Runtime analyses of the population-based univariate estimation of distribution algorithms on LeadingOnes

Lehre, Per Kristian; Nguyen, Hai

DOI:
https://doi.org/10.1007/s00453-021-00862-3

Document Version
Publisher's PDF, also known as Version of record

Citation for published version (Harvard):
Lehre, PK & Nguyen, H 2021, 'Runtime analyses of the population-based univariate estimation of distribution algorithms on LeadingOnes', *Algorithmica*, vol. 83, no. 10, pp. 3238-3280. https://doi.org/10.1007/s00453-021-00862-3

Algorithmica
https://doi.org/10.1007/s00453-021-00862-3

Runtime Analyses of the Population-Based Univariate Estimation of Distribution Algorithms on LeadingOnes

Per Kristian Lehre[1] · **Phan Trung Hai Nguyen**[1,2]

Received: 29 May 2020 / Accepted: 18 July 2021

Abstract
We perform rigorous runtime analyses for the univariate marginal distribution algorithm (UMDA) and the population-based incremental learning (PBIL) Algorithm on LeadingOnes. For the UMDA, the currently known expected runtime on the function is $\mathcal{O}\left(n\lambda \log \lambda + n^2\right)$ under an offspring population size $\lambda = \Omega(\log n)$ and a parent population size $\mu \leq \lambda/(e(1+\delta))$ for any constant $\delta > 0$ (Dang and Lehre, GECCO 2015). There is no lower bound on the expected runtime under the same parameter settings. It also remains unknown whether the algorithm can still optimise the LeadingOnes function within a polynomial runtime when $\mu \geq \lambda/(e(1+\delta))$. In case of the PBIL, an expected runtime of $\mathcal{O}(n^{2+c})$ holds for some constant $c \in (0,1)$ (Wu, Kolonko and Möhring, IEEE TEVC 2017). Despite being a generalisation of the UMDA, this upper bound is significantly asymptotically looser than the upper bound of $\mathcal{O}\left(n^2\right)$ of the UMDA for $\lambda = \Omega(\log n) \cap \mathcal{O}(n/\log n)$. Furthermore, the required population size is very large, i.e., $\lambda = \Omega(n^{1+c})$. Our contributions are then threefold: (1) we show that the UMDA with $\mu = \Omega(\log n)$ and $\lambda \leq \mu e^{1-\varepsilon}/(1+\delta)$ for any constants $\varepsilon \in (0,1)$ and $0 < \delta \leq e^{1-\varepsilon} - 1$ requires an expected runtime of $e^{\Omega(\mu)}$ on LeadingOnes, (2) an upper bound of $\mathcal{O}\left(n\lambda \log \lambda + n^2\right)$ is shown for the PBIL, which improves the current bound $\mathcal{O}\left(n^{2+c}\right)$ by a significant factor of $\Theta(n^c)$, and (3) we for the first time consider the two algorithms on the LeadingOnes function in a noisy environment and obtain an expected runtime of $\mathcal{O}\left(n^2\right)$ for appropriate parameter settings. Our results emphasise that despite the independence assumption in the probabilistic models, the UMDA and the PBIL with fine-tuned parameter choices can still cope very well with variable interactions.

Keywords Estimation of distribution algorithms · Running time analysis · Level-based analysis · Noisy optimisation · Theory of randomised search heuristics

Preliminary versions of this work appeared in the Proceedings of the 2018 Springer Parallel Problem Solving from Nature Conference (PPSN '18) and 2019 ACM Genetic and Evolutionary Computation Conference (GECCO '19)

Extended author information available on the last page of the article

Published online: 28 August 2021

1 Introduction

1.1 Motivations

Estimation of distribution algorithms (EDAs) [24, 38, 40] are randomised search heuristics that look for optimal solutions by building and sampling from probabilistic models. They are known by various other names, including probabilistic model-building genetic algorithms [40] or iterated density estimation algorithms [2]. Unlike traditional evolutionary algorithms (EAs), which use standard genetic operators such as mutation and crossover to create variation, EDAs, on the other hand, generate it via model building and model sampling. The workflow of EDAs is an iterative process. The starting model is a uniform distribution over the search space, from which an initial population of λ individuals is sampled. The algorithm ranks individuals according to a fitness function and selects the $\mu \leq \lambda$ fittest individuals to update the model. The procedure is repeated many times and terminates when a threshold on the number of iterations is exceeded or a solution of good quality is obtained [13, 20]. We call the parameter λ the offspring population size, while the parameter μ is known as the parent population size of the algorithm.

Several EDAs have been proposed over the last decades. They differ in how they learn the variable interplay and build/update the probabilistic models over iterations. In general, EDAs can be categorised into two main classes: univariate and multivariate. Univariate EDAs assume variable independence and usually represent the model as a probability vector (each component is a marginal), encoding a product distribution from which individuals are sampled independently and identically. Typical EDAs in this class are the univariate marginal distribution algorithm (UMDA [38]), the compact genetic algorithm (cGA [19]) and the population-based incremental learning (PBIL [1]). Some ant colony optimisation algorithms like the λ-MMAS [42] can also be cast into this framework (also called n-Bernoulli-λ-EDA [15]). In contrast, multivariate EDAs apply statistics of order two or more to capture the underlying structures of the addressed problems. This paper focuses on univariate EDAs on discrete optimisation, and for that reason we refer the interested readers to [20, 23] for other EDAs on a continuous domain.

