Non-elitist evolutionary algorithms excel in fitness landscapes with sparse deceptive regions and dense valleys

Dang, Duc-Cuong; Eremeev, Anton; Lehre, Per Kristian

DOI:
10.1145/3449639.3459398

Document Version
Peer reviewed version

Citation for published version (Harvard):
Dang, D-C, Eremeev, A & Lehre, PK 2021, Non-elitist evolutionary algorithms excel in fitness landscapes with sparse deceptive regions and dense valleys. in F Chicano (ed.), *GECCO '21: Proceedings of the Genetic and Evolutionary Computation Conference.* Genetic and Evolutionary Computation Conference (GECCO), Association for Computing Machinery (ACM), New York, pp. 1133–1141, 2021 Genetic and Evolutionary Computation Conference, GECCO 2021, Virtual, Online, France, 10/07/21.
https://doi.org/10.1145/3449639.3459398

Non-elitist Evolutionary Algorithms Excel in Fitness Landscapes with Sparse Deceptive Regions and Dense Valleys*

Duc-Cuong Dang†, Anton Eremeev‡, Per Kristian Lehre§

May 5, 2021

Abstract

It is largely unknown how the runtime of evolutionary algorithms depends on fitness landscape characteristics for broad classes of problems. Runtime guarantees for complex and multi-modal problems where EAs are typically applied are rarely available.

We present a parameterised problem class $\textsc{SparseLocalOpt}_{\alpha,\varepsilon}$, where the class with parameters $\alpha, \epsilon \in [0,1]$ contains all fitness landscapes with deceptive regions of sparsity ε and fitness valleys of density α. We study how the runtime of EAs depends on these fitness landscape parameters.

We find that for any constant density and sparsity $\alpha, \varepsilon \in (0,1)$, $\textsc{SparseLocalOpt}_{\alpha,\varepsilon}$ has exponential elitist $(\mu+\lambda)$ black-box complexity, implying that a wide range of elitist EAs fail even for mildly deceptive and multi-modal landscapes. In contrast, we derive a set of sufficient conditions for non-elitist EAs to optimise any problem in $\textsc{SparseLocalOpt}_{\alpha,\varepsilon}$ in expected polynomial time for broad values of α and ε. These conditions can be satisfied for tournament selection and linear ranking selection, but not for (μ,λ)-selection.

1 Introduction

Non-elitist evolutionary algorithms (EAs) have been proved to be competitive with the elitist ones on a variety of benchmark functions Corus et al. (2018); Dang and Lehre (2016a), and to perform well in complex settings such as under incomplete information of the objective function Dang and Lehre (2016a), under noise Dang and Lehre (2015) or in dynamic optimisation Dang et al. (2017). Yet it remains an open question whether a pure non-elitist EA on some significant

*To appear in Proceedings of 2021 Genetic and Evolutionary Computation Conference.

†Southampton Business School, University of Southampton, UK <d.c.dang@soton.ac.uk>

‡Sobolev Institute of Mathematics Novosibirsk Russia, and INION RAN, Moscow, Russia <eremeev@ofim.oscsbras.ru>

§School of Computer Science, University of Birmingham, UK <p.k.lehre@cs.bham.ac.uk>

problem classes can efficiently escape a local optima region once the latter has been discovered. This is a fundamental question as non-elitism means the ability to at some point during the search forget some good solutions, albeit local optima, in favour of exploration. Showing rigorously that a non-elitist EA can escape a local optimum while elitist EAs get stuck has been a fundamental open problem in evolutionary computation. In the rest of the section, we will discuss the previous studies that are related to this open problem and describe the contribution of this paper.

1.1 Previous Work

It was shown in Rudolph (1996) that the use of bit-wise mutation is crucial for the (1+1) EA, in contrast to the fixed mutation of the Randomised Local Search (RLS), to find the global/local optimum of the LongPath function in expected polynomial time. Later, it was shown in Droste et al. (1998) that even the (1+1) EA requires a super-polynomial time to optimise the modified function $\textsc{LongPath}_{\sqrt{n}}$.

Crossover has been a long-time suggested operator that allows faster escaping from local optima Horn et al. (1994). Using Jump as the benchmark function where the performance of mutation-only EAs is restricted to $\Omega(n^k)$ with k being the Hamming distance from the global optimum from the local optima, in Dang et al. (2018) it was shown that the interplay between mutation and crossover operators can increase the diversity of the population in (μ+1) EA, hence enabling the algorithm to leave the local optima in a shorter time. An $\Omega(\sqrt{n})$ speed-up in the overall expected runtime is proved for the standard mutation rate $1/n$, and it is increased to $\Omega(n)$ with mutation rate $2/n$. In the same line of research, it was shown in Dang et al. (2016) that artificially enforcing population diversity with common mechanisms found in Evolutionary Computing (EC) literature enables the (μ+1) EA to escape the local optima faster and hence optimise Jump efficiently. Similar arguments relying on the impact of various operators found in EC, e. g. ageing, tabu, hypermutation, to population diversity allowed the local performance proofs of these operators Oliveto et al. (2019); Sudholt (2011); Zarges (2011).

Like the theoretical studies of elitist EAs, the research on the local performance of non-elitist EAs started with simple populations that consist of a single individual. Popular algorithms in this category are Metropolis, and Simulated Annealing (SA). It was shown Hajek (1988) that the convergence of SA to the global optimum depends on the temperature schedule and on the depth of the deepest local optimum which is not the global one. It was also proved Wegener (2005) that SA can beat Metropolis on a class of minimum spanning tree instances.

The $(1, \lambda)$ EA was compared with the (1+λ) EA on the Cliff function in Jägerskupper and Storch (2007), where the $(1, \lambda)$ EA is proved to outperform the $(1 + \lambda)$ EA when lambda is logarithmic in n. However, the $(1, \lambda)$ EA with smaller λ is shown to be inefficient on any pseudo-Boolean function with a unique optimum.

More recently, the scheme of Strong Selection Weak Mutation (SSWM), which is well-studied in Population Genetics, was introduced as a search algorithm for EC in Paixão et al. (2017). The authors proved a speed-up of $e^{\Omega(d)}$ of SSWM over the base expected runtime $\Theta(n^d)$ of the (1+1) EA on the CLIFF function, here d is both the height of the local optima and the distance from them to the global optimum.

The *elitist (μ+λ)-black-box model* introduced in Doerr and Lengler (2016) was used to study inherent performance limits of EAs using (certain) elitist selection mechanisms. The model covers any algorithm which bases its decisions solely on the μ best found solutions found so far. It was shown that some problem-tailored, non-elitist EAs can optimise efficiently some problem classes which have exponential elitist black-box complexity. However, these tailored EAs are unlikely to perform well on other problem classes, so Doerr and Lengler (2016) does not answer whether standard non-elitist EAs can outperform elitist EAs.

Based on the construction of LONGPATH, the VALLEYPATH function was introduced in Oliveto et al. (2018). It was shown that the non-elitist algorithms with population of size 1, such as the SSWM and Metropolis, are able to cross a valley of deceptive fitness, and their ability to escape the current local optimum depends on the depth of the valley. This is in contrast to the (1+1) EA in which case the ability to escape crucially depends on the width of the valley.

In the experimental analysis of meta-heuristics, also a lot of attention has been given to the relationship between the structure of the fitness landscape and the performance of EAs (see e.g. Herrmann (2016); Reeves and Eremeev (2004)), especially the ones with a population of a single individual, elitist and non-elitist. In particular, in Thomson et al. (2017) it was demonstrated that the presence of multiple sub-optimal funnels in fitness landscapes contributes to lower success for Randomised Local Search and SA.

Related to our work, a simple modification of the LEADINGONES function was studied in Dang and Lehre (2016b), where a single peak/local optimum at height m is placed at the all zeroes bitstring. Under the assumption of the initial population on the peak, a normal (μ+λ) EA obviously needs $e^{\Omega(m)}$ expected evaluations to optimise the function. On the other hand, the (μ, λ) EA with two possibilities of mutation parameter embedded and co-evolving with each individual is shown to optimise the function in $\mathrm{O}\big(n\lambda \log \lambda + n^2\big)$ expected evaluations. Furthermore, the same framework of self-adaption for non-elitist populations has been also adopted in Case and Lehre (2020) to include more possibilities for the mutation parameter, allowing the handling of optimisation problems with unknown solution length efficiently.

A common feature of such positive results for the non-elitist EAs is that the mutation rate has to be set below but pretty close to a known limit, above which they are inefficient in optimising functions with unique optimum Lehre (2010). This limit, called the *error threshold*, is studied in both evolutionary computing Doerr (2020); Lehre (2010, 2011), population genetics Wilke (2005), and virology Biebricher and Eigen (2005). The experimental results in Dang et al. (2021) have also indicated that setting the mutation rate close to the error threshold

allows the EAs with tournament selection to succeed on some benchmarks of the Set Cover problem.

In contrast to the upper bounds on runtime of the non-elitist EAs Corus et al. (2018); Lehre (2011); Dang and Lehre (2016b), the tight analysis of the (μ, λ) EA runtime on JUMP, obtained in Doerr (2020), indicates that there is no benefit of the non-elitist comma selection with any setting of mutation rate on this function with multiple local optima.

The first exponential separation between the elitist (μ+λ) EA and a non-elitist, population-based EA was shown for the FUNNEL problem Dang et al. (2021). This result holds for 3-tournament selection, and contrasts the conclusions from Doerr (2020), indicating that results for (μ, λ)-selection cannot be extrapolated to other non-elitist selection mechanisms. In fact, it was shown in Dang et al. (2021) that the non-elitist (μ,λ)-EA has exponential expected runtime on FUNNEL, assuming that the μ best individuals of the initial population are in the basin of attraction of a local optimum and close to it. The (μ,λ) EA considered in the last negative result is indeed a non-elitist algorithm, except that the comma selection is used instead of tournament or linear ranking selection. The intuition to the difference between (μ, λ)-selection and non-elitist selection mechanisms like tournament selection comes from the following difference in selection probabilities. Let P be a population of size λ and sorted in a descending order according to fitness, and define $\beta(\gamma, P)$ to be the probability of selecting an individual at least as good as the $\lceil\gamma\lambda\rceil$-ranked individual of P, see e.g. Lehre (2011). In the case of (μ, λ)-selection, before reaching 1.0, $\beta(\gamma, P)$ is essentially linear Lehre (2011): $\beta(\gamma, P) = \gamma\lambda/\mu$ if $\gamma \leq \mu/\lambda$. The advantage that tournament has over the comma selection is the non-linearity of its $\beta(\gamma, P)$, namely $\beta(\gamma, P) = 1 - (1 - \gamma)^k$, Lehre (2011) where k is the tournament size.

Indeed, let us make a simplistic assumption that p_0 is a constant representing the probability of making copies of some high quality solution x, which could be a promising new incumbent or just a local optimum causing a stagnation, and assume that a γ-fraction of the population are copies of x. If x is a promising new incumbent, we want this fraction to grow, which requires that $\beta(\gamma, P)p_0 > \gamma$ when γ is small. If x is just a local optimum, we also want the fraction to not grow too large so that the remaining part of the population has some chance to reproduce, this requires $\beta(\gamma, P)p_0 < \gamma$ when γ is too large. It is easy to see that within the co-domain $[0, 1)$ the $\beta(\gamma, P)$ mentioned above for tournament selection can display both properties: it can be set to stay above the function $f(\gamma) = \gamma/p_0$ up to some point, then afterwards to move below f. The linear function of comma selection on the other hand can only display either one of the two properties.

