ where } s_* := \min_{j \in [m-1]} \{s_j\},$$

then

$$E[T] < \left(\frac{8}{\delta^2}\right) \sum_{j=1}^{m-1} \left(\lambda \ln \left(\frac{6\delta\lambda}{4 + \gamma_0 s_j \delta \lambda}\right) + \frac{1}{\gamma_0 s_j}\right).$$

Proof. Following the guideline, we apply Theorem 1. Step 1 is skipped because we already have the partition. Step 2: We assume $|P \cap A_{\geq j}| \geq \gamma_0 \lambda$ and $|P \cap A_{\geq j+1}| \geq \gamma \lambda > 0$ for some $\gamma \leq \gamma_0$. Hence, to create an individual in $A_{\geq j+1}$ it suffices to pick an $x \in |P \cap A_k|$ for any $k \geq j + 1$ and mutate it to an individual in $A_{\geq k}$, the probability of such an event according to (M2) and (M3) is at least $(1 - p_c)\zeta(\gamma, P)p_0 \geq (1 + \delta)\gamma$. So (G2) holds for the same p_0 , δ and γ_0 as in (M3).

Step 3: We are given $|P \cap A_j| \geq \gamma_0 \lambda$. Thus, with probability $\zeta(\gamma_0, P)$, the selection mechanism chooses an individual x in either A_j or $A_{\geq j+1}$. If the individual x belongs to A_j , then the mutation operator will by condition (M1) upgrade the individual to $A_{\geq j+1}$ with probability s_j . If the individual belongs to $A_{\geq j+1}$, then by (M2), the mutation operator maintains the individual in $A_{\geq j+1}$ with probability p_0 . Finally, no crossover occurs with probability $1 - p_c$. Hence, the probability of producing an individual in $A_{\geq j+1}$ is at least

$$(1 - p_c)\zeta(\gamma_0, P) \min\{s_j, p_0\} \geq (1 - p_c)\zeta(\gamma_0, P)s_j p_0$$

$$> \gamma_0 s_j =: z_j.$$

Thus (G1) holds for that choice of z_j and $z_* := \gamma_0 s_*$.

Step 4: Given our choice of z_* , we have that condition (M4) implies condition (G3).

For the last step, all conditions (G1-3) are satisfied, and Theorem 1 gives

$$\begin{aligned} E[T] &\leq \frac{8\lambda}{\delta^2} \sum_{j=1}^{m-1} \left(\ln \left(\frac{6\delta\lambda}{4 + z_j\delta\lambda} \right) + \frac{1}{z_j\lambda} \right) \\ &= \frac{8}{\delta^2} \sum_{j=1}^{m-1} \left(\lambda \ln \left(\frac{6\delta\lambda}{4 + \gamma_0 s_j \delta\lambda} \right) + \frac{1}{\gamma_0 s_j} \right). \quad \square \end{aligned}$$

From the proof, we remark that any operator can be used in place of crossover in line 5 of Algorithm 2, and the result still holds. Therefore, the corollary is in fact applicable to a wider range of algorithms than just Algorithm 2.

3.2 Analysis of permanent use of crossover

The following corollary adapts Theorem 1 to the setting of Algorithm 2 with $p_c = 1$.

Corollary 6. *Given a partition (A_1, \dots, A_m) of \mathcal{X} , let $T := \min\{t\lambda \mid |P \cap A_m| > 0\}$ be the first point in time that Algorithm 2 with $p_c = 1$ obtains an element of A_m . If there exist $s_1, \dots, s_{m-1}, s_*, p_0, \varepsilon, \delta \in (0, 1]$, and a constant $\gamma_0 \in (0, 1)$ such that*

(C1) *for each level $j \in [m-1]$*

$$p_{\text{mut}}(y \in A_{\geq j+1} \mid x \in A_j) \geq s_j$$

(C2) *for each level $j \in [m]$*

$$p_{\text{mut}}(y \in A_{\geq j} \mid x \in A_j) \geq p_0$$

(C3) *for each level $j \in [m-2]$*

$$p_{\text{xor}}(x \in A_{\geq j+1} \mid u \in A_{\geq j}, v \in A_{\geq j+1}) \geq \varepsilon$$

(C4) *for any population $P \in (\mathcal{X} \setminus A_m)^\lambda$ and $\gamma \in (0, \gamma_0]$*

$$\zeta(\gamma, P) \geq \gamma \sqrt{\frac{1 + \delta}{p_0 \gamma_0 \varepsilon}}$$

(C5) *the population size satisfies*

$$\lambda \geq \left(\frac{4}{\gamma_0 \delta^2} \right) \ln \left(\frac{128m}{\gamma_0 \delta^2 s_*} \right) \text{ where } s_* := \min_{j \in [m-1]} \{s_j\}$$

then

$$E[T] < \left(\frac{8}{\delta^2} \right) \sum_{j=1}^{m-1} \left(\lambda \ln \left(\frac{6\delta\lambda}{4 + \gamma_0 s_j \delta\lambda} \right) + \frac{1}{\gamma_0 s_j} \right).$$

Proof. We apply Theorem 1 following the guideline. Again, Step 1 is skipped because the partition is already defined.

Step 2: We are given $|P \cap A_{\geq j}| \geq \gamma_0 \lambda$ and $|P \cap A_{\geq j+1}| \geq \gamma \lambda > 0$. To create an individual in $A_{\geq j+1}$, it suffices to pick the individual u in $A_{\geq j}$ and v in $A_{\geq j+1}$, then to produce an individual in A_k for any $k \geq j+1$ by crossover and not destroy the produced individual by mutation. The probability of such an event according to (C2), (C3) and (C4) is bounded from below by $\zeta(\gamma_0, P)\zeta(\gamma, P)\varepsilon p_0 \geq (1 + \delta)\gamma$. Condition (G2) is then satisfied with the same γ_0 and δ as in (C4).

Step 3: We assume $|P \cap A_j| \geq \gamma_0 \lambda$. Note that condition (C3) written for level $j-1$ is $p_{\text{xor}}(x \in A_{\geq j} \mid u \in A_{\geq j-1}, v \in A_{\geq j}) \geq \varepsilon$, and because $A_{\geq j} \subset A_{\geq j-1}$ then $p_{\text{xor}}(x \in A_{\geq j} \mid u \in A_{\geq j}, v \in A_{\geq j}) \geq \varepsilon$. To create an individual in $A_{\geq j+1}$, it then suffices to pick both u and v from $A_{\geq j}$ in line 4, then to produce an individual in A_k for any $k \geq j$ by crossover, now if $k = j$ we need to improve the produced individual by mutation, i.e. relying on (C1), otherwise if $k > j$ it suffices not to destroy the produced individual by mutation, i.e. relying on (C2). It then follows from (C4) that the probability of producing an individual in $A_{\geq j+1}$ is at least

$$\zeta(\gamma_0, P)^2 \varepsilon \min\{s_j, p_0\} \geq \zeta(\gamma_0, P)^2 \varepsilon s_j p_0 > \gamma_0 s_j =: z_j.$$

Condition (G1) then holds for that choice of z_j and $z_* := \gamma_0 s_*$.

Step 4: It follows from the above definition of z_* that (C5) implies (G3).

In the last step, since all conditions (G1-3) are satisfied, Theorem 1 guarantees that

$$E[T] \leq \frac{8}{\delta^2} \sum_{j=1}^{m-1} \left(\lambda \ln \left(\frac{6\delta\lambda}{4 + \gamma_0 s_j \delta \lambda} \right) + \frac{1}{\gamma_0 s_j} \right). \quad \square$$

The corollary shares many similarities with Corollary 5, except that condition (C2) has to additionally hold for level A_m , that (C3) is a new condition on the Cross operator, and that condition (C4) on Sel operator is different from (M3).

3.3 Analysis of selection mechanisms

We show how to parameterise the following selection mechanisms such that condition (M3) of Corollary 5 and (C4) of Corollary 6 are satisfied. In *k-tournament selection*, k individuals are sampled uniformly at random with replacement from the population, and the fittest of these individuals is returned. In (μ, λ) -*selection*, parents are sampled uniformly at random among the fittest μ individuals in the population. A function $\alpha : \mathbb{R} \rightarrow \mathbb{R}$ is a ranking function [28] if $\alpha(x) \geq 0$ for all $x \in [0, 1]$, and $\int_0^1 \alpha(x) dx = 1$. In ranking selection with ranking function α , the probability of selecting individuals ranked γ or better is $\int_0^\gamma \alpha(x) dx$. In *linear ranking selection* parametrised by $\eta \in (1, 2]$, the ranking function is $\alpha(\gamma) := \eta(1 - 2\gamma) + 2\gamma$. We define *exponential ranking selection* parametrised by $\eta > 0$ with $\alpha(\gamma) := \eta e^{\eta(1-\gamma)} / (e^\eta - 1)$.

Lemma 7. *Assuming that (A_1, \dots, A_m) is a partition of \mathcal{X} with (A_1, \dots, A_{m-1}) being an f -based partition of $\mathcal{X} \setminus A_m$, for any constants $\delta' > 0$, $p_0 \in (0, 1)$, $\varepsilon \in (0, 1)$, and for any non-negative parameter $p_c = 1 - \Omega(1)$, there exists a constant $\gamma_0 \in (0, 1)$ such that all the following selection mechanisms*

1. k -tournament selection,
2. (μ, λ) -selection,
3. linear ranking selection,
4. exponential ranking selection

with their parameters k , λ/μ and η being set to no less than $\frac{1 + \delta'}{(1 - p_c)p_0}$ satisfy (M3), i. e. $\zeta(\gamma, P) \geq \frac{(1 + \delta'')\gamma}{p_0(1 - p_c)}$ for any $\gamma \in (0, \gamma_0]$, any $P \in (\mathcal{X} \setminus A_m)^\lambda$ and some constant $\delta'' > 0$.

Proof. Since (M3) only concerns with the restricted subspace $\mathcal{X} \setminus A_m$ we only need to focus on this subspace, and because the partition is f -based on it, due to (16) it suffices to prove the results for β function instead of ζ function.

The results for k -tournament, (μ, λ) -selection and linear ranking follow by applying Lemma 13 in [16] (with its p_0 being set as our $p_0(1 - p_c)$). For exponential ranking, we first remark the following lower bound,

$$\begin{aligned} \beta(\gamma, P) &\geq \int_0^\gamma \frac{\eta e^{\eta(1-x)} dx}{e^\eta - 1} = \left(\frac{e^\eta}{e^\eta - 1} \right) \left(1 - \frac{1}{e^{\eta\gamma}} \right) \\ &\geq 1 - \frac{1}{1 + \eta\gamma}. \end{aligned}$$

Then the rest of the proof is similar to k -tournament with η in place of k . \square

Lemma 8. *Assuming that (A_1, \dots, A_m) is a partition of \mathcal{X} with (A_1, \dots, A_{m-1}) being an f -based partition of $\mathcal{X} \setminus A_m$, for any constants $\delta' > 0$, $p_0 \in (0, 1)$ and $\varepsilon \in (0, 1)$, there exists a constant $\gamma_0 \in (0, 1)$ such that the following selection mechanisms*

1. k -tournament selection with $k \geq 4(1 + \delta')/(\varepsilon p_0)$,
2. (μ, λ) -selection with $\lambda/\mu \geq (1 + \delta')/(\varepsilon p_0)$, and
3. exponential ranking selection with $\eta \geq 4(1 + \delta')/(\varepsilon p_0)$

satisfy (C4), i. e. $\zeta(\gamma, P) \geq \gamma \sqrt{\frac{1 + \delta'}{p_0 \varepsilon \gamma_0}}$ for any $\gamma \in (0, \gamma_0]$ and any $P \in (\mathcal{X} \setminus A_m)^\lambda$.

Proof. Similar to the proof of Lemma 7, we only focus on the subspace $\mathcal{X} \setminus A_m$ where the partition is f -based, and based on (16) we consider β function instead of ζ function.

Define $\varepsilon' := \varepsilon p_0$.

1. Consider k -tournament selection and let $\gamma \in (0, \gamma_0]$. By the definition of f -based partition, to select an individual from the same level as the γ -ranked individual or higher it is sufficient that the randomly sampled tournament contains at least one individual with rank γ or higher. Hence,

$$\beta(\gamma, P) \geq 1 - (1 - \gamma)^k > 1 - \frac{1}{1 + \gamma k},$$

because $(1 - \gamma)^k < e^{-\gamma k} < \frac{1}{1 + \gamma k}$. So for $k \geq 4(1 + \delta')/\varepsilon'$,

$$\beta(\gamma, P) \geq 1 - \frac{1}{1 + \frac{4\gamma(1+\delta')}{\varepsilon'}} = \frac{\frac{4\gamma(1+\delta')}{\varepsilon'}}{1 + \frac{4\gamma(1+\delta')}{\varepsilon'}}.$$

If $\gamma_0 := \varepsilon'/(4(1 + \delta'))$, then for all $\gamma \in (0, \gamma_0]$ it holds that $4(1 + \delta')/\varepsilon' \leq 1/\gamma$ and

$$\begin{aligned} \beta(\gamma, P) &\geq \frac{\gamma 4(1 + \delta')/\varepsilon'}{\gamma(1/\gamma) + 1} = \frac{2(1 + \delta')\gamma}{\varepsilon'} \\ &= \sqrt{\frac{(1 + \delta')}{\varepsilon'(\varepsilon'/4(1 + \delta'))}}\gamma = \sqrt{\frac{(1 + \delta')}{\varepsilon'\gamma_0}}\gamma. \end{aligned}$$

2. In (μ, λ) -selection, again by f -based property of the partition, we have $\beta(\gamma, P) = \lambda\gamma/\mu$ if $\gamma\lambda \leq \mu$, and $\beta(\gamma, P) = 1$ otherwise. It suffices to pick $\gamma_0 := \mu/\lambda$ so that with $\lambda/\mu \geq (1 + \delta')/\varepsilon'$, for all $\gamma \in (0, \gamma_0]$. Then

$$\beta(\gamma, P) \geq \frac{\lambda\gamma}{\mu} = \sqrt{\frac{\lambda^2}{\mu^2}}\gamma = \sqrt{\frac{\lambda}{\mu\gamma_0}}\gamma \geq \sqrt{\frac{1 + \delta'}{\varepsilon'\gamma_0}}\gamma.$$

3. Similar to the proof of Lemma 7, we remark that $\beta(\gamma, P) \geq 1 - \frac{1}{1 + \eta\gamma}$, thus the rest of the proof is similar to k -tournament selection. \square

4 Applications to Genetic Algorithms for Different Problems

This section applies the results from the previous section to derive bounds on the expected runtime of GAs for optimising pseudo-Boolean functions and solving combinatorial Optimization problems.