The UMDA is probably the most famous univariate EDA. In each so-called iteration, the algorithm updates marginals to the 'empirical' frequencies of 1s sampled among the μ fittest individuals. In 2015, Dang and Lehre [5], via the level-based theorem [3], obtained an upper bound of $\mathcal{O}\big(n\lambda \log \lambda + n^2\big)$ on the expected runtime for the algorithm on the LeadingOnes function when the offspring population size is $\lambda = \Omega(\log n)$ and the parent population size $\mu \leq \lambda/(e(1+\delta))$ for any constant $\delta > 0$. For $\lambda = \Omega(\log n) \cap \mathcal{O}(n/\log n)$, the above bound becomes $\mathcal{O}\big(n^2\big)$. Under a selective pressure $\mu/\lambda \geq (1+\delta)/e$, it is still unknown whether the UMDA could optimise the function in polynomial expected runtime. Furthermore, we are also missing a lower bound on the expected runtime, that is necessary to understand how the algorithm copes with variable dependencies.

Another univariate EDA is the PBIL [1]—a generalisation of the UMDA—which updates the marginals using a convex combination with a smoothing parameter

Table 1 Summary of running times of the UMDA and PBIL on LeadingOnes

Algorithm	Constraints	Expected runtime	References
UMDA	$\mu/\lambda \leq (1-\delta)/e, \lambda = \Omega(\log n)$	$\mathcal{O}(n\lambda \log \lambda + n^2)$	Dang et al. [5]
	$\delta \in (0,1)$		
	$\mu/\lambda \leq (1-\delta)/e, \mu = \Omega(\log n)$	$\Omega(n\lambda/\log \lambda)$	[Theorem 18]
	$\delta \in (0,1)$		
	$\mu/\lambda \geq (1+\delta)/e^{1-\epsilon}, \mu = \Omega(\log n)$	$e^{\Omega(\mu)}$	[Theorem 15]
	$\epsilon \in (0,1), 0 < \delta \leq e^{1-\epsilon} - 1$		
	Prior noise, $\mathcal{O}(1) \ni p \in (0,1)$	$\mathcal{O}\left(n\lambda \log \lambda + n^2\right)$	[Theorem 26]
	$\mu/\lambda \leq 1/(e(1+\delta)), \mu = \Omega(\log n), \delta > 0$		
PBIL	$\lambda = \Omega(n^{1+c}), \mu = \mathcal{O}\left(n^{c/2}\right)$	$\mathcal{O}\left(n^{2+c}\right)$	Wu et al. [47]
	$\mathcal{O}(1) \ni \rho \in (0,1], c \in (0,1)$		
	$\lambda = \Omega(\log n), \mathcal{O}(1) \ni \rho \in (1/e, 1]$	$\mathcal{O}\left(n\lambda \log \lambda + n^2\right)$	[Theorem 21]
	$\mathcal{O}(1) \ni \mu/\lambda \leq c(\rho) < 1$		
	Prior noise, $\mathcal{O}(1) \ni p \in (0,1)$	$\mathcal{O}\left(n\lambda \log \lambda + n^2\right)$	[Theorem 27]
	$\mathcal{O}(1) \ni \mu/\lambda = c(\rho, p), \mu = \Omega(\log n)$		
λ-MMAS	$\lambda \geq c \log n, c > 0$	$\mathcal{O}\left(n\lambda \log \lambda + n^2\right)$	[Corollary 24]

$\rho \in (0,1]$ between the current marginals and the empirical frequencies of 1s sampled among the μ fittest individuals (also called incremental learning). Unlike the UMDA, runtime results for the PBIL are very limited. The only rigorous analysis on test functions has been published recently in [47], where the authors argued that the algorithm with a sufficiently large population size can avoid making wrong decisions early even when the smoothing parameter is large. They also showed an expected runtime of $\mathcal{O}(n\lambda) = \mathcal{O}\left(n^{2+\epsilon}\right)$ on the LeadingOnes function for some small constant $\epsilon \in (0,1)$. And yet the required offspring population size still remains large, i.e., $\lambda = \Omega(n^{1+\epsilon})$ [47]. It remains open whether a tighter upper bound can be obtained for the PBIL on the LeadingOnes function. The answer to this question is of special interest because it might be considered as the first step towards showing the substantial advantage of incremental learning over the update mechanism used by the UMDA. Furthermore, more bounds on the expected runtime of the PBIL on test functions with well-known structures possibly shed light on the behaviour of the algorithm on other problems, especially those with a separably additive decomposition property [37] because many sub-functions may have fitness landscapes that resemble those of test functions, and in these situations the behaviours of the algorithms can be quickly deduced.

See Table 1 for a summary of the latest runtime results of the UMDA and the PBIL on the LeadingOnes function.

1.2 Our Contributions

The contributions of this paper are three-fold.