1.2 Contributions of this Paper

We study how the runtime of evolutionary algorithms depend on properties of the fitness landscape Wright (1932); Stadler (2002). We classify fitness landscapes based on the *sparsity* of deceptive regions ("local optima") and the *density* of surrounding "fitness valleys" (see Defs. 1 and 2).

For any parameters $\alpha, \varepsilon \in [0,1]$ and fitness function f, we say that f belongs to the problem class $\textsc{SparseLocalOpt}_{\alpha,\varepsilon}$ if all its deceptive regions are ε-sparse, and all its fitness valleys are α-dense (Definition 4). Thus, the parameters induce a problem hierarchy: whenever $0 \leq \alpha' \leq \alpha \leq 1$ and $0 \leq \varepsilon \leq \varepsilon' \leq 1$,

$$\textsc{SparseLocalOpt}_{\alpha,\varepsilon} \subseteq \textsc{SparseLocalOpt}_{\alpha',\varepsilon'}.$$

Intuitively, increasing the parameter ε relaxes the sparsity requirement for the deceptive regions, and decreasing the parameter α relaxes the density requirement on the fitness valleys. In the limit when $\alpha = 0$ and $\varepsilon = 1$, the problem class contains almost any function (barring a technical condition (SP2) from Def. 2). In the other limit, when $\alpha = 1$ and $\varepsilon = 0$, the class contains only completely non-deceptive problems, such as OneMax and LeadingOnes.

We prove the following runtime results with respect to this classification:

- $\textsc{SparseLocalOpt}_{\alpha,\varepsilon}$ has an exponential *elitist* black box complexity (in the sense of Doerr and Lengler (2016)) for any constant levels of denseness α and sparsity ε, which we demonstrate on the BBFunnel problem sub-class. This negative result (Theorem 8) implies that a large set of elitist EAs, including those with one-point, bit-wise or heavy-tailed mutation, crossover etc. are inefficient on problems with even mild degrees of deception.
- Standard diversity mechanisms do not help the elitist (μ+1) GA optimise BBFunnel efficiently (Theorem 9).
- Non-elitist EAs with bit-wise mutation have expected polynomial runtime on $\textsc{SparseLocalOpt}_{\alpha,\varepsilon}$, given the appropriate values of α, ε, and selection and mutation parameters (Theorem 11). In particular, this applies to EAs with 3-tournament selection and linear ranking selection.
- The non-elitist (μ,λ) EA has an exponential expected runtime on the BBFunnel problem, assuming the standard initialisation of the EA population (Theorem 14). This standard assumption improves the negative result on (μ,λ) EA from Dang et al. (2021).

2 Preliminaries

The natural and base-2 logarithms are denoted $\ln(\cdot)$, and $\log(\cdot)$ respectively. $\mathbb{N}$ is the set of natural numbers. For any $n \in \mathbb{N}$, define $[n] := \{1, \ldots, n\}$. The Iverson bracket is denoted by $[\cdot]$. The *Hamming distance* is denoted by $\mathrm{H}(\cdot,\cdot)$. The *Hamming sphere* with radius $r \in [n]$ around a bitstring $x \in \{0,1\}^n$ is defined by $S_r(x) := \{y \in \{0,1\}^n \mid \mathrm{H}(x,y) = r\}$. Clearly, $|S_r(x)| = \binom{n}{r}$.

Given a partition of a search space $\mathcal{X}$ into m ordered "levels" $(A_1, \ldots, A_m)$, we define $A_{\geq j} := \cup_{i=j}^{m} A_i$.

Given two bitstrings x and y, we let $x \cdot y$ and xy denote the concatenated bitstrings.

A *simple path* is any sequence of bitstrings $x_1, \ldots, x_\ell \in \{0,1\}^m$, where for all $i, j \in [\ell]$, if $x_i = x_j$ then $i = j$ (i. e. uniqueness of each string), and for every $i \in [\ell - 1]$, $\mathrm{H}(x_i, x_{i+1}) = 1$ (i. e. consecutively separated by a Hamming distance of 1) (Droste et al., 2006). Clearly, simple paths contain no cycles.

Some definitions use the standard pseudo-Boolean functions $\textsc{OneMax}(x) := \textsc{Om}(x) := \sum_{i=1}^{n} x_i$, and y, $\textsc{Lo}_y(x) := \sum_{i=1}^{n} \prod_{j}^{i} [x_j = y_j]$ for a bitstring y.

For an event E and random variable X, we use the notation $E[X; E] := E[X 1_E]$, using the random variable $1_E(\omega) := [\omega \in E]$ (see e.g. Section 6.3 in Williams (1991)).

A *population* is a vector $P \in \mathcal{X}^\lambda$, the i-th individual of P is denoted $P(i)$. Given $x \in \mathcal{X}$, define $\mathrm{H}(x, P) := \min_{j \in [|P|]} \{\mathrm{H}(P(j), x)\}$, and for $A \subseteq \mathcal{X}$, we let $|P \cap A| := |\{i \mid P(i) \in A\}|$, i. e. the number of individuals of P belonging to A.

All non-elitist EAs with *unary variation operators* can be cast in the framework of Algorithm 1 Dang and Lehre (2016a). A new population P_{t+1} is generated by independently sampling λ individuals from an existing population P_t according to a selection mechanism p_{sel}, then by perturbing each of the selected individuals with a unary variation operator p_{mut}. The algorithm in turn is a special case of a more general framework of Algorithm 2, for which the *level-based theorem* Corus et al. (2018) was developed. To prove the positive result in this paper, we derive a variant of that theorem, i. e. see Theorem 10.

We will characterise selection mechanisms in the following way. If the individuals in a population P are ordered by decreasing fitness such that that $f(P(1)) \geq f(P(2)) \geq \cdots \geq f(P(\lambda))$, then $\beta(\psi, \gamma, P)$ where $0 \leq \psi \leq \gamma \leq 1$ is the probability that the selection mechanism chooses an individual ranked between $\lceil \psi\lambda \rceil$ and $\lceil \gamma\lambda \rceil$ in the population. We omit the symbol P from the notation when the population is clear from the context. Note also that the 2-argument variant of the definition of β, such as the one used in Lehre (2011) or in our introduction, is the special case of the above notation with $\psi = 0$.

Algorithm 1 Non-elitist EA with unary variation operator Dang and Lehre (2016a)

Require: Initial population $P_0 \in \mathcal{X}^\lambda$ where $\mathcal{X} = \{0,1\}^n$, and a mutation rate parameter $\chi \in [0, n]$.

1: **for** $t \in \mathbb{N}$ until a termination cond. is met **do**
2: **for** $i = 1$ to λ **do**
3: Sample $I_t(i) \sim p_{\mathrm{sel}}(P_t)$, and set $x := P_t(I_t(i))$.
4: Sample $x' \sim p_{\mathrm{mut}}(x, \chi)$, and set $P_{t+1}(i) := x'$.
5: **end for**
6: **end for**

In *k-tournament selection*, p_{sel} returns $\mathrm{argmax}_{i \in S} f(P_t(i))$ where S is a set of k random numbers is drawn independently and uniformly from $[\lambda]$. The corresponding β is non-linear, with $\beta(\gamma_1, \gamma_2) = (1 - \gamma_1)^k - (1 - \gamma_2)^k$. In (μ, λ)-*selection* (comma-selection), the set of indices $S = [\lambda]$ is first sorted according to fitness, then p_{sel} returns $S[i]$ where $i \sim \mathrm{Unif}([\mu])$. In this case, before reaching 1.0,

Algorithm 2 Population-based algorithm Corus et al. (2018).

Require: A finite state space $\mathcal{X}$, and population size $\lambda \in \mathbb{N}$,
a mapping D from $\mathcal{X}^\lambda$ to the space of prob. dist. over $\mathcal{X}$,
and an initial population $P_0 \in \mathcal{X}^\lambda$.

for $t = 0, 1, 2, \ldots$ until termination condition met **do**
  Sample $P_{t+1}(i) \sim D(P_t)$ independently for all $i \in [\lambda]$.
end for

Algorithm 3 $(\mu + \lambda)$ elitist black-box algorithm Doerr and Lengler (2016), f is unknown.

$P_0 \leftarrow \emptyset$
for $i \in [\mu]$ **do**
  Depending only on the multiset P_0 and the ranking $\rho(P_0, f)$ of P_0 induced by f, choose a probability distribution $p^{(i)}$ over $\{0,1\}^n$ and sample $x^{(i)}$ according to $p^{(i)}$
  $P_0 \leftarrow P_0 \cup \{x^{(i)}\}$
end for
for $t = 0, 1, 2, \ldots$ **do**
  Depending only on the multiset P_t and the ranking $\rho(P_t, f)$ of P_t induced by f choose a probability distribution $p^{(t)}$ on $(\{0,1\}^n)_{i=1}^\lambda$ and sample $(y^{(1)}, \ldots, y^{(\lambda)})$ according to $p^{(t)}$
  $P_{t+1} \leftarrow P_t \cup \{y^{(1)}, \ldots, y^{(\lambda)}\}$
  for $i \in [\lambda]$ **do**
    Select $x \in \operatorname{argmin} f(P_{t+1})$ and update $P_{t+1} \leftarrow P_{t+1} \setminus \{x\}$
  end for
end for

β is a linear function Lehre (2011) satisfying $\beta(\psi, \psi+\gamma) = \gamma\lambda/\mu$ if $(\psi+\gamma) \leq \mu/\lambda$. *Linear ranking selection* (Corollary 13) is defined in Goldberg and Deb (1991).

We consider the standard *bitwise mutation operator* as p_{mut} and it is configured by a parameter $\chi \in (0, n/2]$ so that for any pair of bitstrings $x, x' \in \{0,1\}^n$, the probability of obtaining x' from x is $\Pr(x' = p_{\text{mut}}(x, \chi)) = (\chi/n)^{\mathrm{H}(x,x')}(1-\chi/n)^{n-\mathrm{H}(x,x')}$.

The elitist black-box model Doerr and Lengler (2016) covers all algorithms with the outline of Algorithm 3. The initial μ search points may be sampled adaptively, i.e. the i-th sample may depend on the ranking of the first $i-1$ samples w.r. t. f. In each of the main iterations, this algorithm samples λ new search points from distributions that depend only on the current population P_t and the ranking of it. In each of these iterations, the λ offspring do not need to be independent of each other. However, it is required that all of the λ offspring are created before the fitness of any of them is evaluated.

Given a class $\mathcal{F}$ of pseudo-Boolean functions, the complexity of an algorithm A for this class is the maximum expected number of fitness evaluations made by A, before it evaluates an optimal solution for the first time, where the maximum is taken over all fitness functions $f \in \mathcal{F}$. The $(\mu+\lambda)$ elitist black-box complexity

of a class $\mathcal{F}$ is the minimum complexity, taken over all $(\mu+\lambda)$ elitist black-box algorithms A with the outline of the Algorithm 3.

In the elitist $(\mu+\lambda)$ EA (which is a special case of Algorithm 3), a new offspring population P is created by selecting uniformly those from P_t and perturbing them with p_{mut}. The surviving population P_{t+1} of the next generation then composes of the μ best individuals among both parent and offspring populations $P_t \cup P$.