In what follows, by bitwise mutation operator we mean an operator that given a bitstring x , computes a bitstring y , where independently of other bits, each bit y_i is set to $1 - x_i$ with probability p_m and with probability $1 - p_m$ it is set equal to x_i . The tunable parameter p_m is called a *mutation rate*.

4.1 Optimisation of pseudo-Boolean functions

In this subsection, we consider the expected runtime of non-elitist GAs in Algorithm 2 on the following functions,

$$\text{ONEMAX}(x) := \sum_{i=1}^n x_i = |x|_1 = \text{OM}(x),$$

$$\text{LEADINGONES}(x) := \sum_{i=1}^n \prod_{j=1}^i x_j = \text{LO}(x),$$

$$\text{JUMP}_r(x) := \begin{cases} n + 1 & \text{if } |x|_1 = n \\ r + |x|_1 & \text{if } |x|_1 \leq n - r, \\ n - |x|_1 & \text{otherwise} \end{cases}$$

$$\text{ROYALROAD}_r(x) := \sum_{i=0}^{n/r-1} \prod_{j=1}^r x_{ir+j},$$

$$\text{LINEAR}(x) := \sum_{j=1}^n c_j x_j.$$

Note that our results on these functions also hold for their generalised classes, i.e. the meaning of 0-bit and 1-bit in each position can be exchanged, and/or x is rearranged according to a fixed permutation before each evaluation. For LINEAR, w.o.l.g. we can assume $c_1 \geq c_2 \geq \dots \geq c_n > 0$ [16]. We also consider the class of ℓ -UNIMODAL functions, for which each function has exactly ℓ distinctive fitness values $f_1 < f_2 < \dots < f_\ell$, and each bitstring x of the search space is either optimal or it has a Hamming-neighbour y with a better fitness, i.e. $f(y) > f(x)$.

For a moderate use of crossover, i.e. $p_c = 1 - \Omega(1)$, Corollary 5 is applicable and provides upper bounds on the expected runtime for all these functions and classes.

Theorem 9. *The expected runtime of the GA in Algorithm 2, with $p_c = 1 - \Omega(1)$ using any crossover operator, a bitwise mutation with mutation rate χ/n for any fixed constant $\chi > 0$ and one of the selection mechanisms: k -tournament selection, (μ, λ) -selection, linear or exponential ranking selection, with their parameters k , λ/μ and η being set to no less than $(1 + \delta)e^\chi/(1 - p_c)$ where $\delta > 0$ being any constant, is*

- $\mathcal{O}(n\lambda)$ on ONEMAX if $\lambda \geq c \ln n$,
- $\mathcal{O}(n^2 + n\lambda \ln \lambda)$ on LEADINGONES if $\lambda \geq c \ln n$,
- $\mathcal{O}(n\ell + \ell\lambda \ln \lambda)$ on ℓ -UNIMODAL if $\lambda \geq c \ln(\ell n)$,
- $\mathcal{O}(n^2 + n\lambda \ln \lambda)$ on LINEAR if $\lambda \geq c \ln n$,
- $\mathcal{O}\left(\left(\frac{n}{\chi}\right)^r + n\lambda + \lambda \ln \lambda\right)$ on JUMP $_r$ if $\lambda \geq cr \ln n$,
- $\mathcal{O}\left(\left(\frac{n}{\chi}\right)^r \ln\left(\frac{n}{r}\right) + \frac{n\lambda \ln \lambda}{r}\right)$ on ROYALROAD $_{r \geq 2}$ if $\lambda \geq cr \ln n$,

for some sufficiently large constant c .

Proof. We apply Corollary 5 with the canonical partition $A_j := \{j \mid f(x) = j\}$ for all functions¹, except for LINEAR, the fitness-based partition [16]:

$$A_j := \left\{ x \mid \sum_{i=1}^j c_i \leq f(x) < \sum_{i=1}^{j+1} c_i \right\}$$

for $j \in \{0\} \cup [n-1]$ and $A_n := \{1^n\}$, is used.

The choices of s_j and s_* to satisfy (M1) are the following.

¹The first level can be A_0 instead of A_1 for some functions but that does not matter as far as we compute the sums correctly later on.

- For ONEMAX, we set

$$s_j := \binom{n-j}{1} \left(\frac{\chi}{n}\right) \left(1 - \frac{\chi}{n}\right)^{n-1} = \Omega\left(\frac{n-j}{n}\right),$$

i.e. the probability of flipping a 0-bit while keeping all the other bits unchanged, and $s_* := \Omega\left(\frac{1}{n}\right)$.

- For LEADINGONES, ℓ -UNIMODAL and LINEAR, we set

$$s_j := \frac{\chi}{n} \left(1 - \frac{\chi}{n}\right)^{n-1} = \Omega\left(\frac{1}{n}\right) =: s_*,$$

i.e. the probability of flipping a specific bit to create a Hamming neighbour solution with better fitness while keeping all the other bits unchanged. In ℓ -UNIMODAL, the bit to flip must exist by the definition of the function. In LEADINGONES, the 0-bit at position $j+1$ should be flipped. For LINEAR, the partition satisfies that among the first $j+1$ bits there must be at least a 0-bit, thus it suffices to flip the left most 0-bit will produce a search point at a higher level.

- For JUMP $_r$, as similar to ONEMAX for $j \in [n-1]$ we use

$$s_j := \binom{n-j+1}{1} \left(\frac{\chi}{n}\right) \left(1 - \frac{\chi}{n}\right)^{n-1} = \Omega\left(\frac{n-j+1}{n}\right),$$

but $s_n := \left(\frac{\chi}{n}\right)^r \left(1 - \frac{\chi}{n}\right)^{n-r} = \Omega\left(\left(\frac{\chi}{n}\right)^r\right)$, i.e. the probability of flipping the r remained 0-bits, so $s_* := \Omega\left(\left(\frac{1}{n}\right)^r\right)$.

- For ROYALROAD $_r$, we use

$$s_j := \binom{n/r-j}{1} \left(\frac{\chi}{n}\right)^r \left(1 - \frac{\chi}{n}\right)^{n-r} = \Omega\left(\left(\frac{\chi}{n}\right)^r \left(\frac{n}{r} - j\right)\right),$$

i.e. the probability of flipping an entire unsolved block of length r (in the worst case) while keeping the other bits unchanged, and $s_* := \Omega\left(\left(\frac{1}{n}\right)^r\right)$.

It follows from Lemma 19 that the probability of not flipping any bit position by mutation is $(1 - \chi/n)^n \geq \left(1 - \frac{\delta/2}{1+\delta/2}\right) e^{-\chi} = \frac{e^{-\chi}}{1+\delta/2}$ for n sufficiently large, thus choosing $p_0 := \frac{e^{-\chi}}{1+\delta/2}$ satisfies (M2).

We now look at (M3). In k -tournament selection, we have

$$k \geq \frac{(1+\delta)e^\chi}{1-p_c} = \left(1 + \frac{\delta/2}{1+\delta/2}\right) \frac{1}{(1-p_c)p_0}.$$

Hence, it follows from Lemma 7 that (M3) is satisfied with constant $\delta' := \frac{\delta/2}{1+\delta/2}$. The same conclusion can be drawn for the other three selection mechanisms.

In (M4), since γ_0 and δ' are constants, there should exist a constant $c > 0$ for each function such that the condition is satisfied given the minimum requirement on population size related to c .

Since all conditions are satisfied, Corollary 5 gives the desired result for each function. For ONEMAX and JUMP $_r$, optimisation time can be saved at early levels, i.e. s_j is not small at the beginning, thus the evaluation of the sum $\sum_{j=1}^{m-1} \ln\left(\frac{6\delta\lambda}{4+\gamma_0 s_j \delta\lambda}\right)$ has to be precise:

- For ONEMAX, simplifying each term by $\ln\left(\frac{6}{\gamma_0 s_j}\right)$ gives

$$\mathcal{O}\left(\ln\left(\frac{6^n n^n}{\gamma_0^n \prod_{j=0}^{n-1} (n-j)}\right)\right) = \mathcal{O}\left(\ln\left(\frac{6^n n^n}{n! \gamma_0^n}\right)\right),$$

and by Stirling's approximation $n! = \Theta(n^{n+\frac{1}{2}}/e^n)$, this is no more than $\mathcal{O}(n)$. The expected runtime is then $\mathcal{O}(\lambda n + n \ln n)$. Since we already require $\lambda = \Omega(\ln n)$, this can be written shortly as $\mathcal{O}(n\lambda)$.

- For JUMP_r, we use the simplification $\ln\left(\frac{6}{\gamma_0 s_j}\right)$ for the first $m-2$ terms of the sum, and $\ln(3\lambda/2)$ for the last term, so this gives

$$\begin{aligned} \mathcal{O}\left(\ln\left(\frac{6^n n^{n-1}}{\gamma_0^n \prod_{j=1}^{n-1} (n-j+1)}\right) + \ln \lambda\right) &= \mathcal{O}\left(\ln\left(\frac{6^n n^n}{n! \gamma_0^n}\right) + \ln \lambda\right) \\ &= \mathcal{O}(n + \ln \lambda). \end{aligned}$$

The expected runtime is then $\mathcal{O}\left(\left(\frac{n}{\chi}\right)^r + n\lambda + \lambda \ln \lambda\right)$.

For the other functions, s_j is already small at early levels, thus there is no benefit of considering the gradual sum of \ln . Hence, the simplification $\mathcal{O}(m \ln \lambda)$ for the sum $\sum_{j=1}^{m-1} \ln\left(\frac{6\delta\lambda}{4+\gamma_0 s_j \delta\lambda}\right)$ gives the corresponding results. \square

In the case of regular use of crossover, i.e. $p_c = 1$, the relationship between the crossover operator and the structure of the search space becomes non-negligible. In the following, we consider a general *mask-based* crossover as follows. Given two parent genotypes u, v , the operator consists in first choosing (deterministically or randomly) a binary string $\tilde{m} = (m_1, \dots, m_n)$ to produce two offspring vectors x', x'' as

$$x'_i = \begin{cases} u_i, & \text{if } m_i = 1 \\ v_i, & \text{otherwise,} \end{cases} \quad x''_i = \begin{cases} v_i, & \text{if } m_i = 1 \\ u_i, & \text{otherwise.} \end{cases}$$

Then one element of $\{x', x''\}$ chosen uniformly at random is returned. For example, the uniform crossover is a mask-based crossover for which $\tilde{m} \sim \text{Unif}(\{0, 1\}^n)$, and a k -point crossover is a mask-based crossover for which

$$\tilde{m} \sim \text{Unif}\left(\left\{0^{a_1} 1^{a_2} 0^{a_3} \dots \mid a_i \in \mathbb{N}, \sum_{i=1}^{k+1} a_i = n\right\}\right).$$

The following lemma shows that all mask-based crossover operators satisfy (C3) with $\varepsilon = 1/2$ for OM and LO functions.

Lemma 10. *If $x \sim p_{\text{xor}}(u, v)$, where p_{xor} is a mask-based crossover, then:*

1. *If $\text{LO}(u) = \text{LO}(v) = j$, then $\Pr(\text{LO}(x) \geq j) = 1$, otherwise $\Pr(\text{LO}(x) > \min\{\text{LO}(u), \text{LO}(v)\}) \geq 1/2$.*
2. *$\Pr(\text{OM}(x) \geq \lceil (\text{OM}(u) + \text{OM}(v))/2 \rceil) \geq 1/2$.*

Proof. 1) When $\text{LO}(u) = \text{LO}(v) = j$, in mask-based crossover operators, the two bitstrings x' , x'' have j leading ones. So does the returned bitstring, i.e. with probability 1.

If $\text{LO}(u) \neq \text{LO}(v)$, we can assume w.o.l.g. that $\text{LO}(v) = j$ and $\text{LO}(u) > \text{LO}(v)$. Then v has a 0 while u has a 1 at position $j + 1$. So, one of the bitstrings x' , x'' in the mask-based crossover will inherit the 1 at that position and the other will inherit the 0. This implies that one of them has fitness at least $j + 1$ and with probability $1/2$ it is returned as output.