1. We analyse the expected runtime of the UMDA. Together with previous results [5, 6], our results provide a clearer picture of the runtime of the algorithm on the LeadingOnes function. We show that under a low selective pressure the algorithm fails to optimise the function in polynomial expected runtime. This result essentially reveals the limitations of probabilistic models based on probability vectors as the algorithm hardly stays in promising states when the selective pressure is not high enough, while the optimum cannot be sampled with high probability. On the other hand, when the selective pressure is sufficiently high, we obtain a lower bound of $\Omega(n\lambda/\log\lambda)$ on the expected runtime under any offspring population sizes $\lambda = \Omega(\log n)$.
2. We obtain an expected runtime of $\mathcal{O}\big(n\lambda\log\lambda + n^2\big)$ for the PBIL on the LeadingOnes function under any population sizes $\lambda = \Omega(\log n)$. For $\lambda = \mathcal{O}(n/\log n)$, the runtime bound becomes $\mathcal{O}\big(n^2\big)$, making it relatively comparable to the performance of the class of univariate unbiased black-box algorithms in the sense of Lehre and Witt [31]—a general framework covering many well-known randomised search heuristics in evolutionary computation. More importantly, the new upper bound improves the previously best known upper bound of $\mathcal{O}\big(n^{2+c}\big)$ [47] by a factor of $\Theta(n^c)$ for some constant $c \in (0, 1)$. Our bound only requires a population of size $\lambda = \Omega(\log n)$ as opposed to $\lambda = \Omega(n^{1+\varepsilon})$ as in [47]. To do this, we make use of the level-based theorem [3] with some additional arguments. By taking advantage of the Dvoretzky–Kiefer–Wolfowitz (DKW) inequality [33], we observe that with high probability, the empirical frequency does not deviate far from the model probability. We believe that it is the first time that the DKW inequality has been used in the runtime analysis of model-based algorithms.
3. We introduce noise to the LeadingOnes function, where a single bit is flipped before evaluating the fitness with a constant probability $p \in (0, 1)$ (also called prior noise). We note that the same noise model is first considered in [11, 41] for the $(1 + 1)$ EA and [4, 17] for population-based EAs. We show that if the selective pressure μ/λ is sufficiently high, the algorithms optimise the noisy LeadingOnes function within an expected runtime of $\mathcal{O}\big(n^2 + n\lambda\log\lambda\big)$. To the best of our knowledge, this is also the first time that the UMDA and the PBIL are rigorously studied in a noisy environment, while the cGA is already considered in [16] under Gaussian posterior noise. Despite the simplicity of the noise model considered, this can be viewed as the first step towards understanding the behaviour of these EDAs in a noisy environment.

1.3 Outline of the Paper

The paper is structured as follows. Section 2 introduces the studied algorithms and general tools used in the paper. Section 3 provides a detailed analysis for an exponential expected runtime for the UMDA on the LeadingOnes function in case of low selective pressure, followed by the analysis under a high selective pressure in Sect. 4. We also present in this section an improved upper bound on the expected runtime of the PBIL on LeadingOnes. In Sect. 5, we consider the LeadingOnes function under a prior noise model and obtain an upper bound $\mathcal{O}\big(n^2\big)$ on the expected runtime for

appropriate parameter settings. Section 6 presents an empirical study to complement theoretical results derived earlier. The paper ends in Sect. 7, where we give our concluding remarks and speak of potential future work.

2 Preliminaries

We first recall that a random variable Y is said to follow a Bernoulli distribution with success probability $p \in [0,1]$, denoted as $Y \sim \text{Ber}(p)$, if and only if $\Pr(Y=1)=p$ and $\Pr(Y=0)=1-p$ [36, p. 445]. If there are $n \in \mathbb{N}$ such random variables (with the same success probability p), then the sum of them (i.e., a random variable X) follows a binomial distribution with n trials and success probability p, denoted as $X \sim \text{Bin}(n,p)$ [36, p. 445]. An extension of the binomial distribution is the Poisson binomial distribution, in which each of n random variables can have a different success probability [14, p. 263]. More formally, we write $X \sim \text{PB}(p_1, p_2, \ldots, p_n)$ if and only if $X = \sum_{i=1}^{n} X_i$, where $\Pr\left(X_i = 1\right) = p_i$ and $\Pr\left(X_i = 0\right) = 1 - p_i$ for all $i \in [n]$.

2.1 The Studied Fitness Function

In evolutionary computing, we represent a solution to an optimisation problem as a bitstring (or an individual) $x = (x_1, x_2, \ldots, x_n)$ of length $n \in \mathbb{N}$, where $x_i \in \{0,1\}$ for all $i \in \{1,2,\ldots,n\} =: [n]$. We consider in the paper the problem of maximising the number of leading 1s in a bitstring. The fitness value of such a bitstring can be obtained by

$$\text{LeadingOnes}(x) := \sum_{i=1}^{n} \prod_{j=1}^{i} x_j. \tag{1}$$

This is a uni-modal function with a maximum fitness value of n when the input is the all-ones bitstring (i.e., the global optimum). In essence, the bits in this particular function are highly correlated, so it is often used to study the ability of EDAs to cope with variable dependency [22]. We call n the problem instance size and $\mathcal{X} = \{0,1\}^n$ the finite binary search space consisting of all bitstrings of length n.

2.2 Population-Based Univariate EDAs

The UMDA, defined in Algorithm 1, maintains a probabilistic model that is represented as an n-vector $p_t := (p_{t,1}, \ldots, p_{t,n})$, where each so-called marginal $p_{t,i} \in [0,1]$ for $i \in [n]$ is the probability of sampling a one at the i-th bit position in the offspring. The probability of sampling a particular individual $x = (x_1, x_2, \ldots, x_n) \in \mathcal{X}$ from the given probability vector p_t equals

$$\Pr\left(x = (x_1, x_2, \ldots, x_n) \mid p_t\right) = \prod_{i=1}^{n} \left[\left(p_{t,i}\right)^{x_i} \left(1 - p_{t,i}\right)^{1-x_i} \right]. \tag{2}$$

We often call the distribution defined in Eq. 2 a product distribution. The starting model is the uniform distribution $p_0 := (1/2, \dots, 1/2)$. The algorithm, in each iteration $t \in \mathbb{N}$, samples an offspring population of λ individuals, denoted as $P_t = (x_t^{(1)}, x_t^{(2)}, \dots, x_t^{(\lambda)})$, and sorts them in descending order according to fitness to obtain a sorted population $\tilde{P}_t = (\tilde{x}_t^{(1)}, \tilde{x}_t^{(2)}, \dots, \tilde{x}_t^{(\lambda)})$. A parent population consisting of the μ fittest individuals in $\tilde{P}_t$ participates in the update of the probabilistic model.