3 Problems with Sparse Optima and Dense Fitness Valleys

We introduce a class of fitness landscapes $\textsc{SparseLocalOpt}_{\alpha,\varepsilon}$ which we claim separates elitist from non-elitist evolutionary algorithms. The class contains all functions which satisfy the following requirements. We first require that from any point in the search space, there must exist a not too long directed path to the global optimum, where consecutive steps on the path are near each other in the search space. Along any path, we distinguish between "deceptive regions" and "fitness valleys". Any region of the path with higher fitness than a later part of the path is called deceptive. Conversely, any region with lower fitness than an earlier part of the path is called a fitness valley. We only impose the constraint that deceptive regions must be sparse, while fitness valleys must be dense. Informally, a set is called dense if every member of the set has many neighbours in that set, and a set is called sparse if there are few set members in any direction. (In the following definitions, recall that $S_r(x)$ refers to the Hamming-sphere of radius r around bitstring x, as defined in Section 2.)

Definition 1. *For $\alpha \in [0,1]$, a subset $C \subseteq \{0,1\}^n$ is called α-dense if $\forall x \in C,\ |S_1(x) \cap C| \geq \alpha n$.*

Definition 2. *For $\varepsilon \in [0,1]$, a subset $B \subseteq \{0,1\}^n$ is ε-sparse if*

(SP1) $\forall x \in B, \forall r \in [n],\ |S_r(x) \cap B| \leq \varepsilon \cdot \binom{n}{r}$, *and*
(SP2) $\forall x \in \{0,1\}^n \setminus B, \forall r \in [n],\ |S_r(x) \cap B| = O\big(\frac{1}{n}\binom{n}{r}\big)$.

To make the notion of deceptive regions and fitness valleys more precise, we formally define deceptive pairs.

Definition 3. *Given a function $f : \{0,1\}^n \to \mathbb{R}$ and a partition $(A_1, \ldots, A_m)$ of $\{0,1\}^n$, a pair (A_i, A_j) is called f-deceptive if $1 \leq i < j \leq m$ and there are elements $x \in A_i$, $y \in A_j$ such that $f(x) \geq f(y)$.*

We can now state the definition of the problem class.

Definition 4. *An objective function $f : \{0,1\}^n \to \mathbb{R}$ belongs to the problem class $\textsc{SparseLocalOpt}_{\alpha,\varepsilon}$ if there exists a partition of $\{0,1\}^n$ into $m \in \text{poly}(n)$, subsets $(A_1, \ldots, A_m)$ such that*

- $A_m = \{x \in \{0,1\}^n \mid \forall y \in \{0,1\}^n, f(x) \geq f(y)\}$

- $\forall j \in [m-1], \forall x \in A_j, \exists y \in A_{\geq j+1}$ *s.t.* $\mathrm{H}(x,y) = O(1)$,

and if $(A_{i_1}, A_{j_i}), \ldots, (A_{i_u}, A_{j_u})$ *are* f*-deceptive pairs then*

- $\cup_{v=1}^{u} A_{i_v}$ *is* ε*-sparse, and*
- $A_{\geq j_v}$ *is* α*-dense for all* $v \in [u]$.

4 Elitist Black-box Algorithms

4.1 Elitist black-box complexity of BBFunnel

To prove that $\textsc{SparseLocalOpt}_{\alpha,\varepsilon}$ has exponential elitist black-box complexity, we consider a sub-class BBFunnel (see Definition 5 and Theorem 7). The sub-class captures the same features of the problem Funnel introduced in Dang et al. (2021), however it is a broader class. It is partly defined in terms of the functions $\textsc{Lo}_z$ and $\textsc{Om}$ (see Section 2).

Definition 5. *For any integers* $1 \leq u < v < w \leq n$*, and a simple path* $p_1, p_2, \ldots, p_\ell \in \{0,1\}^{v-u}$ *of length* $\ell \in \mathrm{poly}(n)$ *starting from* $p_1 := 0^{v-u}$*, let* $\textsc{BBFunnel}_\ell(x) :=$

$$\begin{cases} \textsc{Lo}_y(x) + \ell & \text{if } w < \textsc{Lo}_y(x) \leq n & (D) \\ \textsc{Lo}_y(x) & \text{if } v < \textsc{Lo}_y(x) \leq w & (C) \\ i + w & \text{if } x = 1^u p_i 0^{n-v} \text{ where } i \in [\ell] & (B) \\ -n - \textsc{Om}(x) & \text{if } \textsc{Lo}_y(x) \geq u \text{ and } x \notin B \cup C \cup D & (B') \\ \textsc{Lo}_y(x) & \text{if } \textsc{Om}(x) \leq u & (A) \\ -\textsc{Om}(x) & \text{if } x \notin A \cup B' \cup B \cup C \cup D & (A') \end{cases}$$

where a bitstring $z \in \{0,1\}^{w-v}$*, and* $y := 1^u \cdot p_\ell \cdot z \cdot 1^{n-w}$.

To prove a worst-case runtime for any randomised, elitist black-box algorithm, we will apply Yao's principle.

Theorem 6 (Yao's Principle (Yao, 1977))**.** *Let* Π *be a problem with a finite set* $\mathcal{I}$ *of input instances of fixed size permitting a finite set* $\mathcal{A}$ *of deterministic algorithms. Let* I_p *be a randomly chosen instance with a probability distribution* p *over* $\mathcal{I}$ *and let* A_q *be a randomly chosen algorithm with a probability distribution* q *over* $\mathcal{A}$*. Then* $\min_{A \in \mathcal{A}} E\left[T(I_p, A)\right] \leq \max_{I \in \mathcal{I}} E\left[T(I, A_q)\right]$*, where* $T(I, A)$ *denotes the running time of* $A \in \mathcal{A}$ *on* $I \in \mathcal{I}$.

We fix $\mathcal{A}$ to be the set of deterministic algorithms, which may become a realisation of the randomised elitist $(\mu + \lambda)$ black-box algorithms (see Algorithm 3). As pointed out by Doerr and Lengler Doerr and Lengler (2016), to account for the possibility that a deterministic elitist black-box algorithm can enter an infinite loop, it is necessary to extend the class $\mathcal{A}$ with algorithms that know the number of queries they have made. Clearly, lower bounds that hold in this more general class, also hold for the original class.

We define a random instance I_p from the class BBFunnel as follows. The bitstring z is chosen uniformly at random. Furthermore, the simple path p is constructed by the following randomised algorithm (see (Droste et al., 2006)): Define a sequence R of $b = n^{2\delta}\mu$ points for any constant $\delta \in (0,1)$ and where $m := v - u$ and $r_0 := 0^m$, and for all $i \leq b$, r_{i+1} equals r_i except for one bit-position chosen uniformly at random. From this random sequence, we obtain a simple path where $p_0 := r_0$, and for $i \geq 0$, if $p_i = r_j$, then $p_{i+1} := r_{k+1}$ where $k \geq j$ is the largest index where $r_j = r_k$. This construction ensures that $p_0, \ldots, p_\ell$ is a simple path, where ℓ is the (random) length of the induced path p. From Lemma 1, it follows that the path length satisfies $\ell \geq n^\delta \mu$ with probability $1 - 2^{-\Omega(n^\delta)}$. The following lemma is a variant of Theorem 8 in Droste et al. (2006), but uses a different proof idea.

Lemma 1. *For $d \leq m/5$, and any bitstring $z \in \{0,1\}^m$,*

$$\Pr\left(H(z, r_{i+k}) \leq d \mid H(z, r_i)\right) = 2^{d - H(z,r_i) - ck} + e^{-\Omega(m)},$$

where the random sequence $r_0, \ldots, r_b$, is distributed as described above, $i \in [0..b-1]$, $k \in [1..b-i]$, and $c > 0$ is a constant.

Lemma 1 can be used to prove that BBFunnel belongs to the class $\textsc{SparseLocalOpt}_{\alpha,\varepsilon}$. Due to the page limit, the proof is omitted.

Theorem 7. *For any constants $\sigma, \varepsilon \in (0,1)$, with probability $1-\sigma$, a function sampled from the class* BBFunnel *with $w \in \Theta(n), v - u \leq n/2$ and $\ell \in \mathrm{poly}(n)$ according to distribution I_p belongs to the problem class $\textsc{SparseLocalOpt}_{\alpha,\varepsilon}$ with parameter $\alpha = 1 - w/n$.*

In the following, let $\mathbf{A}$ be any deterministic, elitist $(\mu+\lambda)$ black box algorithm. All instances of BBFunnel define the fitness values in regions $A \cup A' \cup B'$ identically. So we can assume that an optimal deterministic algorithm $\mathbf{A}$ has hard coded the fitness values in these regions, and that it knows the starting point p_1 of the path in B.

In order to reach the optimum $y = 1^u \cdot p_\ell \cdot z \cdot 1^{n-w}$, the algorithm must discover the end point p_ℓ of the random path and the hidden sub-string z. Lemma (2) implies that any black-box algorithm must explore large parts of the path before the end point p_ℓ is discovered.

Lemma 2. *For each $t \in \mathbb{N}$, let M_t be the set of points visited up to step t of algorithm $\mathbf{A}$. Assume that an instance of* BBFunnel *is sampled from the distribution I_p and the algorithm after step $t \leq e^{cd(n)}$ knows the path points $p_0, \ldots, p_i$ but no further points on the path. Then $\Pr\left(p_{i+d(n)} \in M_{t+1}\right) = e^{-\Omega(d(n))}$, if $c > 0$ is small enough.*

Proof. For any $z \in M_{t+1}$, Lemma 1 with $d = 0$ implies

$$\begin{aligned}\Pr\left(z = p_{i+d(n)}\right) &= \Pr\left(H(z, p_{i+d(n)}) \leq 0\right) \\ &= 2^{-H(z,p_i) - \Omega(d(n))} + e^{-\Omega(m)} = e^{-\Omega(d(n))}.\end{aligned}$$

By a union bound over all $t+1 \leq e^{cd(n)}+1$ elements in M_{t+1},

$$\Pr\left(p_{i+d(n)} \in M_{t+1}\right) \leq (t+1)e^{-\Omega(d(n))} = e^{-\Omega(d(n))}. \quad \square$$

Unless the algorithm knows the end point p_ℓ, the fitness function does not reveal any information about the hidden sub-string z. However, if the path length is at least $n^\delta\mu$, any elitist $(\mu+\lambda)$ EA must with high probability produce at least μ path points before the end point p_ℓ is discovered. From this point, the only way to discover the hidden sub-string z is to evaluate search points in region C. However, due to the elite nature of the algorithm, it is prevented from evaluating any search point. The only way for the algorithm to reach region D is to guess the bitstring z correctly.

Theorem 8. *For any constants $\alpha, \varepsilon \in (0,1)$ and $\mu, \lambda \in \text{poly}(n)$, the elitist $(\mu+\lambda)$ black-box complexity of* $\textsc{SparseLocalOpt}_{\alpha,\varepsilon}$ *is $e^{\Omega(n)}$.*

Proof. We apply Theorem 6 and consider the average case runtime of any deterministic elitist $(\mu+\lambda)$ black-box algorithm $\mathbf{A}$ wrt the following distribution I'_p over $\textsc{SparseLocalOpt}_{\alpha,\varepsilon}$.

We construct I'_p from the distribution I_p over BBFunnel with path length $b = \mu n^{2\delta} \in \text{poly}(n)$, and parameters $w := (1-\alpha)n, v := (2/3)(1-\alpha)n$, and $u := \max(1/3(1-\alpha)n, v - n/2)$. Note that $m := v - u \leq n/2$. Given a function f sampled according to distribution I_p, let F be the event that f has simple path length $\ell < n^\delta\mu$ or $f \notin \textsc{SparseLocalOpt}_{\alpha,\varepsilon}$. By Theorem 7, Lemma 1 and a union bound, the probability of event F is less than $\sigma + 2^{-\Omega(n^\delta)}$ for any constant $\sigma \in (0,1)$. To sample from distribution I'_p, we sample f according to I_p, and return $\textsc{OneMax} \in \textsc{SparseLocalOpt}_{1,0}$ if F occurs, and f otherwise.