2) Each bit of u and v is copied either to x' or to x'' , therefore $|x'|_1 + |x''|_1 = |u|_1 + |v|_1$, which means that $\max\{|x'|_1, |x''|_1\} \geq \lceil (|u|_1 + |v|_1)/2 \rceil$. The output is chosen with probabilities $1/2$ to be copied either from x' or x'' , and the result follows. \square

Theorem 11. *Assume that the GA in Algorithm 2 with $p_c = 1$ uses any mask-based crossover operator, a bitwise mutation with mutation rate χ/n for any fixed constant $\chi > 0$, and one of the following selection mechanisms:*

- *k -tournament selection with $k \geq 8(1 + \delta)e^\chi$,*
- *(μ, λ) -selection with $\lambda/\mu \geq 2(1 + \delta)e^\chi$, or*
- *exponential ranking selection with $\eta \geq 8(1 + \delta)e^\chi$,*

for any constant $\delta > 0$. Then there exists a constant $c > 0$, such that the expected runtime of the GA is

- $\mathcal{O}(n\lambda)$ on ONEMAX if $\lambda \geq c \ln n$,
- $\mathcal{O}(n^2 + n\lambda \ln \lambda)$ on LEADINGONES if $\lambda \geq c \ln n$.

Proof. We apply Corollary 6 this time, but again using the canonical partition of the search space for both functions. We also assume that n is large enough so that by Lemma 19 the probability of not flipping any bit by mutation is $(1 - \chi/n)^n \geq \left(1 - \frac{\delta/2}{1+\delta/2}\right) e^{-\chi} = \frac{e^{-\chi}}{1+\delta/2} =: p_0$, and so (C2) is satisfied with this choice of p_0 . In addition, we use the same upgrade probabilities s_j and their smallest value s_* for each of the two functions as in the proof of Theorem 9 to satisfy (C1).

It follows from Lemma 10 that (C3) is satisfied for constant $\varepsilon_1 := 1/2$. We now look at condition (C4). For k -tournament, we get $k \geq 8(1 + \delta)e^\chi = 4 \left(1 + \frac{\delta/2}{1+\delta/2}\right) / (p_0 \varepsilon_1)$. So condition (C4) is satisfied with constant $\delta' := \frac{\delta/2}{1+\delta/2}$ for k -tournament by Lemma 8. The same reasoning can be applied so that (C4) is also satisfied for the other selection mechanisms.

Since δ' and γ_0 are constants, thus condition (C5) is satisfied given $\lambda \geq c \ln n$ and for some constant c . Since all conditions are satisfied, the result follows from Corollary 6. \square

Note that the upper bounds in Theorem 11 match the upper bounds of Theorem 9. The latter is a generalisation of the results in [16] which were limited to EAs without crossover.

In the next sections, we further demonstrate the generality of Theorem 1 through Corollary 5 by deriving bounds on the expected runtime of GAs with $p_c = 1 - \Omega(1)$ to optimise or to approximate the optimal solutions of combinatorial optimisation problems. We start with a simple problem of sorting n elements from a totally ordered set.

4.2 Optimisation on permutation space

Given n distinct elements from a totally ordered set, we consider the problem of ordering them so that some *measure of sortedness* is maximised. Several measures were considered by [52] in the context of analysing the (1+1)EA. One of those is $\text{INV}(\pi)$ which is defined to be the number of pairs (i, j) such that $1 \leq i < j \leq n$, $\pi(i) < \pi(j)$ (i.e. pairs in correct order). We show that with the method introduced in this paper, i.e. Corollary 5 analysing GAs on Sorting problem with INV measure, denoted by $\text{SORTING}_{\text{INV}}$, is not much harder than analysing the (1+1)EA.

For the mutation we use the $\text{Exchange}(\pi)$ operator [52], which consecutively applies N pairwise exchanges between uniformly selected pairs of indices, where N is a random number drawn from a Poisson distribution with parameter 1.

Theorem 12. *If the GA in Algorithm 2 with $p_c = 1 - \Omega(1)$ uses any crossover operator, the Exchange mutation operator, one of the selection mechanisms: k -tournament selection, (μ, λ) -selection, and linear or exponential ranking selection, with their parameters k , λ/μ and η being set to no less than $(1+\delta)e/(1-p_c)$, then there exists a constant $c > 0$ such that if the population size is $\lambda \geq c \ln n$, the expected time to obtain the optimum of $\text{SORTING}_{\text{INV}}$ is $\mathcal{O}(n^2\lambda)$.*

Proof. Define $m := \binom{n}{2}$. We apply Corollary 5 with the canonical partition, $A_j := \{\pi \mid \text{INV}(\pi) = j\}$ for $j = \{0\} \cup [m]$. The probability that mutation exchanges 0 pairs is $1/e$. Hence, condition (M2) is satisfied for $p_0 := 1/e$.

To show that (M1) is satisfied, we first define $s_j := \frac{m-j}{em}$ for each $j \in \{0\} \cup [m-1]$. Since $x \in A_j$, then the probability that the exchange operator exchanges exactly one pair is $1/e$, and the probability that this pair is incorrectly ordered in x , is $(m-j)/m$. Thus, (M1) is satisfied with the defined s_j .

In (M3), for k -tournament we have that $k \geq \frac{(1+\delta)e}{1-p_c} = \left(1 + \frac{\delta/2}{1+\delta/2}\right) \frac{1}{(1-p_c)p_0}$, thus the condition is satisfied for constant $\delta' := \frac{\delta/2}{1+\delta/2}$ and some constant $\gamma_0 \in (0, 1)$ by Lemma 7. The same conclusion can be drawn by the other selection mechanisms. Finally, since γ_0, δ' are constants, there exists a constant $c > 0$ such that (M4) is satisfied for any $\lambda \geq c \ln(n)$.

It therefore follows by Corollary 5 that the expected runtime of the GA on $\text{SORTING}_{\text{INV}}$ is $\mathcal{O}(n^2\lambda)$, i.e. this is similar to ONEMAX except that we have $m = O(n^2)$ levels. \square

4.3 Search for Local Optima

A great interest in the area of combinatorial optimisation is to find approximate solutions to NP-hard problems, because exact solutions for such problems are unlikely to be computable in polynomial time under the so-called $\text{P} \neq \text{NP}$ hypothesis. In the case of maximisation problems, a feasible solution is called a ρ -approximate solution if its objective function value is at least ρ times the optimum for some $\rho \in (0, 1]$. Local search is one method among others to approximate solutions for combinatorial optimisation problems through finding local optima (a formal definition is given below). For a number of well-known problems it was shown [1] that any local optimum is guaranteed to be a ρ -approximate solution with a constant ρ .

Suppose that a *neighbourhood* $\mathcal{N}(x) \subseteq \mathcal{X}$ is defined for every $x \in \mathcal{X}$. The mapping $\mathcal{N} : \mathcal{X} \rightarrow 2^{\mathcal{X}}$ is called the *neighbourhood mapping* and all elements of $\mathcal{N}(x)$ are called *neighbours* of x . For example, a frequently used neighbourhood mapping in the case of binary search space $\mathcal{X} = \{0, 1\}^n$ is defined by the Hamming distance $H(\cdot, \cdot)$ and a radius r as

$$\mathcal{N}(x) = \{y \mid H(x, y) \leq r\}.$$

If $f(y) \leq f(x)$ holds for all neighbours y of $x \in \mathcal{X}$, then x is called a *local optimum* w.r.t. \mathcal{N} . The set of all local optima is denoted by \mathcal{LO} (note that global optima also belong to \mathcal{LO}).

A local search method starts from some initial solution y_0 . Each iteration of the algorithm consists in moving from the current solution to a new solution in its neighbourhood, so that the value of the fitness function is increased. The way to choose an improving neighbour, if there are several of them, will not matter in this paper. The algorithm continues until a local optimum is reached. Let m be the number of different fitness values attained by solutions from $\mathcal{X} \setminus \mathcal{LO}$ plus 1. Then starting from any point, the local search method finds a local optimum within at most $m - 1$ steps.

Alternatively, one can use GAs to solve the optimisation problem, and possibly find local optima. The following result provides sets of sufficient conditions for a performance guaranteed GA to find local optima.

Corollary 13. *Given some positive constants p_0, ε_0 and δ , define the following conditions:*

(X1) $p_{\text{mut}}(y \mid x) \geq s$ for any $x \in \mathcal{X}$, $y \in \mathcal{N}(x)$.

(X2) $p_{\text{mut}}(x \mid x) \geq p_0$ for all $x \in \mathcal{X}$.

(X3) $p_{\text{xor}}(f(x') \geq \max\{f(u), f(v)\} \mid u, v) \geq \varepsilon_0$ for any $u, v \in \mathcal{X}$.

(X4.1) *the non-elitist GA in Algorithm 2 is set with $p_c = 1$, and it uses one of the following selection mechanisms:*

- *k-tournament selection with $k \geq \frac{4(1+\delta)}{\varepsilon_0 p_0}$,*
- *(μ, λ) -selection with $\frac{\lambda}{\mu} \geq \frac{(1+\delta)}{\varepsilon_0 p_0}$,*
- *exponential ranking selection with $\eta \geq \frac{4(1+\delta)}{\varepsilon_0 p_0}$.*

(X4.2) *the non-elitist GA is set with $p_c = 1 - \Omega(1)$, and it uses one of the following selection mechanisms: k-tournament selection, (μ, λ) -selection, linear or exponential ranking selection, with their parameters k , λ/μ and η being set to no less than $\frac{(1+\delta)}{(1-p_c)p_0}$.*

If (X1-3) and (X4.1) hold, or exclusively (X1-2) and (X4.2) hold, then there exists a constant c , such that for $\lambda \geq c \ln\left(\frac{m}{s}\right)$, a local optimum is reached for the first time after $\mathcal{O}\left(m\lambda \ln \lambda + \frac{m}{s}\right)$ fitness evaluations in the expectation.

Condition (X4.1) or (X4.2) characterises the setting of selection mechanisms, while (X1-3) bear the properties of the variation operators over the neighbourhood structure \mathcal{N} . Particularly, (X1) assumes a lower bound s on the probability that the mutation operator transforms an input solution into a specific

neighbour. To illustrate this condition, we note that in most of the local search algorithms the neighbourhood $\mathcal{N}(x)$ may be enumerated in polynomial time of the problem input size. For such neighbourhood mappings, a mutation operator that generates the uniform distribution over $\mathcal{N}(x)$ will select any given point in $\mathcal{N}(x)$ with probability at least s , so that $1/s$ is polynomially bounded in the problem input size.

If crossover is frequently used, i. e. $p_c = 1$, we also need to satisfy condition (X3) on the crossover operator. It requires that the fitness of solution x on the output of crossover is not less than the fitness of parents with probability at least ε_0 . Note that such a requirement is satisfied with $\varepsilon_0 = 1$ for the optimized crossover operators, where the offspring is computed as a solution to the *optimal recombination problem* (see e.g [23]). This supplementary problem is known to be polynomially solvable for Maximum Clique [4], Set Packing, Set Partition and some other NP-hard problems [23].

When a set of conditions, i. e. depending on whenever $p_c = 1$ or $p_c = 1 - \Omega(1)$, is satisfied and given a sufficiently large population w. r. t. to m and s , an upper bound on the expected number of fitness evaluations that GA performs to find a local optimum is guaranteed. The proof directly follows from Corollaries 6 and 5 of Theorem 1.

Proof of Corollary 13. We use the following partition

$$A_j := \{x \in \mathcal{X} | f(x) = f_j\} \setminus \mathcal{LO}, \quad j \in [m-1], \text{ and} \\ A_m := \mathcal{LO}.$$

We note that (A_1, \dots, A_{m-1}) is a fitness-based partition of $\mathcal{X} \setminus \mathcal{LO}$. Thus, applications of Corollary 5 and Lemma 7 for the set of conditions (X1-2) and (X4.2), or alternatively, Corollary 6 and Lemma 8 for the set of conditions (X1-3) and (X4.1), yield the required result. \square

A similar result for GAs with very high selection pressure was obtained in [25]. In particular, the result from [25] implies that a GA with tournament selection or (μ, λ) -selection, given certain settings of parameters, reaches a local optimum after $\mathcal{O}(m \ln(m)/s)$ fitness evaluations in expectation. The upper bound from Corollary 13 has an advantage to the bound from [25] if $1/s$ is at least linear in m .

The effect of the corollary can be seen in the following example setting. Let us consider the binary search space $\{0, 1\}^n$ with Hamming neighbourhood mapping of a constant radius r , a fitness function f such that $m \in \text{poly}(n)$, and assume that GA uses bitwise mutation operator and $p_c = 1 - \Omega(1)$. The bitwise mutation operator outputs a string y , given a string x , with probability $p_m^{H(x,y)}(1 - p_m)^{n-H(x,y)}$. Note that probability $p_m^j(1 - p_m)^{n-j}$, as a function of p_m attains its maximum at $p_m = j/n$. It is easy to show (see e.g. [25]) that for any $x \in \mathcal{X}$ and $y \in \mathcal{N}(x)$, the bitwise mutation operator with $p_m = r/n$ satisfies the condition $p_{\text{mut}}(y | x) = \mathcal{O}(1/n^r)$. Besides that, for a sufficiently large n and any $x \in \mathcal{X}$ holds $p_{\text{mut}}(x | x) \geq e^{-r}/2 = \Omega(1)$. Therefore, Corollary 13 implies that a GA with the above mentioned operators given appropriate settings of parameters λ, p_m and p_c , first visits a local optimum w. r. t. a Hamming neighbourhood of constant radius after a polynomially bounded number of fitness evaluations in expectation.