Let $x_{t,i}^{(j)}$ denote the value sampled at bit position i in the j-th individual in the offspring population P_t (and analogously $\tilde{x}_{t,i}^{(j)}$ for the parent population $\tilde{P}_t$). Then, $X_{t,i} := \sum_{j=1}^{\mu} \tilde{x}_t^{(j)}$ is the number of 1s sampled at bit position $i \in [n]$ across the parent population. The algorithm first sets each marginal to the value $X_{t,i}/\mu$ and then adjusts them to be within the interval $[1/n, 1 - 1/n]$, where the two values, $1/n$ and $1 - 1/n$, are called the lower and upper borders (or margins), respectively. In summary, the updating process can be written as follows.

$$p_{t+1,i} \leftarrow \max\left\{\frac{1}{n}, \min\left\{1 - \frac{1}{n}, \frac{X_{t,i}}{\mu}\right\}\right\} \quad \text{for all } i \in [n], \tag{3}$$

Furthermore, the ratio of $\mu/\lambda \in (0, 1]$ is known as the selective pressure of the algorithm. The whole procedure is repeated until some terminal condition has been fulfilled. Some common choices are a threshold on the number of iterations allowed to run or a lower bound on the fitness quality of the fittest individual in the population. However, for theoretical analysis, we halt the algorithm only after a global optimum has been found for the first time.

Algorithm 1: UMDA with an offspring population size λ and a parent population size μ for the maximisation of a function $f(x)$ where x is of length n.

$t \leftarrow 0$; initialise $p_t \leftarrow (1/2, 1/2, \dots, 1/2)$
repeat
  for $j = 1, 2, \dots, \lambda$ **do**
    $x_{t,i}^{(j)} \leftarrow \mathrm{Ber}(p_{t,i})$ for each $i \in [n]$
  $P_t \leftarrow (x_t^{(1)}, x_t^{(2)}, \dots, x_t^{(\lambda)})$
  $\tilde{P}_t \leftarrow (\tilde{x}_t^{(1)}, \tilde{x}_t^{(2)}, \dots, \tilde{x}_t^{(\lambda)})$ by sorting P_t in descending order of fitness, where ties are broken uniformly at random
  for $i = 1, 2, \dots, n$ **do**
    $X_{t,i} \leftarrow \sum_{j=1}^{\mu} \tilde{x}_{t,i}^{(j)}$
    $p_{t+1,i} \leftarrow \max\{1/n, \min\{1 - 1/n, X_{t,i}/\mu\}\}$
  $t \leftarrow t + 1$
until *termination condition is fulfilled*

A generalisation of the UMDA is the PBIL. While most operations in the PBIL are similar to those of the UMDA, the algorithm makes use of a new smoothing parameter $\rho \in (0, 1]$ and updates the model via a convex combination as follows.

$$p_{t+1,i} \leftarrow \max\left\{\frac{1}{n}, \min\left\{1-\frac{1}{n}, (1-\rho)p_{t,i} + \rho \cdot \frac{X_{t,i}}{\mu}\right\}\right\} \quad \text{for all } i \in [n], \tag{4}$$

In essence, the PBIL takes into account the current marginals when updating the probabilistic model. We also note that the UMDA is the PBIL with a maximum smoothing parameter $\rho = 1$.

2.3 Level-Based Analysis

First proposed in [25], the level-based theorem is a general tool that provides upper bounds on the expected runtime of many non-elitist population-based algorithms on a wide range of optimisation problems [3, 5, 6, 8, 26, 27, 29]. The theorem assumes that the studied algorithm can be cast into the framework in Algorithm 2, which maintains a population $P_t \in \mathcal{X}^\lambda$, where $\mathcal{X}^\lambda$ is the space of all populations of size λ. We write $P_{t,i}$ to denote the i-th individual in the population P_t. The theorem also assumes the existence of a mapping $\mathcal{D}$ from the space of populations $\mathcal{X}^\lambda$ to the space of the probability distribution over the search space. In iteration t, the mapping $\mathcal{D}$ depends only on the population P_t and involves in the production of a new population for the next iteration [3].

Algorithm 2: Non-elitist population-based algorithm

```
t ← 0; create initial population P_t
repeat
    for i = 1, ..., λ do
        sample P_{t+1,i} ~ D(P_t)
    t ← t + 1
until termination condition is fulfilled
```

However, the theorem never assumes specific fitness functions, selection mechanisms, or generic operators like mutation and crossover, but it assumes that the search space $\mathcal{X}$ can be partitioned into m disjoint subsets $A_1, \ldots, A_m$, which we call levels, and the last level A_m consists of all global optima of the objective function. Let $A_{\geq j} := \cup_{k=j}^{m} A_k$. The following theorem is taken from [3, Theorem 1].

Theorem 1 (Level-Based Theorem) *Given a partition* $(A_i)_{i\in[m]}$ *of* $\mathcal{X}$, *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 of Algorithm 2 in iteration* t. *Let* $y \sim \mathcal{D}(P_t)$. *If there exist* $z_1, \ldots, z_{m-1}, \delta \in (0,1]$, *and* $\gamma_0 \in (0,1)$ *such that for any population* $P_t \in \mathcal{X}^\lambda$,

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

$$\Pr\left(y \in A_{\geq j+1}\right) \geq z_j.$$

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

$$\Pr\left(y \in A_{\geq j+1}\right) \geq (1+\delta)\gamma.$$

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

$$\lambda \geq \left(\frac{4}{\gamma_0 \delta^2}\right) \ln\left(\frac{128m}{z_* \delta^2}\right),$$

where $z_* := \min_{j \in [m-1]}\{z_j\}$, *then*

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

2.4 Dvoretzky–Kiefer–Wolfowitz Inequality

The DKW inequality [33] provides an estimate on how close an empirical distribution will be to the true distribution from which the samples are drawn. Let $\mathbb{1}_A(x)$ be the indicator function, where $\mathbb{1}_A(x) = 1$ if $x \in A$ and 0 otherwise. The following theorem is derived by replacing $\varepsilon = \varepsilon' \sqrt{\lambda}$ into [33, Corollary 1].