When event F occurs, we use the lower bound $T(I'_p, \mathbf{A}) \geq 0$. Otherwise, if event F does not occur, then by by Lemma 2, with probability $1 - e^{-\Omega(n^\delta)}$, algorithm $\mathbf{A}$ must query at least $\ell/n^\delta \geq \mu$ of the points in the path region B before it obtains the final point p_ℓ. The fitness of these search points is higher than any search point in C, hence once the algorithm has obtained path point p_ℓ, the population contains only points in B. With μ elements in B, the algorithm cannot base any further decisions on the fitness values in region C. Hence, in order to reach region D, the algorithm must find the hidden sub-string z. The probability that z is found in any of the $e^{\Omega(n)}$ next queries made by the algorithm is exponentially small. Thus, if F does not occur, $T(I'_p, \mathbf{A}) = e^{\Omega(n)}$.

By the law of total probability wrt F, $\min_{\mathbf{A}\in\mathcal{A}} T(I'_p, \mathbf{A}) = e^{\Omega(n)}$, which by Yao's principle implies the theorem. $\square$

4.2 Elitist EAs with diversity mechanisms

It has been shown in Dang et al. (2016) that on the $\textsc{Jump}_k$ function, diversity mechanisms can improve the performance of $(\mu+1)$ GA. This base algorithm is outlined as Algorithm 4. As a corollary of Theorem 8, it holds that many of those mechanisms, denoted by the set $L11$ as they alter the tie-breaking rule in line 11,

are not helpful to escape the local optimum of BBFunnel. Specifically, the set $L11 := \{$*duplicate elimination*, *duplicate minimisation*, *convex hull maximisation*, *Hamming distance maximisation*, *deterministic crowding*$\}$.

Algorithm 4 The (μ+1) GA on $\{0,1\}^n$ space.

Require: State space $\{0,1\}^n$, population size $\mu \in \mathbb{N}$, crossover probability p_c, and mutation parameter χ.

Initialise P_0 with μ individuals uniformly at random in $\{0,1\}^n$
for $t = 0, 1, 2, \ldots$ until termination condition met **do**
  Sample $x \sim \text{Unif}(P_t)$ and sample $p \sim \text{Unif}([0,1])$
  if $p \leq p_c$ **then**
    Sample $y \sim \text{Unif}(P_t)$.
    Set $z \leftarrow$ by copying independently bit-by-bit either from x or from y with equal probability.
  else
    Set $z \leftarrow x$.
  end if
  Flip each bit of z with probability χ/n independently.
  Set $P_{t+1} \leftarrow P_t \cup \{z\}$, then remove one element from P_{t+1} with the lowest fitness, breaking ties at random.
end for

Theorem 9. *There exists a* BBFunnel *function such that the expected optimisation time using any (μ+1) GA as in Algorithm 4 with $\mu = \text{poly}(n)$ and any parameters p_c and χ, and using one of the diversity mechanisms from the set $L11$, or none of them, is exponential.*

5 Non-elitist Algorithms

We now develop a set of sufficient conditions for non-elitist EAs using bitwise mutation to be efficient on $\textsc{SparseLocalOpt}_{\alpha,\varepsilon}$ (see Theorem 11). These conditions require that the selection mechanism has a non-linear β-function. As examples, we demonstrate that these conditions hold for Algorithm 1 (non-elitist EA) using 3-tournament selection (Corollary 12) and linear ranking selection (Corollary 13), and mutation rate close to the error threshold.

5.1 Conditions for Efficiency of Non-Elitist EAs

The level-based theorem Corus et al. (2018) is a tool for deriving upper bounds on the expected runtime of Algorithm 2. To derive sufficient conditions for the efficiency of non-linear EAs on $\textsc{SparseLocalOpt}_{\alpha,\varepsilon}$, we generalise the theorem by introducing a "deceptive region" B. The new conditions (G1) and (G2) are weakened (compared to those in Corus et al. (2018)), and only required to hold if there few individuals in the deceptive region. A new condition (G0) requires that

the number of deceptive individuals reduces in expectation by a $(1-\delta)$-factor if it is above some threshold $\psi_0\lambda$. This variant of the level-based theorem has been implicitly used before, including Dang and Lehre (2016b) and Case and Lehre (2020).

Theorem 10. *Consider Algorithm 2 with population size λ. Given a partition $(A_1,\dots,A_m)$ of $\mathcal{X}$ and a subset $B \subset \mathcal{X}$, define $T := \min\{t\lambda \mid |P_t \cap A_m| > 0\}$, where for all $t \in \mathbb{N}$, $P_t \in \mathcal{X}^\lambda$ is the population in generation t. If there exist $z_1,\dots,z_{m-1},\delta \in (0,1]$, and $\psi_0,\gamma_0 \in (0,1)$ such that for any population $P \in \mathcal{X}^\lambda$, $y \sim D(P)$, any $j \le m-1$, any $\gamma \le \gamma_0$, and any $\psi \ge \psi_0$,*

- **(G0)** *If $|P \cap B| \le \psi\lambda$ then $\Pr(y \in B) \le (1-\delta)\psi$,*
- **(G1)** *If $|P \cap B| \le \psi_0\lambda$ and $|P \cap A_{\ge j}| \ge \gamma_0\lambda$, then $\Pr(y \in A_{\ge j+1}) \ge z_j$,*
- **(G2)** *If $|P \cap B| \le \psi_0\lambda$ and $|P \cap A_{\ge j}| \ge \gamma_0\lambda$ and $|P \cap A_{\ge j+1}| \ge \gamma\lambda$, then $\Pr(y \in A_{\ge j+1}) \ge (1+\delta)\gamma$,*
- **(G3)** $\lambda \ge \left(\frac{12}{\gamma_0\delta^2}\right)\ln\left(\frac{300m}{z_*\delta^2}\right)$, *where $z_* := \min_j z_j$,*

then $E[T] \le \frac{12\lambda}{\delta} + \left(\frac{96}{\delta^2}\right)\sum_{j=1}^{m-1}\left(\lambda\ln\left(\frac{6\delta\lambda}{4+z_j\delta\lambda}\right)+\frac{1}{z_j}\right)$.

Proof. We call individuals in B *deceptive individuals*. To apply the level-based theorem Corus et al. (2018), we first prove that the number of deceptive individuals quickly drops below $\psi_0\lambda$. For any $t_0 \in \mathbb{N}$, define $Y_t := |P_{t_0+t} \cap B|$. Hence, by (G0), Y_{t+1} is stochastically dominated by the random variable $Z \sim \mathrm{Bin}(\lambda, p_s)$ where $p_s := \max(\psi_0, Y_t/\lambda)(1-\delta)$. Then by a Chernoff bound (see e.g. Dubhashi and Panconesi (2009)),

$$\begin{aligned}\Pr(Y_{t+1} \ge \max(\psi_0\lambda,(1-\delta/2)Y_t)) &< \Pr(Z \ge (1-\delta/2)\max(\psi_0\lambda, Y_t)) \\ &= \Pr(Z \ge E[Z](1+\delta/(2(1-\delta)))) \\ &\le \exp\left(-\frac{\delta^2\max(\psi_0\lambda,Y_t)}{12(1-\delta)}\right) \le e^{-\frac{\delta^2\psi_0\lambda}{12(1-\delta)}}. \quad (1)\end{aligned}$$

We now consider phases of length $\tau_1 + 2\tau_2$ generations, where

$$\tau_1 := \frac{\log(\psi_0)}{\log(1-\delta/2)} \le \frac{(1-\psi_0)(1-\delta/2)}{\delta/2} < \frac{2}{\delta}. \quad (2)$$

(Note that $\frac{x-1}{x} \le \ln(x)$ for all $x > 0$.) and

$$\tau_2 := \left(\frac{8}{\delta^2}\right)\sum_{j=1}^{m-1}\left(\ln\left(\frac{6\delta\lambda}{4+z_j\delta\lambda}\right)+\frac{1}{z_j\lambda}\right) \quad (3)$$

$$\le \frac{8m}{\delta^2}\left(\ln(6/z_*)+\frac{1}{\lambda z_*}\right) < \frac{8m}{\delta^2 z_*}\left(6+\frac{1}{\lambda}\right) < \frac{146m}{3\delta^2 z_*} \quad (4)$$

where the last inequality applied $\lambda > 12$ which follows from condition (G3). Note that τ_2 is the expected time to reach level m in the original level-based theorem Corus et al. (2018). Now by Eqs. (1),(2),(4), condition (G3), and a union bound, with probability no more than

$$(\tau_1 + 2\tau_2)e^{-\frac{\delta^2\psi_0\lambda}{12(1-\delta)}} \leq \left(\frac{2}{\delta} + \frac{292m}{3\delta^2 z_*}\right)\left(\frac{z_*\delta^2}{300m}\right) < \frac{1}{3}, \tag{5}$$

after τ_1 generations of the phase, there are still $\lambda\,(1-\delta/2)^{\tau_1} = \psi_0\lambda$ deceptive individuals. By the Markov's inequality and the level-based theorem Corus et al. (2018), with probability no more than $1/2$, the algorithm fails to reach level m after $2\tau_2$ generations of the phase. Inversely, by a union bound and Eq. (5), a phase is successful with probability $1 - 1/2 - 1/3 = 1/6$. If the phase is unsuccessful, we can apply the same arguments to the next phase, since we have not assumed anything about the initial state of the population.

Hence, a successful phase occurs in expectation after at most 6 phases, i.e., $6(\tau_1 + 2\tau_2)$ generations. The theorem now follows because each generation produces λ individuals. □

The following lemmas describe the behaviour of the bitwise mutation operator in dense and sparse regions of the search space.

Lemma 3. *If C is an α-dense set, then*

$$\Pr\left(p_{\text{mut}}(x) \in C \mid x \in C\right) > (1-\chi/n)^n\,(1+\alpha\chi). \tag{6}$$

Lemma 4. *If B is ε-sparse with $\varepsilon = \frac{\rho-(1-\chi/n)^n}{1-(1-\chi/n)^n}$, then*

- $\Pr\left(p_{\text{mut}}(x) \in B \mid x \in B\right) \leq \rho$, *and*
- $\Pr\left(p_{\text{mut}}(x) \in B \mid x \notin B\right) = O(1/n)$.

Theorem 11. *If there exist constants $\varepsilon, \psi_0, \gamma_0, \delta, \alpha \in (0,1)$ such that Algorithm 1 with the bitwise mutation operator with rate χ, and a selection mechanism with β, and population size λ satisfying*

(SM0) $\beta(0,\gamma) \leq \frac{\gamma}{\frac{\varepsilon}{1-\varepsilon}+\left(1-\frac{\chi}{n}\right)^n}$ *for all* $\gamma \in [\psi_0, 1]$,

(SM2a) $\beta(0,\gamma) \geq \frac{\gamma(1+\delta)}{\left(1-\frac{\chi}{n}\right)^n}$ *for all* $\gamma \in (0, \gamma_0]$,

(SM2b) $\beta(\psi, \psi+\gamma) \geq \frac{\gamma(1+\delta)}{\left(1-\frac{\chi}{n}\right)^n(1+\alpha\chi)}$ *for all* $\gamma \in (0,\gamma_0]$, $\psi \in [0, \psi_0]$,

(SM3) $c\ln(n) \leq \lambda \in \text{poly}(n)$ *for a sufficiently large constant c,*

then it has expected polynomial runtime on $\textsc{SparseLocalOpt}_{\alpha,\varepsilon}$.