To give concrete examples, we consider the following unconstrained (and unweighted) problems:

- MAX-SAT: given a CNF formula in n logical variables which is represented by m' clauses $\mathbf{c}_1, \dots, \mathbf{c}_{m'}$ and each clause is a disjunction of logical variables or their negations, it is required to find an assignment of the variables so that the number of satisfied clauses is maximised.
- MAX-CUT: given an undirected graph $G = (V, E)$, it is required to find a partition of V into two sets $(S, V \setminus S)$, so that the number of crossing edges, i.e. $\delta(S) := |\{(u, v) \mid (u, v) \in E, u \in S, v \notin S\}|$, is maximised.

Both problems are NP-hard, and their solutions can be naturally represented by bitstrings. Particularly, any local optimum w.r.t. the neighbourhood defined by Hamming distance 1 has at least half the optimal fitness [1]. Better approximation ratios can be obtained with more sophisticated algorithms, e.g. 0.79-approximation for MAX-SAT [2] based on a time-consuming semi-definite programming relaxation. The local search algorithm with the above neighbourhood however has an advantage of low time complexity, e.g. it only makes $\mathcal{O}(nm')$ tentative solutions for MAX-SAT. Corollary 13 translates such a result into relatively low runtime bound for GAs.

Theorem 14. *Suppose the GA in Algorithm 2 is applied to MAX-SAT or to MAX-CUT using a bitwise mutation with $p_m = \chi/n$, where $\chi > 0$ is a constant, a crossover with $p_c = 1 - \Omega(1)$ and one of the selection mechanisms: k -tournament selection, (μ, λ) -selection, linear or exponential ranking selection, with their parameters k , λ/μ and η being set to no less than $\frac{(1+\delta)e^\chi}{1-p_c}$, where $\delta > 0$ is any constant. Then there exists a constant c , such that for $\lambda \geq c \ln(nm')$, a 1/2-approximate solution is reached for the first time after $\mathcal{O}(m'\lambda \ln \lambda + nm')$ fitness evaluations in expectation for MAX-SAT, and after $\mathcal{O}(|E|\lambda \ln \lambda + |V| |E|)$ for MAX-CUT.*

The proof is analogous to the analysis of ℓ -UNIMODAL function in Theorem 9, combined with Corollary 13 where m is no more than $m' + 1$ for MAX-SAT and no more than $|E| + 1$ for MAX-CUT.

5 Estimation of Distribution Algorithms

As mentioned in the introduction, there are few rigorous runtime results for UMDA and other estimation of distribution algorithms (EDAs). The analytic techniques used in previous analyses of EDAs were often complex, e.g. relying on the machinery of Markov chain theory. Surprisingly, even apparently simple problems, such as the expected runtime of UMDA on ONEMAX, were until recently open.

Algorithm 1 matches closely the typical behaviour of estimation of distribution algorithms: given a current distribution over the search space, sample a finite number of search points, and update the probability distribution. We demonstrate the ease at which the expected runtime of UMDA with margins and truncation selection on the ONEMAX function can be obtained using the level-based theorem without making any simplifying assumptions about the optimisation process.

5.1 Algorithm

If $P \in \mathcal{X}^\lambda$ is a population of λ solutions, let $P(k, i)$ denote the value in the i -th bit position of the k -th solution in P . The Univariate Marginal Distribution Algorithm (UMDA) with (μ, λ) -truncation selection is defined in Algorithm 4.

The algorithm has three parameters, the parent population size μ , the offspring population size λ , and a parameter $m' < \mu$ controlling the size of the margins. It is necessary to set $m' > 0$ to prevent a premature convergence, e.g. without this margin $p_t(i)$ can go to a non-optimal fixation, this prevents further exploration and causes an infinite runtime. Based on insights about optimal mutation rates in the (1+1) EA, we will use the parameter setting $m' = \mu/n$ in the rest of this section.

It is immediately clear that the UMDA in Algorithm 4 is a special case of Algorithm 1 scheme. The probability distribution $D(P_t)$ of y is computed in steps 6-7, and is defined for any search point $x \in \{0, 1\}^n$ by

$$\Pr(y = x) = \prod_{j=1}^n p_t(j)^{x_j} (1 - p_t(j))^{1-x_j}.$$

Algorithm 4 UMDA

- 1: Initialise the vector $p_0 := (1/2, \dots, 1/2)$.
- 2: **for** $t = 1, 2, 3, \dots$ **do**
- 3: **for** $x = 1$ to λ **do**
- 4: Sample the x -th individual $P_t(x, \cdot)$ according to

$$P_t(x, i) \sim \text{Bernoulli}(p_{t-1}(i)) \text{ for all } i \in [n].$$

- 5: **end for**
- 6: Sort the population P_t according to f .
- 7: Calculate a new vector p_t from P_t according to

$$p_t(i) := \begin{cases} \frac{m'}{\mu} & \text{if } X_i < m' \\ \frac{X_i}{\mu} & \text{if } m' \leq X_i \leq \mu - m' \\ 1 - \frac{m'}{\mu} & \text{if } \mu - m' < X_i, \end{cases}$$

for all $i \in [n]$, where $X_i := \sum_{k=1}^{\mu} P_t(k, i)$.

- 8: **end for**
-

Note that the sampling from vector $(p_i)_{i \in [n]}$ in UMDA is analogous to a population-wise crossover, i.e. the bit sampling at position i with a non-marginal probability (line 4 in Algorithm 4) is equivalent to picking uniformly at random a bit from the μ bits at position i of the μ selected individuals of the previous generation. In some other randomised search heuristics such as ant colony optimisation (ACO) and compact genetic algorithms (cGA), the sampling distribution D_t does not only depend on the current population, but also on additional information, such as pheromone values. The level-based theorem does not apply to such algorithms.

It is well-known that the (1+1) EA solves ONEMAX problem in expected time $\Theta(n \ln n)$, and this is optimal for the class of unary, unbiased black-box algorithms. Surprisingly, no previous runtime analysis of UMDA seems available for ONEMAX. We demonstrate that the expected runtime can be obtained relatively easy with our methods. To obtain lower bounds on the tail of the level-distribution, we make use of the Feige inequality [26] (or see Lemma 20 in the appendix).

Theorem 15. *Given any positive constants $\delta \in (0, 1)$, and $\gamma_0 \leq \frac{1}{(1+\delta)13e}$, the UMDA with offspring population size λ with $b \ln(n) \leq \lambda \leq n/\gamma_0$ for some constant $b > 0$, parent population size $\mu = \gamma_0 \lambda$ and margins $m' = \mu/n$, has expected optimisation time $\mathcal{O}(n \lambda \ln \lambda)$ on ONEMAX.*

Proof. Step 1: We use the canonical partition into $m = n + 1$ levels, where level $j \in [m]$ is defined by

$$A_j := \{x \in \{0, 1\}^n \mid \text{ONEMAX}(x) = j - 1\}.$$

We use the parameter $\gamma_0 := \mu/\lambda$ and let Y be the number of one-bits in the sampled solution.

The choice $m' = \mu/n$ and $\mu \leq n$ implies that the margins for $p_t(i)$ are simplified to $1/n$ and $1 - 1/n$, and that these margins are only used when the bit values at position i of the μ selected individuals are identical. We categorise the probabilities $p_t(i)$ into three groups: those at the upper margin $1 - 1/n$, those at the lower margin $1/n$, and intermediary values in the closed interval $[1/\mu, 1 - 1/\mu]$. Due to linearity of the fitness function, the components of p_t can be rearranged without changing the distribution of Y . We assume w.l.o.g. a rearrangement so that there exists integers $k, \ell \geq 0$ satisfying

$$\begin{aligned} 1 \leq X_i < \mu & \quad \text{and} \quad p_t(i) = X_i/\mu & \quad \text{if } 1 \leq i \leq k, \\ X_i = \mu & \quad \text{and} \quad p_t(i) = 1 - 1/n & \quad \text{if } k < i \leq k + \ell, \text{ and} \\ X_i = 0 & \quad \text{and} \quad p_t(i) = 1/n & \quad \text{if } k + \ell < i \leq n. \end{aligned}$$

By these assumptions, it follows that

$$\sum_{i=k+1}^{k+\ell} X_i = \mu \ell \quad \text{and} \quad \sum_{i=k+\ell+1}^n X_i = 0. \quad (17)$$

In the following, we define $Y_{i,k}$ to be the number of sampled one-bits due to $(p_t(i), \dots, p_t(k))$ in the rearranged p_t .

For any population P_t and any $\gamma \in [0, \gamma_0]$, let $j \in [n]$ be any integer such that $|P_t \cap A_{\geq j}| \geq \gamma_0 \lambda = \mu$ and $|P_t \cap A_{\geq j+1}| \geq \gamma \lambda$. This implies that among the μ fittest individuals in the current population, there are at least $\gamma \lambda$ individuals with at least j one-bits, and the remaining among the μ fittest individuals have at least $j - 1$ one-bits. Hence, the total number of one-bits among the fittest μ individuals must satisfy

$$\sum_{i=1}^n X_i \geq \gamma \lambda j + (\mu - \gamma \lambda)(j - 1) = \gamma \lambda + \mu(j - 1). \quad (18)$$

Combining Eqs. (17) and (18), when $k \geq 1$, we get

$$E[Y_{1,k}] = \sum_{i=1}^k p_t(i) = \frac{1}{\mu} \sum_{i=1}^k X_i \geq \frac{\gamma\lambda}{\mu} + j - 1 - \ell. \quad (19)$$

Step 2: We first verify condition (G2), i.e. checking if $\Pr(Y \geq j) \geq (1 + \delta)\gamma$ for any level j defined like above with $\gamma > 0$. We distinguish between two cases, either $k = 0$ or $k \geq 1$.

Case 1: If $k \geq 1$, then Eq. (19) and Lemma 20 give

$$\begin{aligned} \Pr(Y_{1,k} \geq j - \ell) &= \Pr\left(Y_{1,k} > j - 1 - \ell + \frac{\gamma\lambda}{\mu} - \frac{\gamma\lambda}{12\mu}\right) \\ &\geq \Pr\left(Y_{1,k} > E[Y_{1,k}] - \frac{\gamma\lambda}{12\mu}\right) \\ &\geq \min\left\{\frac{1}{13}, \frac{\frac{\gamma\lambda}{12\mu}}{\frac{\gamma\lambda}{12\mu} + 1}\right\} \\ &= \min\left\{\frac{1}{13}, \frac{\gamma\lambda}{\gamma\lambda + 12\mu}\right\} \geq \frac{\gamma\lambda}{13\mu}. \end{aligned}$$

The probability of sampling an individual with at least j one-bits in the next generation is therefore lower-bounded by

$$\begin{aligned} \Pr(Y \geq j) &\geq \Pr(Y_{1,k} \geq j - \ell) \Pr(Y_{k+1,k+\ell} = \ell) \\ &\geq \frac{\gamma\lambda}{13\mu} \left(1 - \frac{1}{n}\right)^\ell \geq \frac{\gamma\lambda}{13\mu} \left(1 - \frac{1}{n}\right)^{n-1} \\ &\geq \frac{\gamma\lambda}{13e\mu} \geq (1 + \delta)\gamma. \end{aligned}$$

Case 2: If $k = 0$, then all the μ best individuals in the population must be identical. By assumption, there are $\gamma\lambda \geq 1$ individuals with at least j 1-bits, hence all the μ best individuals must have at least j 1-bits. In this case, there are $\ell \geq j$ probabilities at the upper margin, and we get

$$\begin{aligned} \Pr(Y \geq j) &\geq \Pr\left(Y_{k+1}^{k+j} \geq j\right) \\ &= \left(1 - \frac{1}{n}\right)^j \geq \frac{1}{e} \geq 13\gamma_0(1 + \delta) > (1 + \delta)\gamma, \end{aligned}$$

and condition (G2) is therefore satisfied also in this case.

Step 3: We now consider condition (G1) for any j defined with $\gamma = 0$. Again we check the two cases $k = 0$, and $k \geq 1$.