Theorem 2 (DKW Inequality) *Let* $X_1, \ldots, X_\lambda$ *be* λ *i.i.d. real-valued random variables with cumulative distribution function* F. *Let* $\widetilde{F}_\lambda$ *be the empirical distribution function which is defined by*

$$\widetilde{F}_\lambda(x) := \frac{1}{\lambda} \sum_{i=1}^{\lambda} \mathbb{1}_{\{X_i \leq x\}}.$$

For any $\lambda \in \mathbb{N}$ *and* $\varepsilon > 0$, *we always have*

$$\Pr\left(\sup_{x \in \mathbb{R}} \left|\widetilde{F}_\lambda(x) - F(x)\right| > \varepsilon\right) \leq 2e^{-2\lambda\varepsilon^2}.$$

We note that the advantage of the DKW inequality comes from the fact that the upper bound $\exp\{-2\lambda\varepsilon^2\}$ depends only on the number of samples λ, which in our case is the offspring population size of the algorithms.

2.5 Majorisation

We also exploit the properties of majorisation, defined in Definition 3 [18, p. 183], between two vectors.

Definition 3 (*Majorisation*) Given two vectors $p := (p_1, \ldots, p_n)$ and $q := (q_1, \ldots, q_n)$, where $p_1 \geq p_2 \geq \ldots \geq p_n$ and analogously for the q_i. Vector p is said to majorise vector q

if

$$\sum_{i=1}^{k} p_i \geq \sum_{i=1}^{k} q_i \quad \text{for all } k \in [n-1],$$

and

$$\sum_{i=1}^{n} p_i = \sum_{i=1}^{n} q_i.$$

Majorisation is a powerful tool in runtime analysis of univariate EDAs because the algorithms operate on a probability vector-based model (see [6, 10, 27, 45, 47]). The following lemma shows one of the properties of majorisation, which will be used frequently in the main parts of the paper.

Lemma 4 *Let* $X \sim \mathrm{PB}(p_1, \ldots, p_n)$ *and* $Y \sim \mathrm{PB}(q_1, \ldots, q_n)$. *If the vector* $(p_1, \ldots, p_n)$ *majorises the vector* $(q_1, \ldots, q_n)$, *then*

$$\Pr(X = n) \leq \Pr(Y = n).$$

Proof We obtain from [32, Proposition F.1.a.] that if the vector $(p_1, \ldots, p_n)$ majorises the vector $(q_1, \ldots, q_n)$, then

$$\prod_{i=1}^{n} p_i \leq \prod_{i=1}^{n} q_i.$$

The proof is complete by noting that $\Pr(X = n) = \prod_{i=1}^{n} p_i$ and $\Pr(Y = n) = \prod_{i=1}^{n} q_i$. □

2.6 Other Tools

Lemma 5 (Chernoff Bound [36]) *Let* $X \sim \mathrm{PB}(p_1, p_2, \ldots, p_n)$. *Let* $\mu = \mathbb{E}[X] = \sum_{i=1}^{n} p_i$. *Then*

(a) $\Pr(X \geq (1+\delta)\mu) \leq e^{-\delta^2\mu/3}$ *for any* $\delta > 0$.
(b) $\Pr(X \leq (1-\delta)\mu) \leq e^{-\delta^2\mu/2}$ *for any* $0 \leq \delta \leq 1$.

Lemma 6 (*Chernoff–Hoeffding Bound* [12]) *Let* $X \sim \mathrm{PB}(p_1, p_2, \ldots, p_n)$. *Let* $\mu = \mathbb{E}[X] = \sum_{i=1}^{n} p_i$. *Then*, $\Pr(|X - \mu| \geq t) \leq 2e^{-2t^2/n}$.

We also recall that a random variable X is said to stochastically dominate another random variable Y (defined on the same probability space) if for all $k \in \mathbb{R}$ we have $\Pr(X \geq k) \geq \Pr(Y \geq k)$ [7, Definition 1.8.1], and as a result $\mathbb{E}[X] \geq \mathbb{E}[Y]$ [7, Corollary 1.8.3]. We write $\ln(\cdot)$ to denote the natural logarithm (with base e) and $\log(\cdot)$ for logarithm with base two.

3 Runtime Analysis Under Low Selective Pressure

Before we get to analysing the function, we introduce some notation. Let $C_{t,i}$ for all $i \in [n]$ denote the number of individuals having at least i leading 1s in iteration t, and $D_{t,i}$ is the number of individuals having exactly $i-1$ leading 1s. For the special case $i = 1$, $D_{t,i}$ consists of individuals that do not have any leading 1s.

Once the population has been sampled, the algorithm invokes truncation selection to select the μ fittest individuals (out of a population of λ) to update the probability vector. We take this μ-cutoff into account by defining a random variable

$$Z_t := \max\{i \in \{0\} \cup [n] \,:\, C_{t,i} \geq \mu\}, \tag{5}$$

which tells us how many marginals, counting from bit position one, are set to the upper border $1 - 1/n$ in iteration t. Furthermore, we define another random variable

$$Z_t^* := \max\{i \in \{0\} \cup [n] \,:\, C_{t,i} > 0\} \tag{6}$$

to be the number of leading 1s of the fittest individual(s).