Proof. Consider any function f in the problem class, with an associated partition $(A_1, \ldots, A_m)$ and f-deceptive pairs (A_{i_1}, A_{j_1}), ..., (A_{i_u}, A_{j_u}). We will apply Theorem 10 with respect to $B := \cup_{v=1}^{u} A_{j_v}$, which by the definition of the problem class is ε-sparse. Assume that $x = P(i)$ where $i \sim p_{\mathrm{sel}}(P)$ and $y \sim p_{\mathrm{mut}}(x)$. For any $\gamma \geq \psi_0$, we have by Lemma 4, condition (SM0), and the law of total probability

$$\begin{aligned}\Pr(y \in B) &\leq \Pr(x \in B) \Pr(y \in B \mid x \in B) \\ &\quad + \Pr(x \notin B) \Pr(y \in B \mid x \notin B) \\ &\leq \beta(0,\gamma)\varepsilon + \mathrm{O}(1/n) \leq \gamma(1-\varepsilon) + \mathrm{O}(1/n).\end{aligned}$$

Hence, for large n, condition (G0) is satisfied for any constant $\varepsilon' < \varepsilon$.

Assume that $|P \cap B| \leq \psi_0\lambda$ and $|P \cap A_{\geq j}| \geq \gamma_0\lambda$. Then, to produce an individual y in $A_{\geq j+1}$, it suffices to select an individual $x \in A_{\geq j}$, and flip $d = \mathrm{O}(1)$ bit-positions. Except for at most $\psi_0\lambda$ individuals in B, the individuals in the set $A_{\geq j}$ have higher fitness than all other individuals. Thus, by condition (SM2b), we get

$$\begin{aligned}\Pr(y \in A_{\geq j+1}) &\geq \Pr(x \in A_{\geq j}) \Pr(y \in A_{\geq j+1} \mid x \in A_{\geq j}) \\ &\geq \beta(\psi_0, \psi_0+\gamma_0) (\chi/n)^d (1-\chi/n)^{n-d} \\ &\geq \frac{\gamma_0(1+\delta)}{(1-\chi/n)^n (1+\alpha\chi)} (\chi/n)^d (1-\chi/n)^{n-d} =: z_j.\end{aligned}$$

Hence, condition (G1) is satisfied for the parameter $z_j = n^{-O(1)}$.

Assume that $|P \cap B| \leq \psi_0\lambda$, $|P \cap A_{\geq j}| \geq \gamma_0\lambda$ and $|P \cap A_{\geq j+1}| \geq \gamma\lambda$, then except for at most $\psi_0\lambda$ individuals in B, the individuals in $A_{\geq j+1}$ have higher fitness than all others in P. We consider two cases. If A_{j+1} is a part of a deceptive pair, then by the problem definition, $A_{\geq j+1}$ is an α-sparse set. Lemma 3 and (SM2b) then imply

$$\begin{aligned}\Pr(y \in A_{\geq j+1}) &\geq \Pr(x \in A_{\geq j+1}) \Pr(y \in A_{\geq j+1} \mid x \in A_{\geq j+1}) \\ &\geq \beta(\psi_0, \psi_0+\gamma) (1-\chi/n)^n (1+\alpha\chi) \geq \gamma(1+\delta).\end{aligned}$$

In case A_{j+1} is not part of a deceptive pair, the individuals in $A_{\geq j+1}$ are fitter than any other individual in the population, and to produce an individual in $A_{\geq j+1}$, it suffices to select a one in $A_{\geq j+1}$ and not flip any bits, which by (SM2a) occurs with probability

$$\begin{aligned}\Pr(y \in A_{\geq j+1}) &\geq \Pr(x \in A_{\geq j+1}) \Pr(y = x) \\ &\geq \beta(0,\gamma) (1-\chi/n)^n \geq \gamma(1+\delta).\end{aligned}$$

Hence, condition (G2) of Theorem 10 is satisfied. Finally, (G3) follows immediately from (SM3) for some constant c since γ_0 and δ are constants and $m \in \mathrm{poly}(n)$. All conditions are satisfied, thus the expected runtime on f is $\mathrm{O}\left(\sum_{j=1}^{m-1} \lambda \ln(1/z_j) + 1/z_j\right) \in \mathrm{poly}(n)$. □

The following corollaries of Theorem 11 give example configurations where non-elitist EAs are efficient on $\textsc{SparseLocalOpt}_{\alpha,\varepsilon}$. The proofs are omitted due to the page limit. Future work should provide a more comprehensive analysis of the algorithmic configurations which satisfy the conditions of Theorem 11.

Corollary 12. *Algorithm 1 with 3-tournament selection, population size* $c\ln(n) \leq \lambda \in \text{poly}(n)$ *for a sufficiently large c, and mutation rate* $\chi = 1.09812$ *has polynomial worst case expected runtime on* $\textsc{SparseLocalOpt}_{\alpha,\varepsilon}$ *with* $\alpha = 1/4$ *and* $\varepsilon = 7/10^5$.

Corollary 13. *Algorithm 1 with linear ranking selection Lehre and Yao (2012) for* $\eta = 2$*, population size* $c\ln(n) \leq \lambda \in \text{poly}(n)$ *for a sufficiently large c, and mutation rate* $\chi = 0.693146$ *has polynomial worst case expected runtime on* $\textsc{SparseLocalOpt}_{\alpha,\varepsilon}$ *with* $\alpha = 4/9$ *and* $\varepsilon = 1/100$.

5.2 The (μ, λ) EA is inefficient

The (μ, λ)-selection mechanism has a linear β-function, and therefore does not satisfy the requirements imposed by Theorem 11. We will now show that the non-elitist (μ, λ) EA is inefficient on BBFunnel, thus by extension on the problem $\textsc{SparseLocalOpt}_{\varepsilon,\alpha}$. To achieve this, it suffices to prove that (μ, λ) EA requires exponential runtime with overwhelmingly high probability on a specific BBFunnel function, denoted as f in the rest of the section, where $z = 1^{w-v}$ and the simple path is $(p_i)_{0\leq i\leq v-u}$ with $p_i = 1^i 0^{v-u-i}$.

Algorithm 1 has exponential runtime on any function with a unique optimum if the mutation rate χ exceeds $\ln(\alpha_0)+\sigma$ (called the error threshold), where α_0 is the reproductive rate of the algorithm, and for an arbitrarily small constant $\sigma \in (0, 1)$ Lehre (2010). For the (μ, λ) EA, the error threshold occurs when $\chi \geq \ln(\lambda/\mu)+\sigma$. We therefore consider (μ, λ) EA for mutation rates $\chi \leq \ln(\lambda/\mu) - \sigma$, for any constant $\sigma \in (0, 1)$. We will show that the population will quickly produce μ individuals in region B. From then on, all individuals will with high probability have a parent in B, and it will take exponential time to produce an offspring in D.

To simplify the derivations, we will only count generations where the population has an individual in region $\mathcal{Y} := A \cup B \cup C \cup D$. This will only under-estimate the runtime of the algorithm. Lemmas 5 and 6 show that it is unlikely that (μ, λ) EA optimises the B-region before the population contains μ individuals in B.

Lemma 5. *If* $\chi \leq \ln(\lambda/\mu) - \sigma$ *for any constant* $\sigma > 0$*, then with probability* $1 - t_0 e^{-\frac{\mu\sigma^2}{2(1+\sigma)}} - \lambda e^{-\Omega(n)}$*, every population* P_t *of* (μ, λ) *EA, where* $t \leq t_0$*, contains at least* μ *individuals in* $\mathcal{Y} := A \cup B \cup C \cup D$ *of function* f *with* $u \geq (1 + \varepsilon)n/2$ *for any constant* $\varepsilon > 0$.

Proof. We prove the statement by induction on t. By a Chernoff bound and a union bound, all individuals in P_0 belong to A with probability $1 - \lambda e^{-\Omega(n)}$. Assume that at least μ individuals in generation $t \leq t_0 - 1$ belong to region

$\mathcal{Y}$. By the definition of the BBFunnel and of (μ,λ)-selection, any individual created in P_{t+1} has a parent in region $\mathcal{Y}$. So the probability that an offspring belongs to region $\mathcal{Y}$ is at least the probability that no bit is mutated, i.e., at least $(1-\chi/n)^n \geq e^{-\chi}(1-o(1)) \geq (\mu/\lambda)e^{\sigma}(1-o(1)) \geq (\mu/\lambda)(1+\sigma)$ where the last inequality follows from the fact that $e^x \geq 1+x+\frac{x^2}{2}$ for any $x \geq 0$. Hence, the expected number of offspring in region $\mathcal{Y}$ is at least $\mu(1+\sigma)$, and by a Chernoff bound, the probability that there are less than μ such individuals is less than $e^{-\frac{\mu\sigma^2}{2(1+\sigma)}}$. The statement now follows by a union bound over t_0 generations. □

Lemma 6. *Let $u \geq (n/2)(1+\varepsilon)$ for any constant $\varepsilon > 0$, $v-u = \Omega(n)$ and define $\alpha := (v-u)/n$. Let P_t be the population in generation t of the (μ,λ) EA with $\lambda = \text{poly}(n)$ and $\chi \leq \ln(\lambda/\mu) - \sigma$ for a constant $\sigma > 0$ on function f. Define $X_t := \max_{i\in[\lambda]} \textsc{Lo}(P_t(i))$. For $T := \min\{t \in \mathbb{N} \mid X_t \geq v - \alpha n/3\}$, it holds $\Pr\left(T \leq n^{1-2\delta}\right) = e^{-\Omega(n^\delta)} + e^{-\frac{\mu\sigma^2}{2(1+\sigma)}+(1-2\delta)\ln(n)}$ for any constant $\delta \in (0,1/2)$.*

Proof. We say that *failure event 1* occurs if there exists a generation $t \leq n^{1-2\delta}$ where less than μ individuals in P_t belong to $\mathcal{Y} := A \cup B \cup C \cup D$. By Lemma 5, the probability of failure event 1 is less than $e^{-\frac{\mu\sigma^2}{2(1+\sigma)}+(1-2\delta)\ln n} + e^{-\Omega(n)}$ given that $\lambda = \text{poly}(n)$.

We say that *failure event 2* occurs if any individual in A is mutated into region $C \cup D$ within the first $n^{1-2\delta}$ generations. In a single step, such a mutation has probability of at most $\left(\frac{\chi}{n}\right)^{\alpha n} \leq 2^{-\Omega(n)}$ since at least αn 0-bits have to be flipped to 1. By a union bound, the failure probability of the event is no more than $\lambda n^{1-2\delta} 2^{-\Omega(n)} = e^{-\Omega(n)}$.

We now assume that failure events 1 and 2 did not occur. Let t_1 be the first generation where the population contains an individual in region B. Firstly, we notice the probability that X_{t_1} exceeds $u+\alpha n/3$ is no more than $\lambda 2^{-\Omega(n)} = e^{-\Omega(n)}$ by a union bound since the easiest way to create an individual x in B with $\textsc{Lo}(x) \geq u + \alpha n/3$ from one in A still requires to flip at least $\alpha n/3$ specific bits. We call such excess *failure event 3*. Secondly, for a lower bound we can assume optimistically that for any generation $t \geq t_1$, some individual $x \in B$ with $\textsc{Lo}(x) = X_t$ is always picked as parent. Even with this, the probability of making a large improvement of n^δ in fitness within B is less than $(\chi/n)^{n^\delta} = 2^{-\Omega(n^\delta)}$ because n^δ specific bit-positions have to be flipped. We call any such mutation within the first $n^{1-2\delta}$ generations *failure event 4* and by a union bound, its probability is no more than $\lambda n^{1-2\delta} 2^{-\Omega(n^\delta)} = e^{-\Omega(n^\delta)}$.