Case 1: If $k = 0$, then with our assumption, the $\ell \geq j - 1$ first probabilities are at the upper margin $1 - 1/n$, and the last $n - \ell \leq n - j + 1$ probabilities are at the lower margin $1/n$. In order to obtain a search point with at least j one-bits, it is sufficient to sample exactly $\ell \geq j - 1$ one-bits in the first ℓ positions and exactly one 1-bit in the last $n - \ell \leq n - j + 1$ positions. Hence,

$$\Pr(Y \geq j) \geq \Pr(Y_{1,\ell} \geq \ell) \Pr(Y_{\ell+1,n} \geq 1)$$

$$\begin{aligned}
&\geq \left(1 - \frac{1}{n}\right)^\ell \binom{n-\ell}{n} \\
&\geq \left(1 - \frac{1}{n}\right)^{n-1} \binom{n-j+1}{n} \geq \binom{n-j+1}{en}.
\end{aligned}$$

Case 2: When $k \geq 1$, we note from Eq. (19) that

$$E[Y_{2,k}] = E[Y_{1,k}] - p_t(1) \geq j - 1 - \ell - p_t(1)$$

Again, by Lemma 20 we get

$$\begin{aligned}
&\Pr(Y_{2,k} \geq j - 1 - \ell) \\
&= \Pr(Y_{2,k} > j - 1 - \ell - p_t(1) - (1 - p_t(1))) \\
&\geq \Pr(Y_{2,k} > E[Y_{2,k}] - (1 - p_t(1))) \\
&\geq \min\left\{\frac{1}{13}, \frac{1 - p_t(1)}{2 - p_t(1)}\right\} > \frac{1 - p_t(1)}{13}.
\end{aligned}$$

The probability of sampling an individual with at least j one-bits in this configuration is bounded from below as

$$\begin{aligned}
&\Pr(Y \geq j) \\
&> \Pr(Y_1 = 1) \Pr(Y_{2,k} \geq j - 1 - \ell) \Pr(Y_{k+1,k+\ell} = \ell) \\
&\geq \left(\frac{p_t(1)(1 - p_t(1))}{13}\right) \left(1 - \frac{1}{n}\right)^\ell \\
&\geq \left(\frac{(1/\mu)(1 - 1/\mu)}{13}\right) \left(1 - \frac{1}{n}\right)^\ell \geq \frac{1}{14e\mu}.
\end{aligned}$$

The last inequality holds for $\mu \geq 14$, which in turn only requires n to be larger than some constant. Hence, combining the cases $k = 0$ and $k > 0$, for all $j \in [n]$ we get

$$\begin{aligned}
\Pr(Y \geq j) &\geq \min\left\{\frac{1}{14e\mu}, \frac{n-j+1}{en}\right\} \\
&\geq \frac{n-j+1}{14e\mu(n-j+1) + en} =: z_j.
\end{aligned}$$

Clearly, there exists a z_* with $1/z_* \in \text{poly}(n)$ such that $\Pr(Y \geq j) \geq z_*$ for all $j \in [n]$ and condition (G1) is satisfied.

Step 4: We consider condition (G3) regarding the population size. The parameters δ and $\gamma_0 = \mu/\lambda$ are constants with respect to n , therefore the variables a, ε and c in condition (G3) are also constants, and $1/z_* \in \text{poly}(n)$. Hence, there must exist a constant $b > 0$ such that condition (G3) is satisfied when $\lambda \geq b \log(n)$.

Step 5: To conclude, the expected optimisation time is

$$\begin{aligned}
&\mathcal{O}\left(n\lambda \ln(\lambda) + \sum_{j=1}^n \frac{1}{z_j}\right) \\
&= \mathcal{O}\left(n\lambda \ln(\lambda) + 14e\mu n + \sum_{i=0}^{n-1} \frac{en}{n-i}\right) = \mathcal{O}(n\lambda \ln \lambda). \quad \square
\end{aligned}$$

A similar analysis for LEADINGONES [15] yields an upper bound on the runtime of $\mathcal{O}(n\lambda \ln \lambda + n^2)$ with offspring population size $\lambda \geq b \ln(n)$ for some constant $b > 0$ without use of Feige's inequality. The previous result [9] on LEADINGONES requires a larger population size and gives a longer runtime bound.

Table 1 summarises the runtime bounds for the example applications of the tools presented in this paper and the above mentioned result for UMDA on LEADINGONES.

Table 1: Summary of results for GA (Algo. 2 with $p_c = 1 - \Omega(1)$), GA1 (Algo. 2 with $p_c = 1$) and UMDA (Algo. 4 with margin $m' = \mu/n$).

RUNTIME RESULT				CONFIGURATION				
Problem	Algorithm	Min. λ	Runtime	Alg.	Recomb.	Selection	Setting	
ONEMAX	GA, GA1	$c \ln n$	$\mathcal{O}(n\lambda)$	GA	any	k -tournament	$k \geq \frac{(1+\delta)e^x}{1-p_c}$	
	UMDA	$c \ln n$	$\mathcal{O}(n\lambda \ln \lambda)$		any	(μ, λ) -selection	$\frac{\lambda}{\mu} \geq \frac{(1+\delta)e^x}{1-p_c}$	
LEADINGONES	GA, GA1, UMDA	$c \ln n$	$\mathcal{O}(n^2 + n\lambda \ln \lambda)$		any	linear ranking	$\eta \geq \frac{(1+\delta)e^x}{1-p_c}$	
ℓ -UNIMODAL	GA	$c \ln(n\ell)$	$\mathcal{O}(n\ell + \ell\lambda \ln \lambda)$		any	exp. ranking	$\eta \geq \frac{(1+\delta)e^x}{1-p_c}$	
LINEAR	GA	$c \ln n$	$\mathcal{O}(n^2 + n\lambda \ln \lambda)$	GA1	mask-based	k -tournament	$k \geq 8(1 + \delta)e^x$	
JUMP $_r$	GA	$cr \ln n$	$\mathcal{O}\left(\left(\frac{n}{\chi}\right)^r + n\lambda + \lambda \ln \lambda\right)$		mask-based	(μ, λ) -selection	$\frac{\lambda}{\mu} \geq 2(1 + \delta)e^x$	
ROYALROAD $_{r \geq 2}$	GA	$cr \ln n$	$\mathcal{O}\left(\left(\frac{n}{\chi}\right)^r \ln\left(\frac{n}{r}\right) + \frac{n\lambda \ln \lambda}{r}\right)$		mask-based	exp. ranking	$\eta \geq 8(1 + \delta)e^x$	
SORTING _{INV}	GA	$c \ln n$	$\mathcal{O}(n^2\lambda)$	UMDA	n/a	(μ, λ) -selection	$\frac{\lambda}{\mu} \geq 13(1 + \delta)e$	
$\frac{1}{2}$ -approx. MAX-SAT	GA	$c \ln(nm')$	$\mathcal{O}(nm' + m'\lambda \ln \lambda)$					
$\frac{1}{2}$ -approx. MAX-CUT	GA	$c \ln(V E)$	$\mathcal{O}(V E + E \lambda \ln \lambda)$					

On $\{0, 1\}^n$, GA and GA1 use bitwise mutation operator with rate χ/n where χ is any constant. On permutation search space, i.e. Sorting, GA uses Exchange mutation and its setting assumes $\chi = 1$. In the case of MAX-SAT, n is the number of logical variables and m' is the number of clauses. Parameter δ is any positive constant, and c is some constant.

6 The level-based theorem is almost tight

How accurate are the time bounds provided by the level-based theorem? To answer this question, we first interpret the theorem as a universally quantified statement over the operators D satisfying the conditions of the theorem. More formally, given a choice of level-partitioning and set of parameters $z_1, \dots, z_{m-1}, \delta, \gamma_0$, which we collectively denote by Θ , the theorem can be expressed on the form

$$\forall D \in \mathcal{D}_\Theta \ E [T_D] \leq t_\Theta, \quad (20)$$

where \mathcal{D}_Θ is the set of operators D in Algorithm 1 that satisfy the conditions of the level-based theorem with parameterisation Θ , $E [T_D]$ is the expected running time of Algorithm 1 with a given operator D , and t_Θ is the upper time bound provided by the level-based theorem which depends on the parameterisation Θ .

In order to obtain an accurate bound for a specific operator D , for example the (μ, λ) EA applied to the ONEMAX function, it is necessary to choose a parameterisation Θ that reflects this process as tightly as possible. If the bounds on the “upgrade” probabilities z_j for the (μ, λ) EA are too small, then the class \mathcal{D}_Θ includes other processes which are slower than the (μ, λ) EA, and the corresponding bound t_Θ cannot be accurate. Hence, the theorem is limited by the accuracy at which one can describe the process by some class \mathcal{D}_Θ . One could imagine a more accurate variant of the theorem requiring more precise, and possibly harder to obtain, information about the process, such as the variance of $D(P_t)$.

Assuming a fixed parameterisation Θ , it is possible to make a precise statement about the tightness of the upper bound t_Θ . Theorem 16 stated below is an existential statement on the form

$$\exists D \in \mathcal{D}_\Theta \ E [T_D] \geq t'_\Theta, \quad (21)$$

where the lower bound t'_Θ is close to the upper bound t_Θ . Hence, given the information the theorem has about the process through the parameterisation Θ , the runtime bound is close to optimal. More information about the process would be required to obtain a more accurate bound on the runtime.

In some concrete cases, one can prove that the level-based theorem is close to optimal using parallel black-box complexity theory [3, 22]. From Corollary 5 with $p_c = 0$, which specialises the level-based theorem to algorithms with unary mutation operators, one can obtain the bounds $\mathcal{O}(n\lambda + n \ln n)$ for ONEMAX, and $\mathcal{O}(n\lambda \ln(\lambda) + n^2)$ for LEADINGONES for appropriately parameterised EAs. These bounds are within a $\mathcal{O}(\ln \lambda)$ -factor of the lower bounds that hold for any parallel unbiased black-box algorithm [3]. For population sizes $\lambda = \mathcal{O}(n/\ln n)$ and $\lambda = \Omega(\ln n)$, the resulting $\mathcal{O}(n^2)$ bound on LEADINGONES is asymptotically tight, because it matches the lower bound that holds for all black-box algorithms with unary unbiased variation operators [40].

Theorem 16. *Given any partition of \mathcal{X} into m non-empty subsets (A_1, \dots, A_m) , for any $z_1, \dots, z_{m-1}, \delta, \gamma_0 \in (0, 1)$ where $1 \geq \gamma_0(1 + \delta) \geq z_j$ for all $j \in [m - 1]$, and $\lambda \in \mathbb{N}$, there exists a mapping D which satisfies conditions (G1), (G2), and (G3), of Theorem 1, such that Algorithm 1 with mapping D has expected hitting*

time

$$E[T] \geq \left(\frac{2}{3\delta} \sum_{j=1}^{m-2} \lambda \ln \left(\frac{\gamma_0 \lambda}{1 + 2\lambda z_j + 1/\delta^2} \right) \right) + \sum_{j=1}^{m-1} \frac{1}{z_j},$$

where $T := \min\{\lambda t \in \mathbb{N} \mid |P_t \cap A_m| > 0\}$.

Proof of Theorem 16. We construct an operator D which leads to the claimed lower bound. Choose any sequence of search points $(x_1, \dots, x_m) \in A_1 \times \dots \times A_m$, and let the initial population of Algorithm 1 be $P_0 := (x_1, \dots, x_1)$, i.e., λ copies of the search point x_1 belonging to the first level.

For any population $P \in \mathcal{X}^\lambda$, let the *current level* be the largest $i \in [m]$ such that $|P \cap A_{\geq i}| \geq \gamma_0 \lambda$. For any population P with current level $i < m$, define the operator D for all $u \in \mathcal{X}$ by

$$\Pr_{y \sim D(P)}(y = u) := \begin{cases} 1 - \max\{(1 + \delta)\gamma, z_i\} & \text{if } u = x_i \\ \max\{(1 + \delta)\gamma, z_i\} & \text{if } u = x_{i+1}, \\ 0 & \text{otherwise.} \end{cases} \quad (22)$$

where $\gamma := (1/\lambda)|P \cap A_{\geq i+1}| < \gamma_0$.

For all $t \in \mathbb{N}$, define

$$\begin{aligned} T_j &:= \min\{t \mid |P_t \cap A_{\geq j}| > 0\}, \text{ for all } j \in [m], \text{ and} \\ S_j &:= T_{j+1} - T_j \text{ for all } j \in [m-1]. \end{aligned}$$

Then we have $\sum_{j=1}^{m-1} S_j = T_m - T_1 = T$ because $T_1 = 0$. The random variable S_j , for $j \in [m-1]$, describes the number of generations from the time the process has discovered the search point x_j until it has discovered the search point x_{j+1} , and we call this *phase j* . We divide each phase j into two sub-phases. Let S_j^1 be the number of generations where

$$1 \leq |P_t \cap A_{\geq j}| < \gamma_0 \lambda,$$

and call this the first sub-phase, and let S_j^2 be the number of generations where

$$\gamma_0 \lambda \leq |P_t \cap A_{\geq j}| \text{ and } 0 = |P_t \cap A_{\geq j+1}|,$$

and call this the second sub-phase. The duration of the j -th phase is the sum $S_j = S_j^1 + S_j^2$. Remark that $S_1^1 = 0$ due to the choice of the initial population P_0 .

Note also that by the definition of operator D , as long as the process is in sub-phase 1 of phase j , the probability of generating the search point x_{j+1} is 0. Furthermore, the process never returns to sub-phase 1 once the process has entered sub-phase 2. To estimate the duration of sub-phase 1, we consider the stochastic process $(X_t)_{t \in \mathbb{N}}$ where $X_t := |P_{T_j+t} \cap A_{\geq j}|$, and a corresponding filtration $(\mathcal{F}_t)_{t \in \mathbb{N}}$ where $\mathcal{F}_t := \sigma(P_1, \dots, P_{T_j+t})$.

During sub-phase 1 of phase $j > 1$, it holds that $X_{t+1} \sim \text{Bin}(\lambda, p_{t+1})$, where $p_{t+1} = \max\{(1 + \delta)\frac{X_t}{\lambda}, z_{j-1}\}$.