3.1 On the Distributions of $C_{t,i}$ and $D_{t,i}$

In order to analyse the distributions of the random variables $C_{t,i}$ and $D_{t,i}$, we shall take an alternative view on the sampling process at an arbitrary bit position $i \in [n]$ in iteration $t \in \mathbb{N}$ via the principle of deferred decisions [36, p. 55]. We imagine that the process samples the values of the first bit for λ individuals. Once this has finished, it moves on to the second bit and so on until the population is sampled.

To be more specific, we now look at the first bit in iteration t. The number of 1s sampled at the first bit position follows a binomial distribution with parameters λ and $p_{t,1}$, i.e., $C_{t,1} \sim \mathrm{Bin}\big(\lambda, p_{t,1}\big)$. Thus, the number of 0s at the first bit position is $D_{t,1} = \lambda - C_{t,1}$. For completeness, we always assume that $C_{t,0} = \lambda$.

Having sampled the first bit for λ individuals, note that the bias due to selection at the second bit position comes into play only if the first bit is a 1. If this is the case, then a 1 is more preferred to a 0 at the second-bit position. Among the $C_{t,1}$ fittest individuals, the probability of sampling a 1 at the second bit position is $p_{t,2}$; thus, the number of individuals having at least 2 leading 1s is binomially distributed with parameters $C_{t,1}$ and $p_{t,2}$, that is, $C_{t,2} \sim \mathrm{Bin}\big(C_{t,1}, p_{t,2}\big)$, and the number of 0s equals $D_{t,2} = C_{t,1} - C_{t,2}$. Among the $D_{t,1}$ last individuals, since for these individuals the first bit is a 0, there is no bias between a 1 and a 0. The number of 1s sampled at the second bit position among the $D_{t,1}$ last individuals follows a binomial distribution with parameters $D_{t,1}$ (or $\lambda - C_{t,1}$) and $p_{t,2}$.

We can generalise this result for an arbitrary bit position $i \in [n]$. The number of individuals having at least i leading 1s follows a binomial distribution with $C_{t,i-1}$ trials and success probability $p_{t,i}$, i.e.,

$$C_{t,i} \sim \mathrm{Bin}\big(C_{t,i-1}, p_{t,i}\big), \tag{7}$$

and

$$D_{t,i} = C_{t,i-1} - C_{t,i} \sim \mathrm{Bin}\big(C_{t,i-1}, 1 - p_{t,i}\big). \tag{8}$$

Furthermore, the number of 1s sampled among the $\lambda - C_{t,i-1}$ remaining individuals is binomially distributed with $\lambda - C_{t,i-1}$ trials and success probability $p_{t,i}$. Let $(\mathcal{F}_t : t \in \mathbb{N})$ be a filtration induced from the population $(P_t : t \in \mathbb{N})$ [44, p. 93]. If we consider the expectations of these random variables, by the tower property of conditional expectation (or tower rule) [44, p. 88] and the fact that $p_{t,i}$ is $\mathcal{F}_{t-1}$-measurable [44, p. 93], we then get

$$\begin{aligned} \mathbb{E}[C_{t,i} \mid \mathcal{F}_{t-1}] &= \mathbb{E}[\mathbb{E}[C_{t,i} \mid C_{t,i-1}, \mathcal{F}_{t-1}] \mid \mathcal{F}_{t-1}] \\ &= \mathbb{E}[C_{t,i-1} \cdot p_{t,i} \mid \mathcal{F}_{t-1}] \\ &= \mathbb{E}[C_{t,i-1} \mid \mathcal{F}_{t-1}] \cdot p_{t,i}, \end{aligned} \tag{9}$$

and similarly

$$\mathbb{E}[D_{t,i} \mid \mathcal{F}_{t-1}] = \mathbb{E}[C_{t,i-1} \mid \mathcal{F}_{t-1}] \cdot (1 - p_{t,i}). \tag{10}$$

We note that by the end of this sampling process, we will obtain a population that is sorted in descending order according to the LeadingOnes-values.

Recall that we aim at showing that the UMDA takes exponential time to optimise the LeadingOnes function when the selective pressure is not sufficiently high, as required in [6, Theorem 7]. Let $\psi := \mu/\lambda$ denote the selective pressure of the algorithm. For any constant $\delta \in (0, 1)$, we define

$$\alpha = \alpha(n) := \log_{1-1/n}(\psi/(1-\delta)) \tag{11}$$

$$\beta = \beta(n) := \log_{1-1/n}(\psi/(1+\delta)). \tag{12}$$

Clearly, we always get $\alpha \leq \beta$. We also define a stopping time. $\tau := \min\{t \in \mathbb{N} : Z_t \geq \alpha\}$ to be the first hitting time of the value α for the random variable Z_t. We then consider two phases: (1) until the random variable Z_t hits the value α for the first time ($t \leq \tau$), and (2) after the random variable Z_t has hit the value α for the first time ($t > \tau$).

3.2 Phase 1: Before the Fitness of the μth Individual Hits the Threshold α

The algorithm starts with an initial population P_0 sampled from a uniform distribution $p_0 = (1/2, \ldots, 1/2)$. An initial observation is that the all-ones bitstring cannot be sampled in the population P_0 with high probability since the probability of

sampling it from the uniform distribution is 2^{-n}, then by the union bound [14, p. 23] it appears in the population P_0 w.p. at most $\lambda \cdot 2^{-n} = 2^{-\Omega(n)}$ since we only consider the offspring population of size at most polynomial in the problem instance size n (i.e., $\lambda \in \text{poly}(n)$). The following lemma states the expectations of the random variables Z_0^* and Z_0.

Lemma 7 $\mathbb{E}[Z_0] \leq \mathbb{E}[Z_0^*] = \mathcal{O}(\log \lambda)$.