If failure events 1-4 do not occur, then $X_t - (u+\alpha n/3) \leq X_t - X_{t_1} \leq t n^\delta \leq n^{1-\delta} < (1/3)\alpha n$ for all $t \leq n^{1-2\delta}$ and sufficiently large n. The lemma follows by noting that $u + 2\alpha n/3 = v - \alpha n/3$. □

We will estimate the expected "upgrade time" for the number of individuals X_t of (μ,λ) EA in $B \cup C \cup D$ by Lemma 7. If $\chi \leq \ln(\lambda/\mu) - \sigma$, then the probability to produce an individual in $B \cup C \cup D$ is at least $(X_t/\mu)\,(1-\chi/n)^n \geq (X_t/\mu)e^{-\chi}(1-o(1)) \geq (X_t/\lambda)(1+\sigma)$. For a lower bound, we only count the latest sub-sequence of iterations where $X_t \geq 1$ with no preemptions. If $\mu \in \text{poly}(n)$,

Lemma 7 implies that with probability $1 - 2^{-n^c}$ for a constant $c > 0$, region $B \cup C \cup D$ takes over the population within at most $n^{1-2\delta}$ iterations.

Lemma 7. *Let $\mu, \lambda \in \mathbb{N}$ where $\mu \leq \lambda/(1+\delta)$ for some $\delta > 0$. Assume a stochastic process $(X_t)_{t\in\mathbb{N}}$ is defined by $X_t := \max(1, Z_t)$ where for all $t \in \mathbb{N}$ and $X_t < \mu$, $Z_t \sim \text{Bin}(\lambda, (1+\delta)X_{t-1}/\lambda)$, and $X_0 \in \mathbb{N}$. Let $T := \min\{t \mid X_t \geq \mu\}$, then $E[T] \leq (7/\delta^2)\ln(1 + (\delta/2)\mu)$. Furthermore $\Pr\left(T \geq r(14/\delta^2)\ln(1 + (\delta/2)\mu)\right) \leq 2^{-r}$ for any $r \in \mathbb{N}$.*

Proof. Let a distance function be $g(x) := \ln\left(\frac{1+(\delta/2)\mu}{1+(\delta/2)x}\right)$. By Lemma 6 from Corus et al. (2018), the expected drift w.r. t. g is at least $\frac{\delta^2}{7}$. The upper bound on $E[T]$ now follows by Theorem 2 Corus et al. (2018), which is a variant of the additive drift theorem He and Yao (2001). By Markov inequality, $X_t \geq \mu$ at least for one t in any phase of length $2E[T]$ with probability at least $\frac{1}{2}$. The tail bound follows by considering r phases. □

We now show that the (μ, λ) EA is inefficient on BBFunnel.

Theorem 14. *Let $\sigma, \varepsilon \in (0,1)$ be constants. The runtime T of (μ,λ) EA with population sizes $\lambda \in \text{poly}(n)$, $\mu \geq \frac{2(1+\sigma)}{\sigma^2}\ln(n)$, and mutation rate parameter $\chi \leq \ln(\lambda/\mu) - \sigma$ on* BBFunnel *with $v - u = \Omega(n)$, $w - v = \Omega(n)$ and $u \geq (1+\varepsilon)n/2$ satisfies $\Pr(T \leq e^{c\min(\mu,n)}) \leq e^{-\Omega(\mu)} + e^{-\Omega(n^d)}$ for some constants $c, d > 0$.*

Proof. It suffices to give a proof for function f. Let us note that

$$\mu\sigma^2/(2(1+\sigma)) - (1-2\delta)\ln(n) = \Omega(\mu), \tag{7}$$

and count generations with some individuals in $\mathcal{Y} := A \cup B \cup C \cup D$.

We will prove a stronger statement that with probability $1 - e^{-\Omega(n^d)} - e^{-\Omega(\mu)}$, none of the search points produced during the first $e^{c\min(\mu,n)}$ generations are in region D for some constants c, d.

We let *phase 1* be the first $n^{1-2\delta}$ generations. By Lemma 5 (failure probability $e^{-\Omega(\mu)} + e^{-\Omega(n)}$ by (7)), Lemma 6 (failure probability $e^{-\Omega(n^\delta)} + e^{-\Omega(\mu)}$), and Lemma 7 (failure probability $e^{-\Omega(n^c)}$) and a union bound, with probability at least $1 - e^{-\Omega(\mu)} - e^{-\Omega(n^{c'})}$, by the end of Phase 1, the population consists of at least μ individuals in region B, and no individual has more than $v - (1/3)\alpha n$ leading 1-bits, where $\alpha := (v-u)/n$. For a constant $c > 0$ to be chosen later, let *phase 2* to be the $e^{c\min(\mu,n)}$ generations after phase 1. A generation in phase 2 is said to *fail* if there are less than μ offspring in B, or there is a mutation from region B into region D.

Since the μ best individuals are in B, all offspring have parents in B. To create an offspring in B, it suffices to not flip any bit, i.e., with probability $(1-\chi/n)^n \geq (1-\sigma')e^{-\chi} \geq (\mu/\lambda)(1-\sigma')e^{\sigma}$ for any constant $\sigma' \in (0,1)$, if n is large enough. Choosing σ' so that $(1+\sigma')/(1-\sigma') = e^{\sigma}$, this probability is at least $(1+\sigma')(\mu/\lambda)$. By a Chernoff bound, the probability that the population has less than μ individuals from B in the next generation is $e^{-\Omega(\lambda)} = e^{-\Omega(\mu)}$. For

those individuals to be the best of the population, no individual in D must be created. The probability of creating a D-individual by mutating a B-individual is $n^{-\Omega(n)}$. By a union bound, this event occurs with probability less than $\lambda \cdot n^{-\Omega(n)} = n^{-\Omega(n)}$ in any generation. By induction and a union bound, a failure occurs in phase 2 only with probability $e^{-\Omega(\mu)} + e^{-\Omega(n)}$. □

6 Conclusion

We have presented SPARSELOCALOPT$_{\alpha,\varepsilon}$, a general class of functions, in which depending on its parameters (α, ε), the deceptive regions of the landscape are sparse while the fitness valleys are dense. By introducing a new version of the level-based theorem, we have deduced sufficient conditions for non-elitist EAs to have polynomial expected runtime on SPARSELOCALOPT$_{\alpha,\varepsilon}$, and indicated some algorithms satisfying these conditions.

Conversely, we have shown that elitist black-box (μ+λ) EAs can meet a significant difficulty, i. e. get trapped in the local optima of functions from the sub-class BBFUNNEL. This result covers a large set of algorithms, including elitist genetic algorithms with different types of variation operators and even with the use of standard diversity mechanisms. We have shown unconditionally that the non-elitist (μ, λ) EA fails to optimise the BBFUNNEL problems in expected polynomial time, demonstrating an exponential performance differences between different non-elitist selection mechanisms.

In summary, the paper proves that non-elitist EAs can outperform elitist EAs by exponential factors on fitness landscapes with highly rugged and multi-modal structure. The results suggest that non-elitism should be considered more often when applying EAs to complex combinatorial optimisation problems.

Acknowledgements

Eremeev was supported by the Russian Science Foundation grant 17-18-01536. Lehre was supported by a Turing AI Fellowship (EPSRC grant ref EP/V025562/1).

References

Christof K. Biebricher and Manfred Eigen. 2005. The error threshold. *Virus Research* 107, 2 (2005), 117–127. https://doi.org/10.1016/j.virusres.2004.11.002

Brendan Case and Per Kristian Lehre. 2020. Self-adaptation in non-Elitist Evolutionary Algorithms on Discrete Problems with Unknown Structure. *IEEE Transactions on Evolutionary Computation* 24, 4 (2020), 650–663. https://doi.org/10.1109/TEVC.2020.2985450

Dogan Corus, Duc-Cuong Dang, Anton V. Eremeev, and Per Kristian Lehre. 2018. Level-Based Analysis of Genetic Algorithms and Other Search Processes. *IEEE Trans. Evolutionary Computation* 22, 5 (2018), 707–719.

Duc-Cuong Dang, Anton Eremeev, and Per Kristian Lehre. 2021. Escaping Local Optima with Non-Elitist Evolutionary Algorithms. In *Proceedings of AAAI'2021*. http://34.94.61.102/paper_AAAI-6811.html

Duc-Cuong Dang, Tobias Friedrich, Timo Kötzing, Martin S. Krejca, Per Kristian Lehre, Pietro Simone Oliveto, Dirk Sudholt, and Andrew M. Sutton. 2016. Escaping Local Optima with Diversity Mechanisms and Crossover. In *Proceedings of the 2016 Genetic and Evolutionary Computation Conference (GECCO 2016)*. ACM, 645–652. https://doi.org/10.1145/2908812.2908956

Duc-Cuong Dang, Tobias Friedrich, Timo Kötzing, Martin S. Krejca, Per Kristian Lehre, Pietro Simone Oliveto, Dirk Sudholt, and Andrew M. Sutton. 2018. Escaping Local Optima Using Crossover With Emergent Diversity. *IEEE Trans. Evolutionary Computation* 22, 3 (2018), 484–497. https://doi.org/10.1109/TEVC.2017.2724201

Duc-Cuong Dang, Thomas Jansen, and Per Kristian Lehre. 2017. Populations Can Be Essential in Tracking Dynamic Optima. *Algorithmica* 78, 2 (2017), 660–680. https://doi.org/10.1007/s00453-016-0187-y

Duc-Cuong Dang and Per Kristian Lehre. 2015. Efficient Optimisation of Noisy Fitness Functions with Population-Based Evolutionary Algorithms. In *Proceedings of the 2015 Conference on Foundations of Genetic Algorithms (FOGA'2015)*. ACM, 62–68. https://doi.org/10.1145/2725494.2725508

Duc-Cuong Dang and Per Kristian Lehre. 2016a. Runtime Analysis of Non-elitist Populations: From Classical Optimisation to Partial Information. *Algorithmica* 75 (2016), 428–461. https://doi.org/10.1007/s00453-015-0103-x

Duc-Cuong Dang and Per Kristian Lehre. 2016b. Self-adaptation of Mutation Rates in Non-elitist Populations. In *Proceedings of the 2016 Conference on Parallel Problem Solving from Nature (PPSN 2016)*. Springer, Cham, 803–813. https://doi.org/10.1007/978-3-319-45823-6_75

Benjamin Doerr. 2020. Does comma selection help to cope with local optima?. In *Proceedings of the 2020 Genetic and Evolutionary Computation Conference (GECCO 2020)*. ACM, New York, NY, USA, 1304–1313.

Carola Doerr and Johannes Lengler. 2016. Introducing Elitist Black-Box Models: When Does Elitist Behavior Weaken the Performance of Evolutionary Algorithms? *Evolutionary Computation* 25, 4 (Oct. 2016), 587–606. https://doi.org/10.1162/evco_a_00195 Publisher: MIT Press.

Stefan Droste, Thomas Jansen, and Ingo Wegener. 1998. On the Optimization of Unimodal Functions with the (1 + 1) Evolutionary Algorithm. In *Proceedings*

of the 1998 Conference on Parallel Problem Solving from Nature (PPSN'1996). Springer-Verlag, Berlin, Heidelberg, 13–22.