To lower bound the expected duration of sub-phase 1, we apply drift analysis (Lemma 28) with respect to the process $(Z_t)_{t \in \mathbb{N}}$ defined by $Z_t := \ln(\lambda/R_t)$ where

$R_t := \max\{X_t, y_j\}$ and $y_j := \max\{\lambda z_{j-1}, 1/\delta^2\} > 1$. Note that since $z_j < \gamma_0$ by assumption, and $1/\delta^2 < \gamma_0 \lambda$ by condition (G3), it holds that $y_j < \gamma_0 \lambda$. It is therefore clear that sub-phase 1 is only complete if

$$Z_t \leq \ln\left(\frac{\lambda}{\gamma_0 \lambda}\right) = -\ln(\gamma_0) =: a.$$

By Jensen's inequality, the drift of this process can be bounded by

$$\begin{aligned} E[Z_t - Z_{t+1} \mid \mathcal{F}_t] &= E\left[\ln\left(\frac{R_{t+1}}{R_t}\right) \mid \mathcal{F}_t\right] \\ &\leq \ln\left(\frac{E[R_{t+1} \mid \mathcal{F}_t]}{R_t}\right) \end{aligned}$$

and by Lemma 29

$$\leq \ln\left(\frac{\max(\lambda p_{t+1}, y_j) + (\frac{1}{2})\sqrt{\lambda p_{t+1}}}{R_t}\right).$$

When $\lambda p_{t+1} \geq y_j$, we use that $R_t = \max\{X_t, y_j\} \geq \lambda p_{t+1}/(1 + \delta)$ because $p_{t+1} = \max\{X_t(1 + \delta)/\lambda, z_{j-1}\}$, so

$$\begin{aligned} E[Z_t - Z_{t+1} \mid \mathcal{F}_t] &\leq \ln\left(\frac{\lambda p_{t+1} + (1/2)\sqrt{\lambda p_{t+1}}}{\lambda p_{t+1}/(1 + \delta)}\right) \\ &= \ln\left((1 + \delta) \left(1 + \frac{1}{2\sqrt{\lambda p_{t+1}}}\right)\right). \end{aligned}$$

Since $\lambda p_{t+1} \geq y_j > 1$, thus $\sqrt{1/(\lambda p_{t+1})} \leq \sqrt{1/y_j} \leq \delta$ and

$$E[Z_t - Z_{t+1} \mid \mathcal{F}_t] \leq \ln(1 + \delta) + \ln(1 + \delta/2) < (3/2)\delta.$$

Otherwise, when $y_j > \lambda p_{t+1}$, we use that $R_t \geq y_j$ and get

$$\begin{aligned} E[Z_t - Z_{t+1} \mid \mathcal{F}_t] &\leq \ln\left(\frac{y_j + (1/2)\sqrt{\lambda p_{t+1}}}{y_j}\right) \\ &< \ln\left(1 + \frac{\sqrt{y_j}}{2y_j}\right) = \ln\left(1 + \frac{1}{2\sqrt{y_j}}\right) \\ &\leq \ln\left(1 + \frac{\delta}{2}\right) < \delta/2. \end{aligned}$$

Hence, condition 1 in Lemma 28 can be satisfied with the parameter $\varepsilon := (3/2)\delta$. We therefore get the bound

$$E[S_j^1 \mid \mathcal{F}_0] \geq \frac{Z_0 - a}{\varepsilon} = \left(\frac{2}{3\delta}\right) \ln\left(\frac{\gamma_0 \lambda}{\max\{X_0, \lambda z_{j-1}, \frac{1}{\delta^2}\}}\right).$$

By the definition of the process, for $1 < j \leq m$, we have $X_0 \sim (Y \mid Y \geq 1)$ where $Y \sim \text{Bin}(\lambda, z_{j-1})$, i.e., X_0 is binomially distributed random variable conditional on having value at least 1. and by the tower property of expectation,

$$E[S_j^1] = E[E[S_j^1 \mid \mathcal{F}_0]]$$

$$\begin{aligned}
&\geq E \left[\left(\frac{2}{3\delta} \right) \ln \left(\frac{\gamma_0 \lambda}{\max\{X_0, \lambda z_{j-1}, 1/\delta^2\}} \right) \right] \\
&> E \left[\left(\frac{2}{3\delta} \right) \ln \left(\frac{\gamma_0 \lambda}{X_0 + \lambda z_{j-1} + 1/\delta^2} \right) \right]
\end{aligned}$$

and since the function $f(x) = \ln(1/x)$ is convex, Jensen's inequality and Lemma 30 give

$$> \left(\frac{2}{3\delta} \right) \ln \left(\frac{\gamma_0 \lambda}{1 + 2\lambda z_{j-1} + 1/\delta^2} \right).$$

During sub-phase 2, it holds that

$$\Pr_{y \sim D(P_t)}(y = x_j) = 1 - z_j, \text{ and } \Pr_{y \sim D(P_t)}(y = x_{j+1}) = z_j.$$

In each generation of sub-phase 2, the phase ends with probability $q_j := 1 - (1 - z_j)^\lambda < \lambda z_j$, i.e., the probability that at least one individual is produced in $A_{\geq j+1}$. The duration of sub-phase 2 is therefore geometrically distributed with parameter q_j and has expectation $E[S_j^2] = 1/q_j \geq 1/(\lambda z_j)$.

Hence, we get

$$\begin{aligned}
E[T] &= \sum_{j=1}^{m-1} E[S_j^1] + E[S_j^2] \\
&\geq \left(\frac{2}{3\delta} \sum_{j=1}^{m-2} \ln \left(\frac{\gamma_0 \lambda}{1 + 2\lambda z_j + 1/\delta^2} \right) \right) + \sum_{j=1}^{m-1} \frac{1}{\lambda z_j}. \quad \square
\end{aligned}$$

7 Conclusion

Time-complexity analysis of evolutionary algorithms (EAs) has advanced significantly over the last decade, starting from simplified settings such as variants of the (1+1) EA without a real population, crossover or other higher-arity operators. It has been unclear to what extent the time-complexity profiles of these simple EAs considered by theoreticians deviate from those of the more sophisticated, population-based EAs often preferred by practitioners. New techniques tailored to time-complexity of population-based algorithms are required.

This paper introduces a new technique that easily yields upper bounds on the expected runtime of complex, non-elitist search processes. The technique is first illustrated on Genetic Algorithms. We have shown that GAs optimise standard benchmark functions, as well as combinatorial optimisation problems, efficiently. As long as the population size is not overly large, the population does not incur an asymptotic slowdown on these functions compared to standard EAs that do not use populations. Thus, speedups can be achieved by parallellising fitness evaluations. Furthermore, consequent work indicate that non-elitist, population-based EAs have an advantage on more complex problems, including those with noisy [14], dynamic [13], and peaked [17] fitness landscapes.

As a side-effect of the analysis, the conditions of level-based theorem yield settings for algorithmic parameters, such as population size, mutation and crossover rates, selection pressure etc., that are sufficient to guarantee a given

time-complexity bound. This opens up the possibility of theory-led design of EAs with guaranteed runtime, where the algorithm is designed to satisfy the conditions of the level-based theorem [11].

Further demonstrating the generality of the theorem, we also provide time-complexity results for the UMDA algorithm, an Estimation of Distribution Algorithm, for which there are few theoretical results. Finally, we show via lower bounds on the runtime of a concrete process, that given the information the theorem requires about the process, the upper bounds are close to tight, i.e., little improvement of the time bound in the theorem is possible.

Acknowledgment

The research was supported by the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement no 618091 (SAGE) and Russian Foundation for Basic Research grants 15-01-00785 and 16-01-00740. Early ideas were discussed at Dagstuhl Seminars 13271 and 15211 “Theory of Evolutionary Algorithms”.

Appendix A

The following results are known in the literature.

Lemma 17 (Lemma 33 in [16]). *For all $x \geq 0$, $x \geq \ln(1+x) \geq x(1-x/2)$.*

Lemma 18 (Lemma 31 in [16]). *For $n \in \mathbb{N}$ and $x \geq 0$, we have $1 - (1-x)^n \geq 1 - e^{-xn} \geq \frac{xn}{1+xn}$.*

Lemma 19 (Lemma 3 in [17]). *For any $\varepsilon \in (0, 1)$ and $\chi > 0$, if $n \geq (\chi + \varepsilon) \frac{\chi}{\varepsilon}$ then $(1 - \varepsilon)e^{-\chi} \leq (1 - \frac{\chi}{n})^n \leq e^{-\chi}$.*

Lemma 20 (Corollary 3 in [15]). *Let Y_1, \dots, Y_n be n independent random variables with support in $[0, 1]$ and finite expectations and $Y := \sum_{i=1}^n Y_i$. It holds for every $\delta > 0$ that*

$$\Pr(Y > E[Y] - \delta) \geq \min \left\{ \frac{1}{13}, \frac{\delta}{1 + \delta} \right\}.$$

Appendix B

The following lemmas are part of the proof of Theorem 1.

Lemma 21. *The functions g_1 and g_2 defined below are level functions for any $c > 0$, $\kappa \in (0, 1)$, $x \in [\lambda]$, $y \in [m]$ and $\gamma_j, q_j \in (0, 1]$ for each $j \in [m-1]$.*

$$g_1(x, y) := \ln \left(\frac{1 + c\lambda}{1 + c \max\{x, \gamma_y \lambda\}} \right) + \sum_{i=y+1}^{m-1} \ln \left(\frac{1 + c\lambda}{1 + c\gamma_i \lambda} \right)$$

$$g_2(x, y) := \begin{cases} \frac{(1 - \kappa)^x}{q_y} + \sum_{i=y+1}^{m-1} \frac{1}{q_i} & \text{if } y \in [m-1], \\ 0 & \text{if } y = m \end{cases}$$

and $g_1(x, j) := g_2(x, j) := 0$ for $j = m$.

Proof. Both g_1 and g_2 are non-increasing functions in x and y , hence properties 1 and 2 of Definition 3 are satisfied. Property 3 is satisfied because for all $y \in [m-1]$

$$\begin{aligned} g_1(\lambda, y) &= \sum_{i=y+1}^{m-1} \ln \left(\frac{1+c\lambda}{1+c\gamma_i\lambda} \right) \\ &= \ln \left(\frac{1+c\lambda}{1+c\gamma_{y+1}\lambda} \right) + \sum_{i=y+2}^{m-1} \ln \left(\frac{1+c\lambda}{1+c\gamma_i\lambda} \right) \\ &= g_1(0, y+1) \end{aligned}$$

and

$$\begin{aligned} g_2(\lambda, y) &= \frac{(1-\kappa)^\lambda}{q_y} + \sum_{i=y+1}^{m-1} \frac{1}{q_i} > \sum_{i=y+1}^{m-1} \frac{1}{q_i} \\ &= \frac{(1-\kappa)^0}{q_{y+1}} + \sum_{i=y+2}^{m-1} \frac{1}{q_i} = g_2(0, y+1). \quad \square \end{aligned}$$

Drift analysis [30,32] is an important tool in runtime analysis of randomised search heuristics. Here we introduce a variant of the additive drift theorem [32] with its proof.

In the following, “ $(X_{t+1} - X_t + \varepsilon) ; t < T_a$ ” is the short-hand notation for “ $(X_{t+1} - X_t + \varepsilon) \cdot \mathbf{1}_{\{t < T_a\}}$ ” (see page 49 in [35]). Whenever we write an equality or inequality involving conditional expectation w.r.t. a σ -algebra, (e.g. $E[X | \mathcal{F}] \leq Y$), we have the “almost surely” meaning in mind.

Lemma 22 (Additive drift theorem). *Let $(Z_t)_{t \in \mathbb{N}}$ be a discrete-time stochastic process in $[0, \infty)$ adapted to any filtration $(\mathcal{F}_t)_{t \in \mathbb{N}}$. For any $a \geq 0$, define $T_a := \min\{t \in \mathbb{N} \mid Z_t \leq a\}$. If for some $\varepsilon > 0$*

1. $E[Z_{t+1} - Z_t + \varepsilon ; t < T_a \mid \mathcal{F}_t] \leq 0$ for all $t \in \mathbb{N}$,
2. $E[Z_t] < \infty$ for all $t \in \mathbb{N}$, and
3. $E[T_a] < \infty$,

then $E[T_a \mid \mathcal{F}_0] \leq Z_0/\varepsilon$.