Proof Because $Z_0 \leq Z_0^*$, we have $\mathbb{E}[Z_0] \leq \mathbb{E}[Z_0^*]$. We are left to bound the expectation $\mathbb{E}[Z_0^*]$. Recall that $Z_0^* = \max\{i : C_{0,i} > 0\}$ and let f:=LeadingOnes. The probability of sampling an individual with more than k leading 1s (where $k < n$) is $\Pr(f(x) > k) = (1/2)^{k+1} = 2^{-(k+1)}$, thus $\Pr(f(x) \leq k) = 1 - \Pr(f(x) > k) = 1 - 2^{-(k+1)}$. The event $Z_0^* \leq k$ implies that the λ individuals all have at most k leading 1s, i.e.,

$$\Pr(Z_0^* \leq k) = \prod_{i=1}^{\lambda} \Pr(f(x_0^{(i)}) \leq k) = (1 - 2^{-(k+1)})^{\lambda},$$

and $\Pr(Z_0^* > k) = 1 - (1 - 2^{-(k+1)})^{\lambda}$. Given that $\mathbb{E}[Y] \leq \sum_{i=0}^{\infty} \Pr(Y > i)$ for any bounded integer-valued random variable Y [7, Lemma 6.1], the random variable Z_0^* is integer-valued, we then get

$$\begin{aligned}
\mathbb{E}\left[Z_0^*\right] &< \sum_{k=0}^{\infty} \Pr(Z_0^* > k) \\
&= \sum_{k=0}^{\infty} (1 - (1 - 2^{-(k+1)})^{\lambda}) \\
&\leq \log \lambda + \sum_{k=\log \lambda}^{\infty} (1 - (1 - 2^{-(k+1)})^{\lambda}) \\
&\leq \log \lambda + \lambda \sum_{k=\log \lambda}^{\infty} 2^{-(k+1)} \qquad \text{(by Bernoulli's ineq. [8])} \\
&\leq \log \lambda + \lambda \cdot 2^{-\log \lambda + 1} \\
&= \log \lambda + 2,
\end{aligned}$$

which completes the proof. □

Lemma 8 *It holds for any* $t \in \mathbb{N}$ *and* $i \geq Z_t + 2$ *that* $X_{t,i} \sim \text{Bin}\left(\mu, p_{t,i}\right)$.

Proof By the definition of the random variable Z_t, we know that $C_{t,Z_t} \geq \mu$ and $C_{t,Z_t+1} < \mu$. Consider bit position j:=$Z_t + 2$. We then obtain from Eq. 7 that among the $C_{t,j-1}$ fittest individuals there are $C_{t,j} \sim \text{Bin}\left(C_{t,j-1}, p_{t,j}\right)$ individuals with at least j leading 1s. For the $\mu - C_{t,j-1} > 0$ remaining individuals (among the μ fittest individuals), the overall fitness (or the fitness ranking) of these individuals have been already decided by the first $j-1$ bits, and what is sampled at bit position j will not have any impact on the ranking of these individuals. In other words, there is no bias

in bit j among these (remaining) individuals, which also means that the number of 1s sampled here follows a binomial distribution with $\mu - C_{t,j-1}$ trials and success probability $p_{t,j}$, i.e., $\mathrm{Bin}\big(\mu - C_{t,j-1}, p_{t,j}\big)$. Thus, we get:

$$\begin{aligned} X_{t,j} &\sim C_{t,j} + \mathrm{Bin}\big(\mu - C_{t,j-1}, p_{t,j}\big) \\ &\sim \mathrm{Bin}\big(C_{t,j-1}, p_{t,j}\big) + \mathrm{Bin}\big(\mu - C_{t,j-1}, p_{t,j}\big) \\ &\sim \mathrm{Bin}\big(\mu, p_{t,j}\big). \end{aligned}$$

Because the distribution of $X_{t,j}$ depends only on $p_{t,j}$, the same line of arguments can be repeated for each of the remaining bit positions $Z_t + 3, \ldots, n$. The proof is now complete. □

We now show that the value of the random variable Z_t does not decrease during phase 1 with high probability.

Lemma 9 $\Pr\big(\forall t \in [1, \tau] : Z_t \geq Z_{t-1}\big) \geq 1 - \tau e^{-\Omega(\mu)}$.

Proof It suffices to show that w.p. at most $\tau e^{-\Omega(\mu)}$ there exists an iteration $t \in [1, \tau]$ such that $Z_t < Z_{t-1}$. We first note that the value of the random variable Z_t drops in iteration $t + 1$ only if the number of individuals with at least Z_t leading 1s in the next iteration is less than μ. Recall that $Z_t < \alpha$ for any $t < \tau$. The number of individuals with at least Z_t leading 1s, sampled in iteration $t + 1$, follows a binomial distribution with λ trials and success probability $(1 - 1/n)^{Z_t}$. Thus, in expectation the number of such individuals is

$$\lambda\left(1 - \frac{1}{n}\right)^{Z_t} \geq \lambda\left(1 - \frac{1}{n}\right)^{\alpha} = \frac{\mu}{1 - \delta}.$$

By a Chernoff bound (see Lemma 5), the probability of sampling at most $(1 - \delta) \cdot \mu/(1 - \delta) = \mu$ such individuals is at most $e^{-(\delta^2/2)\cdot\mu/(1-\delta)} = e^{-\Omega(\mu)}$ for any constant $\delta \in (0, 1)$. By the union bound, this rare event happens at least once during the first τ iterations w.p. at most $\tau e^{-\Omega(\mu)}$, and the complement event takes place w.p. at least $1 - \tau e^{-\Omega(\mu)}$, which completes the proof. □

3.3 Phase 2: After the Fitness of the μth Individual has Hit Value α for the First Time

By the definition of Z_t, the first Z_t marginals are set to the upper border $1 - 1/n$ in iteration $t \in \mathbb{N}$. Recall that the random variable $X_{t,i}$ denotes the number of 1s at bit position $i \in [n]$ among the μ fittest individuals, which is used to update the probabilistic model of the UMDA.