Stefan Droste, Thomas Jansen, and Ingo Wegener. 2006. Upper and Lower Bounds for Randomized Search Heuristics in Black-Box Optimization. *Theory of Computing Systems* 39, 4 (July 2006), 525–544. https://doi.org/10.1007/s00224-004-1177-z

Devdatt Dubhashi and Alessandro Panconesi. 2009. *Concentration of Measure for the Analysis of Randomized Algorithms*. Cambridge University Press, NY, USA.

David E. Goldberg and Kalyanmoy Deb. 1991. A comparative analysis of selection schemes used in genetic algorithms. In *Foundations of Genetic Algorithms*. Morgan Kaufmann, 69–93.

Bruce Hajek. 1988. Cooling Schedules for Optimal Annealing. *Math. Oper. Res.* 13, 2 (1988), 311–329.

Jun He and Xin Yao. 2001. Drift analysis and average time complexity of evolutionary algorithms. *Artif. Intell.* 127, 1 (2001), 57–85.

Sebastian Herrmann. 2016. Determining the Difficulty of Landscapes by PageRank Centrality in Local Optima Networks. In *Evolutionary Computation in Combinatorial Optimization*, Francisco Chicano, Bin Hu, and Pablo García-Sánchez (Eds.). Springer International Publishing, Cham, 74–87.

Jeffrey Horn, David E. Goldberg, and Kalyanmoy Deb. 1994. Long Path Problems. In *Proceedings of the 1994 Conference on Parallel Problem Solving from Nature (PPSN'1994) (PPSN III)*. Springer, Berlin, Heidelberg, 149–158.

Jens Jägerskupper and Tobias Storch. 2007. When the Plus Strategy Outperforms the Comma Strategy and When Not. In *2007 IEEE Symposium on Foundations of Computational Intelligence*. 25–32.

Per Kristian Lehre. 2010. Negative Drift in Populations. In *Proceedings of the 2010 Conference on Parallel Problem Solving from Nature (PPSN'2010)*. Springer, Berlin, Heidelberg, 244–253.

Per Kristian Lehre. 2011. Fitness-Levels for Non-Elitist Populations. In *Proceedings of the 2011 Genetic and Evolutionary Computation Conference (GECCO 2011)*. ACM, 2075–2082. https://doi.org/10.1145/2001576.2001855

Per Kristian Lehre and Xin Yao. 2012. On the impact of mutation-selection balance on the runtime of evolutionary algorithms. *IEEE Transactions on Evolutionary Computation* 16, 2 (April 2012), 225–241. https://doi.org/10.1109/TEVC.2011.2112665

Pietro Simone Oliveto, Tiago Paixão, Jorge Pérez Heredia, Dirk Sudholt, and Barbora Trubenová. 2018. How to Escape Local Optima in Black Box Optimisation: When Non-elitism Outperforms Elitism. *Algorithmica* 80, 5 (2018), 1604–1633. https://doi.org/10.1007/s00453-017-0369-2

Pietro Simone Oliveto, Dirk Sudholt, and Christine Zarges. 2019. On the benefits and risks of using fitness sharing for multimodal optimisation. *Theor. Comput. Sci.* 773 (2019), 53–70. https://doi.org/10.1016/j.tcs.2018.07.007

Tiago Paixão, Jorge Pérez Heredia, Dirk Sudholt, and Barbora Trubenová. 2017. Towards a Runtime Comparison of Natural and Artificial Evolution. *Algorithmica* 78, 2 (2017), 681–713. https://doi.org/10.1007/s00453-016-0212-1

C. R. Reeves and A. V. Eremeev. 2004. Statistical Analysis of Local Search Landscapes. *The Journal of the Operational Research Society* 55, 7 (2004), 687–693. http://www.jstor.org/stable/4102015

Günter Rudolph. 1996. How Mutation and Selection Solve Long-Path Problems in Polynomial Expected Time. *Evol. Comput.* 4, 2 (1996), 195–205. https://doi.org/10.1162/evco.1996.4.2.195

Peter F. Stadler. 2002. Fitness landscapes. In *Biological Evolution and Statistical Physics*, Michael Lässig and Angelo Valleriani (Eds.). Springer, Berlin, Heidelberg, 183–204. https://doi.org/10.1007/3-540-45692-9_10

Dirk Sudholt. 2011. Hybridizing Evolutionary Algorithms with Variable-Depth Search to Overcome Local Optima. *Algorithmica* 59, 3 (2011), 343–368. https://doi.org/10.1007/s00453-009-9384-2

Sarah L. Thomson, Fabio Daolio, and Gabriela Ochoa. 2017. Comparing Communities of Optima with Funnels in Combinatorial Fitness Landscapes. In *Proceedings of the Genetic and Evolutionary Computation Conference (GECCO '17)*. Association for Computing Machinery, New York, NY, USA, 377–384. https://doi.org/10.1145/3071178.3071211

Ingo Wegener. 2005. Simulated Annealing Beats Metropolis in Combinatorial Optimization. In *Proceedings of the International Colloquium on Automata, Languages and Programming (ICALP'2005)*, Vol. 3580. Springer, Berlin, Heidelberg, 589–601. https://doi.org/10.1007/11523468_48

Claus O. Wilke. 2005. Quasispecies theory in the context of population genetics. *BMC Evolutionary Biology* 5, 1 (2005), 44. https://doi.org/10.1186/1471-2148-5-44

David Williams. 1991. *Probability with Martingales*. Cambridge University Press.

Sewall Wright. 1932. The roles of mutation, inbreeding, crossbreeding and selection in evolution. In *Proceedings of the sixth international congress on genetics*, Vol. 1. 356–366.

Andrew C. Yao. 1977. Probabilistic computations: Toward a unified measure of complexity. In *18th Annual Symposium on Foundations of Computer Science (sfcs 1977)*. 222–227. https://doi.org/10.1109/SFCS.1977.24 ISSN: 0272-5428.

Christine Zarges. 2011. *Theoretical Foundations of Artificial Immune Systems*. Ph.D. Dissertation. Universität Dortmund.

Appendix

This appendix provides the proofs from the GECCO 2021 paper "*Non-elitist evolutionary algorithms excel in fitness landscapes with sparse deceptive regions and dense valleys*" that did not fit within the 8 page limit.

Lemma 8. *For any constants $\varepsilon, \sigma \in (0,1)$ and sufficiently large n, an instance of* BBFunnel *sampled from the distribution I_p has a set $B = \{1^u p_i 0^{n-v} \mid i \in [0..\ell]\}$ with $\ell \in \text{poly}(n)$ and $v - u \leq n/2$ which is ε-sparse with probability $1 - \sigma$.*

Proof of Lemma 1. We apply Theorem 16 (Hajek's drift theorem) w.r. t. the process $Y_k := m - H(z, r_{i+k})$ for the filtration $\mathcal{F}_k := \sigma(\{H(z, r_{i+j}) \mid j \in [0..k]\})$, and the parameters $a := (3/4)m, b := m - d \geq a + m/20$ and $\eta = \ln(2)$. For $Y_k \geq a$, we have $\Pr(Y_{k+1} - Y_k = 1) \leq 1/4$, thus

$$E_{1,k} \leq \frac{1}{4}e^{\eta} + \frac{3}{4}e^{-\eta} = \frac{7}{8} =: \rho.$$

For $Y_k \leq a$, we have $Y_{k+1} \leq a + 1$, and $E_{2,k} \leq e^{\eta} =: D$. Thus, both conditions of the Hajek's drift theorem are satisfied and

$$\begin{aligned}\Pr(H(z, r_{i+k}) \leq d \mid H(z, r_i)) &= \Pr(Y_k \geq b \mid \mathcal{F}_0) \\ &\leq \rho^k e^{\eta(Y_0 - b)} + e^{\Omega(a-b)} \\ &= 2^{d - H(z,r_i) + k\log_2(\rho)} + e^{-\Omega(m)}.\square\end{aligned}$$

Proof of Lemma 8. By assumption, there exists a constant k such that the set B consists of $\ell \leq n^k$ elements, where each p_j is a bitstring of length $m := v - u$.

We first verify condition (SP1) of Definition 2. Choose any $r_j \in B$ and $r \in [n]$. We will prove that with probability $1 - \mathrm{O}(1/n)$, the randomised path satisfies

$$X := |\{y \in B \mid H(r_j, y) = r\}| \leq \varepsilon \binom{n}{r}. \tag{8}$$

For any $r > m$, Eq. (8) trivially holds, because the Hamming-distance between any pair of elements in B is at most m.

For any r with $k + 1 \leq r \leq m < n/2$ it holds for sufficiently large n that

$$\varepsilon \binom{n}{r} \geq \varepsilon \binom{n}{k+1} \geq \varepsilon \left(\frac{n}{k+1}\right)^{k+1} > n^k \geq |B|,$$

so Eq. (8) holds in this case with probability 1.

Finally, for $1 \leq r \leq k$, define $X := \sum_{i=1}^{\ell} X_i$ where for each $i \in [\ell]$, $X_i := 1$ if $H(r_i, r_j) \leq r$ and $X_i = 0$ otherwise. Then, it holds for any constant $c > 0$

$$E[X] \leq \sum_{i=0}^{\ell} E[X_i] \tag{9}$$

$$< \sum_{i=0}^{j} \Pr\left(H(r_i, r_{i+(j-i)}) \leq r\right) + \sum_{i=j+1}^{\ell} \Pr\left(H(r_j, r_{j+(i-j)}) \leq r\right) \tag{10}$$

$$\leq 2r/c + \sum_{i=0}^{j-r/c} \Pr\left(H(r_i, r_{i+(j-i)}) \leq r\right) \tag{11}$$

$$+ \sum_{i=j+1+r/c}^{\ell} \Pr\left(H(r_j, r_{j+(i-j)}) \leq r\right), \tag{12}$$

and Lemma 1 with the corresponding constant c gives

$$\leq 2r/c + \ell \cdot e^{-\Omega(m)} + 2\sum_{i=0}^{\infty} 2^{-ci} = \mathrm{O}(1). \tag{13}$$

By Markov's inequality, we thus have

$$\Pr\left(X \geq \varepsilon \binom{n}{r}\right) \leq \frac{E[X]}{\varepsilon \binom{n}{r}} = \mathrm{O}(1/n).$$

We now verify condition (SP2) of Definition 2. Choose any $x \notin B$ and $r \in [n]$. We will prove that with probability $1 - \sigma$, the randomised path satisfies

$$Y := |\{y \in B \mid H(x, y) = r\}| \leq \frac{c'}{\sigma n} \binom{n}{r}, \tag{14}$$

where $c' > 0$ is a constant independent of x that will be determined later.

In the case $k + 2 \leq r \leq n/2$, it holds for sufficiently large n

$$\frac{c'}{\sigma n} \binom{n}{r} \geq \frac{c'}{\sigma n} \binom{n}{k+2} \geq \frac{c'}{\sigma n} \left(\frac{n}{k+2}\right)^{k+2} > |B| \geq Y,$$

so Eq. (14) is satisfied with probability 1.