Proof. Define the stopped process $S_t := Z_{t \wedge T_a} + \varepsilon(t \wedge T_a)$ where $t \wedge T_a := \min(t, T_a)$. By the definition of this process, it holds for all $t \in \mathbb{N}$ almost surely that

$$|S_t| \leq Z_t + \varepsilon t, \quad (23)$$

and, hence by condition 2 and 3, that for all $t \in \mathbb{N}$,

$$E[|S_t|] \leq E[Z_t + \varepsilon T_a] < \infty. \quad (24)$$

Also, by the definition of the process, for all $t \in \mathbb{N}$ it holds in the case $t \geq T_a$ that,

$$E[S_{t+1} ; t \geq T_a \mid \mathcal{F}_t] = E[S_t ; t \geq T_a \mid \mathcal{F}_t].$$

Furthermore, for all $t \in \mathbb{N}$, it holds in the case $t < T_a$,

$$\begin{aligned} E[S_{t+1}; t < T_a | \mathcal{F}_t] &= E[(Z_{t+1} - Z_t + \varepsilon) + Z_t + \varepsilon t; t < T_a | \mathcal{F}_t] \\ &\leq E[Z_t + \varepsilon t; t < T_a | \mathcal{F}_t] = E[S_t; t < T_a | \mathcal{F}_t] \end{aligned}$$

where the inequality is due to condition 1. Combining both cases, we have for all $t \in \mathbb{N}$,

$$E[S_{t+1} | \mathcal{F}_t] \leq E[S_t | \mathcal{F}_t] = S_t. \quad (25)$$

By (24) and (25), S_t is a super-martingale, implying that for all $t \in \mathbb{N}$,

$$E[S_t | \mathcal{F}_0] \leq E[S_0 | \mathcal{F}_0] = Z_0. \quad (26)$$

By (23) and (24), the dominated convergence theorem (see e.g. [35]) applies, and we get by (26)

$$\begin{aligned} Z_0 &\geq \lim_{t \rightarrow \infty} E[S_t | \mathcal{F}_0] = E\left[\lim_{t \rightarrow \infty} S_t | \mathcal{F}_0\right] \\ &= E[Z_{T_a} + \varepsilon T_a | \mathcal{F}_0]. \end{aligned}$$

By noting that $Z_{T_a} \geq 0$, the proof is now complete. \square

Lemma 23 (Improved version of Lemma 5 in [16]). *If $X \sim \text{Bin}(\lambda, p)$ with $p \geq (i/\lambda)(1 + \delta)$ and $i \geq 1$ for some $\delta \in (0, 1]$, then*

$$E\left[\ln\left(\frac{1 + \delta X/2}{1 + \delta i/2}\right)\right] \geq \frac{\delta^2}{7}.$$

This improvement is due to the following generalisation of the lower bound in Lemma 17.

Lemma 24. *For any $z > 0$, and all $x \geq 0$ we have that*

$$\begin{aligned} \ln(1+x) &\geq x(b(z) + a(z)x) \\ \text{where } a(z) &:= \frac{1}{z(z+1)} - \frac{\ln(1+z)}{z^2}, \\ \text{and } b(z) &:= \frac{2\ln(1+z)}{z} - \frac{1}{1+z}. \end{aligned}$$

Proof. For $x = 0$, the result trivially holds. It then suffices to show that for all $x \in (0, \infty)$

$$h(x) := \frac{\ln(1+x)}{x} - b(z) - a(z)x \geq 0.$$

Note that $h(z) = 0$ and $h'(x) = a(x) - a(z)$. It follows from $\ln(1+x) > 2x/(x+2)$ for $x > 0$ (see (3) in [55]) that

$$\begin{aligned} a'(x) &= \frac{2\ln(1+x)}{x^3} - \frac{2}{x^2(1+x)} - \frac{1}{x(1+x)^2} \\ &\geq \frac{4}{x^2(2+x)} - \frac{2}{x^2(1+x)} - \frac{1}{x(1+x)^2} \end{aligned}$$

$$= \frac{1}{(2+x)(1+x)^2} > 0,$$

thus $a(x)$ is an increasing function.

We separate two cases: for $x \in (0, z]$, we have $a(x) \leq a(z)$ and $h'(x) \leq 0$, thus $h(x)$ is decreasing on $(0, z]$ and $h(x) \geq h(z) = 0$; for $x \in [z, \infty)$ we have $h'(x) = a(x) - a(z) \geq 0$, $h(x)$ is increasing on $[z, \infty)$ and $h(x) \geq h(z) = 0$. We have shown that $h(x) \geq 0$ for $x > 0$. \square

Note that the bound is tight at both $x = 0$ and $x = z$. The lemma does not cover the case $z = 0$, however at the limit, we get $\lim_{z \rightarrow 0^+} b(z) = 1$ and $\lim_{z \rightarrow 0^+} a(z) = -1/2$, and that corresponds to the bound given by Lemma 17.

Corollary 25. *Let $X \sim \text{Bin}(n, p)$ and $\mu := E[X]$, then it holds that for all $c > 0$*

$$E[\ln(1 + cX)] \geq \ln(1 + c\mu) - \frac{c}{2} \cdot \frac{c\mu}{1 + c\mu}.$$

Proof. For $p = 0$ (or $\mu = 0$), the bound is trivial. Otherwise, for $p > 0$, applying Lemma 24 with $z = c\mu$ gives $\ln(1 + cX) \geq b(c\mu)cX + a(c\mu)(cX)^2$, hence

$$\begin{aligned} E[\ln(1 + cX)] &\geq b(c\mu)c\mu + a(c\mu)c^2\mu(1 - p + \mu) \\ &= \ln(1 + c\mu) - c(1 - p) \left(\frac{\ln(1 + c\mu)}{c\mu} - \frac{1}{1 + c\mu} \right) \\ &> \ln(1 + c\mu) - c \left(\frac{1}{2} \cdot \frac{2 + c\mu}{1 + c\mu} - \frac{1}{1 + c\mu} \right) \\ &= \ln(1 + c\mu) - \frac{c}{2} \cdot \frac{c\mu}{1 + c\mu}. \end{aligned}$$

The last inequality is due to $1 - p < 1$ and $\ln(1 + x)/x < (1/2)(x + 2)/(x + 1)$ for $x > 0$ (see (3) in [55]). \square

We now give the formal proof of Lemma 23.

Proof of Lemma 23. Let $Y \sim \text{Bin}(\lambda, (1 + \delta)i/\lambda)$, then $Y \preceq X$. Therefore,

$$E \left[\ln \left(\frac{1 + \delta X/2}{1 + \delta i/2} \right) \right] \geq E \left[\ln \left(\frac{1 + \delta Y/2}{1 + \delta i/2} \right) \right]$$

and it is sufficient to show that $E \left[\ln \left(\frac{1 + \delta Y/2}{1 + \delta i/2} \right) \right] > \delta^2/7$ to complete the proof.

It follows from Corollary 25 (choosing $c = \delta/2$) that

$$\begin{aligned} &E \left[\ln \left(\frac{1 + \delta Y/2}{1 + \delta i/2} \right) \right] \\ &\geq \ln \left(\frac{1 + (1 + \delta)\delta i/2}{1 + \delta i/2} \right) - \frac{\delta}{4} \cdot \frac{(1 + \delta)\delta i/2}{1 + (1 + \delta)\delta i/2} \\ &= \ln \left(1 + \frac{i\delta^2}{2 + i\delta} \right) - \frac{\delta}{4} \cdot \frac{(1 + \delta)\delta i}{2 + (1 + \delta)\delta i} =: h(i). \end{aligned}$$

For all $\delta > 0$ and $i \geq 1$, it holds that

$$h'(i) = \frac{1}{2} \cdot \frac{(6 + \delta(3i - 2) + 3i\delta^2)\delta^2}{(2 + i\delta + i\delta^2)^2(2 + i\delta)} > 0,$$

or $h(i)$ monotonically increases in i .

Define $r(\delta) := 12 + 8\delta + 3\delta^2 + \delta^3 - 2\delta^4$ and $s(\delta) := (2 + \delta)^2(2 + \delta + \delta^2) > 0$, we get

$$\begin{aligned} h(i) &\geq h(1) = \ln\left(1 + \frac{\delta^2}{2 + \delta}\right) - \frac{\delta}{4} \cdot \frac{(1 + \delta)\delta}{2 + (1 + \delta)\delta} \\ &\geq \frac{\delta^2}{2 + \delta} \left(1 - \frac{\delta^2}{2(2 + \delta)}\right) - \frac{\delta}{4} \cdot \frac{(1 + \delta)\delta}{2 + (1 + \delta)\delta} = \frac{\delta^2 r(\delta)}{4s(\delta)}. \end{aligned}$$

The last inequality is due to Lemma 17. We notice that $18r(\delta) - 11s(\delta) = (1 - \delta)(128 + 140\delta + 84\delta^2 + 47\delta^3) \geq 0$ for all $\delta \in (0, 1]$, thus $r(\delta)/s(\delta) \geq 11/18$ and $h(i) \geq (\delta^2/4)(11/18) > \delta^2/7$. \square

Lemma 26 (Lemma 6 in [16]). *If $X \sim \text{Bin}(\lambda, p)$ with $p \geq (i/\lambda)(1 + \delta)$, then $E[e^{-\kappa X}] \leq e^{-\kappa i}$ for any $\kappa \in (0, \delta)$.*

Proof. The value of the moment generating function $M_X(t)$ of the binomially distributed variable X at $t = -\kappa$ is

$$E[e^{-\kappa X}] = M_X(-\kappa) = (1 - p(1 - e^{-\kappa}))^\lambda$$

It follows from By Lemma 18 and from $1 + \kappa < 1 + \delta$ that

$$p(1 - e^{-\kappa}) \geq \frac{i(1 + \delta)}{\lambda} \left(\frac{\kappa}{1 + \kappa}\right) \geq \frac{\kappa i}{\lambda}.$$

Altogether, we get $E[e^{-\kappa X}] \leq (1 - \kappa i/\lambda)^\lambda \leq e^{-\kappa i}$. \square

Lemma 27. *Let $\{X_i\}_{i \in [\lambda]}$ be i.i.d. random variables, define $Y(j) := \sum_{i=1}^{\lambda} \mathbb{1}_{\{X_i \geq j\}}$ for any $j \in \mathbb{R}$. It holds for any $a, b, c, j \in \mathbb{R}$ with $c \geq 0$ and $b \leq \lambda$ that*

$$(i) \Pr(Y(j + c) \geq a \mid Y(j) \geq b) \geq \Pr(Y(j + c) \geq a)$$

and for any non-decreasing function f

$$(ii) E[f(Y(j + c)) \mid Y(j) \geq b] \geq E[f(Y(j + c))]$$

provided that both expectations are well-defined.

Proof. Define $p := \Pr(X_i \geq j)$ and $q := \Pr(X_i \geq j + c)$. For $b \leq 0$ or $p = 0$, the result trivially holds. For $b \in (0, \lambda]$ and $p \in (0, 1]$, we have that $q' := \Pr(X_i \geq j + c \mid X_i \geq j) = q/p \geq q$. Event $Y(j) \geq b$ implies the existence of a set $A \subseteq [\lambda]$ such that $|A| \geq [b]$ and $X_i \geq j$ for all $i \in A$. Define $Y_1 := \sum_{i \in A} \mathbb{1}_{\{X_i \geq j + c\}}$ and $Y_2 := \sum_{i \in [\lambda] \setminus A} \mathbb{1}_{\{X_i \geq j + c\}}$, so $Y(j + c) = Y_1 + Y_2$. Clearly, conditioned on $Y(j) \geq b$, $Y_1 \sim \text{Bin}(|A|, q')$ and $Y_2 \sim \text{Bin}(\lambda - |A|, q)$. Therefore, the distribution of $Y(j + c)$ conditioned on $Y(j) \geq b$ stochastically dominates $\text{Bin}(|A|, q) + \text{Bin}(\lambda - |A|, q) = \text{Bin}(\lambda, q)$, which is the (unconditional or original) distribution of $Y(j + c)$, and part (i) follows.

For part (ii), let $F_1(x) := \Pr(f(Y(j+c)) < x \mid Y(j) \geq b)$ and $F_2(x) := \Pr(f(Y(j+c)) < x)$, i.e. F_1 and F_2 are the conditional and the unconditional distribution functions of $f(Y(j+c))$ respectively. Then from part (i) we conclude that $F_1(x) \leq F_2(x)$ for any $x \in \mathbb{R}$, and by the properties of expectation,

$$\begin{aligned}
& E[f(Y(j+c)) \mid Y(j) \geq b] \\
&= - \int_{-\infty}^0 F_1(x) dx + \int_0^{\infty} (1 - F_1(x)) dx \\
&\geq - \int_{-\infty}^0 F_2(x) dx + \int_0^{\infty} (1 - F_2(x)) dx \\
&= E[f(Y(j+c)) \mid Y(j) \geq b]. \quad \square
\end{aligned}$$

Appendix C

The following results are used to analyse the tightness of the level-based theorem, i.e. Theorem 16.

Lemma 28 (Additive drift theorem (lower bound)). *Let $(Z_t)_{t \in \mathbb{N}}$ be a discrete-time stochastic process in $[0, \infty)$ adapted to any filtration $(\mathcal{F}_t)_{t \in \mathbb{N}}$. For any $a \geq 0$, define $T_a := \min\{t \in \mathbb{N} \mid Z_t \leq a\}$. If for some $\varepsilon > 0$*

1. $E[Z_{t+1} - Z_t + \varepsilon; t < T_a \mid \mathcal{F}_t] \geq 0$ for all $t \in \mathbb{N}$, and
2. $E[Z_t] < \infty$ for all $t \in \mathbb{N}$.
3. $E[T_a] < \infty$,

then $E[T_a \mid \mathcal{F}_0] \geq (Z_0 - a)/\varepsilon$.