Assume that $1 \leq r \leq k + 1$, and that $r_j \in B$ is one of the elements in B with minimal distance $r_0 = H(r_j, x)$ to x. If $r_0 \geq r + 1$, then there is nothing to prove since $Y = 0$. Hence, we consider the case $r_0 \leq r + 1 \leq k + 2$. By the triangle inequality, the Hamming distance between x and any element $r_i \in B$ is $H(x, r_i) \geq H(r_j, r_i) - r_0$. Hence, the number of elements in B in Hamming-distance r to x, is at most the number of elements in B in Hamming-distance at most $r_0 + r$ to r_j. For all $i \in [\ell]$, define $Y_i := 1$ if $H(r_i, r_j) \leq r_0 + r$ and $Y_i := 0$ otherwise. Similarly to above, for $r \leq k + 1$, we obtain $E[Y] \leq \sum_{i=1}^{\ell} E[Y_i] \leq c'$ for some constant $c' > 0$. By Markov's inequality, it holds for $1 \leq r \leq k + 1$ that

$$\Pr\left(Y \geq \frac{c'}{\sigma n} \binom{n}{r}\right) \leq \frac{c'}{\frac{c'}{\sigma n}\binom{n}{r}} \leq \frac{\sigma n}{\left(\frac{n}{r}\right)^r} \leq \sigma.$$

Finally, consider the case $r = n/2 + r_0$ for $r_0 \geq 1$. By a symmetry argument, the number of elements in B in Hamming distance r to x, equals the number of

elements in B in Hamming distance $n-n/2-r_0 = n/2-r_0$ to the complementary bitstring to x. Thus, by the arguments above, it holds with probability $1-\sigma$ that

$$Y \leq \frac{c'}{\sigma n}\binom{n}{n/2-r_0} = \frac{c'}{\sigma n}\binom{n}{r}.$$

□

Proof of Theorem 7. We show that an instance of BBFunnel sampled according to distribution I_p satisfies the criteria of Definition 4 with probability $1-\sigma$. Define the sets[1]

$$\begin{aligned}
\tilde{B}'_i &:= \{x \in B' \mid \textsc{Om}(x) = n-i\}\\
\tilde{A}'_i &:= \{x \in A' \mid \textsc{Om}(x) = n-i\}\\
\tilde{A}_i &:= \{x \in A \mid \textsc{Lo}_y(x) = i\}\\
\tilde{B}_i &:= \{1^u p_i 0^{n-v}\}\\
\tilde{C}_i &:= \{x \in C \mid \textsc{Lo}_y(x) = i-v\}, \text{ and}\\
\tilde{D}_i &:= \{x \in D \mid \textsc{Lo}_y(x) = i-w\}.
\end{aligned}$$

Using these sets, we define the partition of $\{0,1\}^n$ as $(A_1,\ldots,A_m) := (\tilde{B}'_1, \tilde{B}'_2,..., \tilde{A}'_1, \tilde{A}'_2,...., \tilde{A}_1, \tilde{A}_2,\ldots, \tilde{B}_1, \tilde{B}_2,\ldots, \tilde{C}_1, \tilde{C}_2,\ldots, \tilde{D}_1, \tilde{D}_2,\ldots, D_{n-w})$. It is easy to see that for all j and all $x \in A_j$, there exists an element $y \in A_{j+1}$ with $H(x,y) = \mathrm{O}(1)$.

The partition has BBFunnel-deceptive pairs $(\tilde{B}_i, \tilde{C}_j)$ for all i,j. By Lemma 8, the set $B := \cup_i \tilde{B}_i$ is ε-sparse with probability $1-\delta$. Furthermore, every $\tilde{C}_{\geq j}$ is a $1-w/n$-dense set because every search point with $\textsc{Lo}_y(x) \geq j+v$ belongs to $\tilde{C}_{\geq j}$. The function therefore belongs to the class $\textsc{SparseLocalOpt}_{\alpha,\varepsilon}$ with probability $1-\sigma$. □

Proof of Lemma 3. To obtain an element in C from x via mutation, it suffices to either flip no bits, or to mutate into one of the at least αn Hamming-neighbours of x in C, each obtained by flipping exactly one specific bit. The probability of this event is

$$\left(1-\frac{\chi}{n}\right)^n + \alpha n\left(\frac{\chi}{n}\right)\left(1-\frac{\chi}{n}\right)^{n-1} > \left(1-\frac{\chi}{n}\right)^n (1+\alpha\chi).$$

□

Proof of Lemma 4. Define $B_r(x) := \{y \in B \mid H(x,y) = r\}$. If $x \in B$ and $y \sim p_{\mathrm{mut}}(x)$, then by the Binomial theorem

$$\Pr(y \in B) = \left(1-\frac{\chi}{n}\right)^n + \sum_{r=1}^{n}\sum_{z \in B_r(x)} \Pr(y=z)$$

[1]The tilde-notation is introduced to disambiguate the levels in region A of BBFunnel, and the partition $(A_1,\ldots,A_m)$ required by the definition of $\textsc{SparseLocalOpt}_{\alpha,\varepsilon}$.

$$\leq (1-\varepsilon)\left(1-\frac{\chi}{n}\right)^n + \varepsilon \sum_{r=0}^{n} \binom{n}{r} \left(\frac{\chi}{n}\right)^r \left(1-\frac{\chi}{n}\right)^{n-r}$$
$$= (1-\varepsilon)\left(1-\frac{\chi}{n}\right)^n + \varepsilon = \rho.$$

Similarly, if $x \notin B$ and $y \sim p_{\mathrm{mut}}(x)$, then

$$\Pr(y \in B) = \mathrm{O}\left(\frac{1}{n} \sum_{r=0}^{n} \binom{n}{r} \left(\frac{\chi}{n}\right)^r \left(1-\frac{\chi}{n}\right)^{n-r}\right) = \mathrm{O}\left(\frac{1}{n}\right). \square$$

Proof of Corollary 12. Note that 3-tournament selection has β-function

$$\beta(\psi, \psi+\gamma) = (1-\psi)^3 \left(1 - \left(1 - \frac{\gamma}{1-\psi}\right)^3\right).$$

We evaluate the conditions of Theorem 11 numerically with the parameters

$$\psi_0 := \frac{2001 - \sqrt{3995997}}{1334} \approx 1.5 \cdot 10^{-3}$$
$$\gamma_0 := \frac{6003 - \sqrt{36028006}}{4002} \approx 1.7 \cdot 10^{-4}, \text{ and}$$
$$\delta := \frac{1}{3000}.$$

We use the upper bound

$$\left(1-\frac{\chi}{n}\right)^n < e^{-\chi} < \frac{3335}{10000}$$

and for $n \geq 10^4$ the lower bound

$$\left(1-\frac{\chi}{n}\right)^n = \left(1-\frac{\chi}{n}\right)^{\left(\frac{n}{\chi}-1\right)\chi} \left(1-\frac{\chi}{n}\right)^{\chi}$$
$$\geq e^{-\chi}\left(1-\frac{\chi}{10^4}\right)^{\chi} > \frac{3334}{10000}.$$

Note that the function

$$h(\gamma, \psi) := \frac{\beta(\psi, \psi+\gamma)}{\gamma} = \gamma^2 + 3(1-\psi)(1-\gamma-\psi)$$

satisfies $\frac{\partial h(\gamma,\psi)}{\partial \gamma} < 0$ for all $\gamma \leq \gamma_0 < 1 < \frac{3}{2}(1-\psi_0) \leq \frac{3}{2}(1-\psi)$, and $\frac{\partial h(\gamma,\psi)}{\partial \psi} < 0$ for all $\psi \leq \psi_0 < 1/2 < \frac{2-\gamma_0}{2} \leq \frac{2-\gamma}{2}$.

Hence, we have for all $\gamma \in [\psi_0, 1]$

$$\frac{\beta(0,\gamma)}{\gamma} = h(\gamma, 0) \leq h(\psi_0, 0) < \frac{2996}{1000} < \frac{1}{\frac{\varepsilon}{1-\varepsilon} + \left(1-\frac{\chi}{n}\right)^n},$$

which implies that condition (SM0) is satisfied.

Similarly, it holds for all $\gamma \in (0, \gamma_0]$

$$\frac{\beta(0,\gamma)}{\gamma} = h(\gamma, 0) \geq h(\gamma_0, 0) > \frac{29995003}{10000000} > \frac{1+\delta}{\left(1-\frac{\chi}{n}\right)^n},$$

thus condition (SM2a) is satisfied.

Furthermore, for all $\gamma \in (0, \gamma_0]$ and $\psi \in [0, \psi_0]$, it holds

$$\frac{\beta(\psi, \psi+\gamma)}{\gamma} = h(\gamma, \psi) \geq h(\gamma_0, \psi_0) > \frac{299}{100} > \frac{1+\delta}{(1-\chi/n)^n (1+\alpha\chi)},$$

thus condition (SM2b) is also satisfied. Finally, condition (SM3) is satisfied for a sufficiently large constant c. □

Proof of Corollary 13. The proof is by Theorem 11, similarly to Corollary 12 for the parameters $\delta = 1/10^6$, $\psi_0 = 39606/10^6$ and $\gamma_0 = 12/10^8$. □

Lemma 9. *For $x > 0$*

$$\frac{x-1}{x} \leq \ln(x)$$

Lemma 10. *$e^x \geq 1 + x + \frac{x^2}{2}$ for $x \geq 0$*

Theorem 15 (Chernoff)**.** *If $X \sim \mathrm{Bin}(n, p)$, then for $0 \leq \varepsilon \leq 1$,*

$$\Pr\left(X \leq (1-\varepsilon)E\left[X\right]\right) \leq \exp\left(-\frac{\varepsilon^2 E\left[X\right]}{2}\right)$$

Theorem 16 (Hajek's drift theorem, Thm. 2.3 (2.8) Hajek (1988))**.** *Let $(Y_k)_{k\geq 0}$ be a sequence of random variables on a probability space $(\Omega, \mathcal{F}, P)$ adapted to an increasing sequence $(\mathcal{F}_k)_{k\geq 0}$ of sub-σ-algebras of $\mathcal{F}$ where for $a < b$ it holds*

D1) $E_{1,k} := E\left[e^{\eta(Y_{k+1}-Y_k)}; Y_k > a \mid \mathcal{F}_k\right] \leq \rho$, *and*

D2) $E_{2,k} := E\left[e^{\eta(Y_{k+1}-a)}; Y_k \leq a \mid \mathcal{F}_k\right] \leq D$,

then

$$\Pr\left(Y_k \geq b \mid \mathcal{F}_0\right) \leq \rho^k e^{\eta(Y_0-b)} + \frac{1-\rho^k}{1-\rho} D e^{\eta(a-b)}.$$

Theorem 17 (Additive drift theorem)**.** *Let $(Z_t)_{t\in\mathbb{N}}$ be a discrete-time stochastic process in $[0, \infty)$ adapted to any filtration $(\mathcal{F}_t)_{t\in\mathbb{N}}$. Define $T_a := \min\{t \in \mathbb{N} \mid Z_t \leq a\}$ for any $a \geq 0$. For some $\varepsilon > 0$ and constant $0 < b < \infty$, define the conditions*

1.1) $E\left[Z_{t+1} - Z_t + \varepsilon\,; t < T_a \mid \mathcal{F}_t\right] \leq 0$ *for all $t \in \mathbb{N}$,*

1.2) $E\left[Z_{t+1} - Z_t + \varepsilon\,; t < T_a \mid \mathcal{F}_t\right] \geq 0$ *for all $t \in \mathbb{N}$,*

2) $Z_t < b$ *for all* $t \in \mathbb{N}$*, and*

3) $E[T_a] < \infty$*.*

If 1.1), 2), and 3) hold, then $E[T_a \mid \mathcal{F}_0] \leq Z_0/\varepsilon$*.*
If 1.2), 2), and 3) hold, then $E[T_a \mid \mathcal{F}_0] \geq (Z_0 - a)/\varepsilon$*.*

Lemma 11 (Corus et al. (2018))**.** *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*

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