Proof. The proof is similar to that of Lemma 22, i.e. starting by defining the same stopped process S_t . However, because the directions of the inequalities are inverted so S_t is a sub-martingale, and in the end we overestimate X_{T_a} by a . \square

Lemma 29. *If $X \sim \text{Bin}(n, p)$ where $p > 0$, then for all $y \in \mathbb{R}$*

$$E[\max(X, y)] < \max(np, y) + (1/2)\sqrt{np}.$$

Proof. By Jensen's inequality w.r.t. the square root, we have

$$\begin{aligned}
E[|X - y|] &= E\left[\sqrt{(X - y)^2}\right] \leq \sqrt{E[(X - y)^2]} \\
&= \sqrt{np(1-p) + (np - y)^2} \\
&\leq \sqrt{np(1-p)} + |np - y|,
\end{aligned}$$

where the last inequality uses $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$ for $a, b \geq 0$. Therefore, it holds that

$$\begin{aligned}
E[\max(X, y)] &= E[(1/2)(X + y + |X - y|)] \\
&\leq (1/2)(np + y + |np - y| + \sqrt{np(1-p)}) \\
&< \max(np, y) + (1/2)\sqrt{np}. \quad \square
\end{aligned}$$

Lemma 30. *If $X \sim \text{Bin}(n, p)$ where $p > 0$ then it holds that $E[X | X > 0] \leq np + 1$.*

Proof. By definition,

$$\begin{aligned} E[X | X > 0] &= \sum_{i=1}^n i \Pr(X = i | X > 0) \\ &= \frac{1}{\Pr(X > 0)} \sum_{i=1}^n i \Pr(X = i \cap X > 0) \\ &= \frac{1}{\Pr(X > 0)} \sum_{i=0}^n i \Pr(X = i) \\ &= \frac{E[X]}{\Pr(X > 0)} = \frac{np}{1 - (1 - p)^n} \leq np + 1, \end{aligned}$$

where the last inequality follows from Lemma 18. □

References

- [1] G. Ausiello and M. Protasi, “Local search, reducibility and approximability of NP-optimization problems,” *Inform. Process. Lett.*, vol. 54, pp. 73–79, 1995.
- [2] A. Avidor, I. Berkovitch, and U. Zwick, “Improved approximation algorithms for MAX NAE-SAT and MAX SAT,” in *Proc. of 3rd Intl. Conf. Approx. Online Algorithm.* Springer Berlin Heidelberg, 2006, pp. 27–40.
- [3] G. Badkobeh, P. K. Lehre, and D. Sudholt, “Unbiased black-box complexity of parallel search,” in *Proc. of PPSN XIII.* Springer, 2014, pp. 892–901.
- [4] E. Balas and W. Niehaus, “Optimized crossover-based genetic algorithms for the maximum cardinality and maximum weight clique problems,” *J. Heuristics*, vol. 4, no. 2, pp. 107–122, 1998.
- [5] T. Chen, J. He, G. Sun, G. Chen, and X. Yao, “A new approach for analyzing average time complexity of population-based evolutionary algorithms on unimodal problems,” *IEEE Trans. Syst. Man. Cybern. B*, vol. 39, no. 5, pp. 1092–1106, 2009.
- [6] T. Chen, P. K. Lehre, K. Tang, and X. Yao, “When is an estimation of distribution algorithm better than an evolutionary algorithm?” in *Proc. of CEC ’09.* IEEE, 2009, pp. 1470–1477.
- [7] T. Chen, K. Tang, G. Chen, and X. Yao, “On the analysis of average time complexity of estimation of distribution algorithms,” in *Proc. of CEC ’07.* IEEE, 2007, pp. 453–460.
- [8] —, “Rigorous time complexity analysis of univariate marginal distribution algorithm with margins,” in *Proc. of CEC ’09.* IEEE, 2009, pp. 2157–2164.

- [9] —, “Analysis of computational time of simple estimation of distribution algorithms,” *IEEE Trans. Evol. Comput.*, vol. 14, no. 1, pp. 1–22, 2010.
- [10] D. Corus, D.-C. Dang, A. V. Eremeev, and P. K. Lehre, “Level-based analysis of genetic algorithms and other search processes,” in *Proc. of PPSN XIII*. Springer, 2014, pp. 912–921.
- [11] D. Corus and P. K. Lehre, “Theory-driven design of efficient genetic algorithms for a classical graph problem,” 2015, (To appear) *Proc. of MIC ’15*.
- [12] D.-C. Dang, T. Friedrich, T. Koetzing, M. S. Krejca, P. K. Lehre, P. S. Oliveto, D. Sudholt, and A. M. Sutton, “Emergence of diversity and its benefits for crossover in genetic algorithms,” in *Proc. of PPSN XIV*. Springer, 2016, pp. 890–900.
- [13] D.-C. Dang, T. Jansen, and P. K. Lehre, “Populations can be essential in tracking dynamic optima,” *Algorithmica*, 2016, (DOI: 10.1007/s00453-016-0187-y).
- [14] D.-C. Dang and P. K. Lehre, “Efficient optimisation of noisy fitness functions with population-based evolutionary algorithms,” in *Proc. of FOGA XIII*. ACM, 2015, pp. 62–68.
- [15] —, “Simplified runtime analysis of estimation of distribution algorithms,” in *Proc. of GECCO ’15*. ACM, 2015, pp. 513–518.
- [16] —, “Runtime analysis of non-elitist populations: From classical optimisation to partial information,” *Algorithmica*, vol. 75, no. 3, pp. 428–461, 2016.
- [17] —, “Self-adaptation of mutation rates in non-elitist populations,” in *Proc. of PPSN XIV*. Springer, 2016, pp. 803–813, (arXiv:1606.05551).
- [18] B. Doerr and C. Doerr, “Optimal parameter choices through self-adjustment: Applying the 1/5-th rule in discrete settings,” in *Proc. of GECCO ’15*. ACM, 2015, pp. 1335–1342.
- [19] —, “A tight runtime analysis of the $(1+(\lambda, \lambda))$ genetic algorithm on OneMax,” in *Proc. of GECCO ’15*. ACM, 2015, pp. 1423–1430.
- [20] B. Doerr, C. Doerr, and F. Ebel, “From black-box complexity to designing new genetic algorithms,” *Theor. Comput. Sci.*, vol. 567, pp. 87–104, 2015.
- [21] S. Droste, “A rigorous analysis of the compact genetic algorithm for linear functions,” *Nat. Comput.*, vol. 5, no. 3, pp. 257–283, 2006.
- [22] S. Droste, T. Jansen, and I. Wegener, “Upper and Lower Bounds for Randomized Search Heuristics in Black-Box Optimization,” *Theor. Comput. Syst.*, vol. 39, no. 4, pp. 525–544, 2006.
- [23] A. Eremeev and J. Kovalenko, “Optimal recombination in genetic algorithms for combinatorial optimization problems: Part I,” *Yugosl. J. Oper. Res.*, vol. 24, no. 1, pp. 1–20, 2014.

- [24] A. V. Eremeev, “Modeling and analysis of genetic algorithm with tournament selection,” in *Proc. of Artif. Evol.* Springer Berlin Heidelberg, 2000, pp. 84–95.
- [25] —, “Hitting times of local and global optima in genetic algorithms with very high selection pressure,” 2016, (To appear, arXiv:1606.05784) Yugosl. J. Oper. Res.
- [26] U. Feige, “On sums of independent random variables with unbounded variance, and estimating the average degree in a graph,” in *Proc. of 36th STOC*, 2004, pp. 594–603.
- [27] T. Friedrich, T. Kötzing, M. S. Krejca, and A. M. Sutton, “The benefit of recombination in noisy evolutionary search,” in *Proc. of 26th ISAAC*. Springer, 2015, pp. 140–150.
- [28] D. E. Goldberg, *Genetic Algorithms in search, optimization and machine learning*. Addison-Wesley, MA, USA, 1989.
- [29] C. González, J. A. Lozano, and P. Larrañaga, “Analyzing the PBIL algorithm by means of discrete dynamical systems,” *Complex Syst.*, vol. 12, pp. 465–479, 2000.
- [30] B. Hajek, “Hitting-time and occupation-time bounds implied by drift analysis with applications,” *Adv. Appl. Probab.*, vol. 14, no. 3, pp. 502–525, 1982.
- [31] G. R. Harik, F. G. Lobo, and D. E. Goldberg, “The compact genetic algorithm,” *IEEE Trans. Evol. Comput.*, vol. 3, no. 4, pp. 287–297, 1999.
- [32] J. He and X. Yao, “Drift analysis and average time complexity of evolutionary algorithms,” *Artif. Intell.*, vol. 127, no. 1, pp. 57–85, 2001.
- [33] —, “From an individual to a population: an analysis of the first hitting time of population-based evolutionary algorithms,” *IEEE Trans. Evol. Comput.*, vol. 6, no. 5, pp. 495–511, 2002.
- [34] T. Jansen and I. Wegener, “The analysis of evolutionary algorithms – a proof that crossover really can help,” *Algorithmica*, vol. 34, no. 1, pp. 47–66, 2002.
- [35] O. Kallenberg, *Foundations of Modern Probability*. Springer-Verlag, New York, 2002.
- [36] T. Kötzing, D. Sudholt, and M. Theile, “How crossover helps in pseudo-boolean optimization,” in *Proc. of GECCO ’11*. ACM, 2011, pp. 989–996.
- [37] P. Larrañaga and J. A. Lozano, Eds., *Estimation of Distribution Algorithms: A New Tool for Evolutionary Computation*, ser. Genetic Algorithms and Evolutionary Computation. Springer, 2002, vol. 2.
- [38] J. Lässig and D. Sudholt, “General scheme for analyzing running times of parallel evolutionary algorithms,” in *Proc. of PPSN XI*. Springer, 2010, pp. 234–243.

- [39] P. K. Lehre, “Fitness-levels for non-elitist populations,” in *Proc. of GECCO '11*. ACM, 2011, pp. 2075–2082.
- [40] P. K. Lehre and C. Witt, “Black-box search by unbiased variation,” *Algorithmica*, pp. 1–20, 2012.
- [41] P. K. Lehre and X. Yao, “Crossover can be constructive when computing unique input-output sequences,” *Soft Comput.*, vol. 15, no. 9, pp. 1675–1687, 2011.
- [42] —, “On the impact of mutation-selection balance on the runtime of evolutionary algorithms,” *IEEE Trans. Evol. Comput.*, vol. 16, no. 2, pp. 225–241, April 2012.
- [43] A. Moraglio and D. Sudholt, “Principled design and runtime analysis of abstract convex evolutionary search,” *Evol. Comput.*, 2015, posted Online.
- [44] H. Mühlenbein and G. Paaß, “From recombination of genes to the estimation of distributions i. binary parameters,” in *Proc. of PPSN IV*. Springer Berlin Heidelberg, 1996, pp. 178–187.
- [45] H. Mühlenbein, “The equation for response to selection and its use for prediction,” *Evol. Comput.*, vol. 5, no. 3, pp. 303–346, 1997.
- [46] F. Neumann, P. S. Oliveto, and C. Witt, “Theoretical analysis of fitness-proportional selection: landscapes and efficiency,” in *Proc. of GECCO '09*. ACM, 2009, pp. 835–842.
- [47] P. S. Oliveto and C. Witt, “Improved runtime analysis of the simple genetic algorithm,” in *Proc. of GECCO '13*. ACM, 2013, pp. 1621–1628.
- [48] —, “On the runtime analysis of the simple genetic algorithm,” *Theor. Comput. Sci.*, vol. 545, pp. 2–19, 2014.
- [49] —, “Improved time complexity analysis of the simple genetic algorithm,” *Theor. Comput. Sci.*, vol. 605, pp. 21–41, 2015.
- [50] M. Pelikan, K. Sastry, and D. E. Goldberg, “Scalability of the bayesian optimization algorithm,” *Intl. J. Approx. Reasoning*, vol. 31, no. 3, pp. 221–258, 2002.
- [51] A. Prügel-Bennett, J. E. Rowe, and J. Shapiro, “Run-time analysis of population-based evolutionary algorithm in noisy environments,” in *Proc. of FOGA XIII*. ACM, 2015, pp. 69–75.
- [52] J. Scharnow, K. Tinnefeld, and I. Wegener, “The analysis of evolutionary algorithms on sorting and shortest paths problems,” *J. Math. Model. Algorithm.*, vol. 3, no. 4, pp. 349–366, 2004.
- [53] J. L. Shapiro, “Drift and scaling in estimation of distribution algorithms,” *Evol. Comput.*, vol. 13, no. 1, pp. 99–123, 2005.
- [54] D. Sudholt, “Crossover speeds up building-block assembly,” in *Proc. of GECCO '12*. ACM, 2012, pp. 689–702.

- [55] F. Topsøe, “Some bounds for the logarithmic function,” in *Inequality Theory and Applications*, Y. J. Cho, J. K. Kim, and S. S. Dragomir, Eds. Nova Science Publishers, Incorporated, 2007, vol. 4, pp. 137–151.
- [56] M. D. Vose, *The Simple Genetic Algorithm: Foundations and Theory*. MIT Press, Cambridge, MA, 1999.
- [57] M. D. Vose and A. H. Wright, “Stability of vertex fixed points and applications,” in *Proc. of FOGA III*. Morgan Kaufmann, 1995, pp. 103–113.
- [58] I. Wegener, “Methods for the analysis of evolutionary algorithms on pseudo-boolean functions,” in *Evolutionary Optimization*. Springer US, 2002, vol. 48, pp. 349–369.
- [59] C. Witt, “Runtime analysis of the $(\mu+1)$ EA on simple pseudo-boolean functions,” *Evol. Comput.*, vol. 14, no. 1, pp. 65–86, 2006.
- [60] Y. Yu, C. Qian, and Z.-H. Zhou, “Switch analysis for running time analysis of evolutionary algorithms,” *IEEE Trans. Evol. Comput.*, vol. 19, no. 6, pp. 777–792, dec 2015.
- [61] Q. Zhang and H. Mühlenbein, “On the convergence of a class of estimation of distribution algorithms,” *IEEE Trans. Evol. Comput.*, vol. 8, no. 2, pp. 127–136, 2004.