The preceding section shows that the random variable Z_t is non-decreasing during phase 1 w.p. $1 - \tau e^{-\Omega(\mu)}$. The following lemma also shows that its value stays above α afterwards with high probability.

Lemma 10 *It holds for any constant $k > 0$ that*

$$\Pr(\forall t \in [\tau, \tau + e^{k\mu}] : Z_t \geq \alpha) \geq 1 - e^{k\mu} \cdot e^{-\Omega(\mu)}.$$

Proof Consider the worst scenario in which $Z_t = \alpha$ for some $t \in [\tau, \tau + e^{k\mu}]$. We also note that the value of the random variable Z_t drops below α in iteration $t + 1$ if and only if the number of individuals with at least Z_t leading 1s sampled in the next iteration is less than μ. An offspring with at least α leading 1s is still sampled w.p. $(1 - 1/n)^\alpha = \mu/(\lambda(1 - \delta))$ for some constant $\delta \in (0, 1)$, and by a Chernoff bound, there are at most μ such individuals sampled in the next iteration w.p. at most $e^{-\Omega(\mu)}$. By the union bound, this happens at least once in the interval $[\tau, \tau + e^{k\mu}]$ w.p. at most $e^{k\mu} \cdot e^{-\Omega(\mu)}$. The complement event then occurs w.p. at least $1 - e^{k\mu} \cdot e^{-\Omega(\mu)}$, which completes the proof. □

The following lemma further shows that there is also an upper bound on the random variable Z_t.

Lemma 11 *It holds for any constant $k > 0$ that*

$$\Pr(\forall t \in [0, e^{k\mu}] : Z_t \leq \beta) \geq 1 - e^{k\mu} \cdot e^{-\Omega(\mu)}.$$

Proof It suffices to show that w.p. at most $e^{k\mu} \cdot e^{-\Omega(\mu)}$ there exists an iteration $t \in [0, e^{k\mu}]$ (for any constant $k > 0$) such that $Z_t > \beta$. Consider an arbitrary iteration $t \in [0, e^{k\mu}]$. An individual with at least β leading 1s is sampled w.p.

$$\prod_{i=1}^{\beta} p_{t,i} \leq \left(1 - \frac{1}{n}\right)^\beta = \frac{\mu}{\lambda(1 + \delta)}$$

for some constant $\delta \in (0, 1)$. Thus, the number of such individuals sampled in the next iteration will be stochastically dominated by $\mathrm{Bin}(\lambda, \mu/(\lambda(1 + \delta)))$, and thus their expected number is at most $\mu/(1 + \delta)$. By a Chernoff bound, the probability of sampling at least $(1 + \delta) \cdot \mu/(1 + \delta) = \mu$ such individuals in the next iteration is at most $e^{-\Omega(\mu)}$. By the union bound, this rare event happens at least once in the interval $[0, e^{k\mu}]$ w.p. at most $e^{k\mu} \cdot e^{-\Omega(\mu)}$. Thus, the complement event occurs w.p. at least $1 - e^{k\mu} \cdot e^{-\Omega(\mu)}$. The proof is then complete by noting that the value of the random variable Z_t exceeds β if and only if the number of individuals with more than β leading 1s sampled in the next iteration is at least μ. □

Lemmas 10 and 11 together give essential insights about the behaviour of the algorithm. The random variable Z_t will stay well below the threshold β for $e^{\Omega(\mu)}$ iterations w.p. $1 - e^{-\Omega(\mu)}$ for a sufficiently large parent population size μ. More precisely, the random variable Z_t will move back and forth around an equilibrium value

$$\kappa = \kappa(n) := \log_{1-1/n}(\psi). \tag{13}$$

This is because when $Z_t = \kappa$, in expectation there are exactly $\lambda(1 - 1/n)^\kappa = \lambda\psi = \mu$ individuals having at least κ leading 1s.

An exponential lower bound on the runtime is obtained if we can also show that the probability of sampling the $n - \beta$ last bits correctly is exponentially small. We now choose the ratio of μ/λ such that $n - \beta \geq \varepsilon n$ for any constant $\varepsilon \in (0, 1)$, that is equivalent to $\beta \leq n(1 - \varepsilon)$. By (12) and solving for ψ, we then obtain

$$\frac{\psi}{1+\delta} \geq \left(1 - \frac{1}{n}\right)^{n(1-\varepsilon)}.$$

The right-hand side is at most $1/e^{1-\varepsilon}$ as $(1 - 1/n)^n \leq 1/e$ for all $n > 0$ [36], so the above inequality always holds if the selective pressure satisfies $\psi \geq (1 + \delta)/e^{1-\varepsilon}$ for any constants $\delta > 0$ and $\varepsilon \in (0, 1)$.

The remainder of this section shows that the $n - (\beta + 1) = \Omega(n)$ last bits cannot be sampled correctly in any polynomial number of iterations with high probability. We first show that the sampling processes among the $\Omega(n)$ last bits are mutually independent. To ease the analysis, we further define $Y_{t,1}, Y_{t,2}, \ldots, Y_{t,n}$ to be n Bernoulli random variables representing an offspring sampled from the product distribution p_t (see Eq. 2).

Lemma 12 *Let k be any positive constant. It holds w.p. at least $1 - e^{k\mu} \cdot e^{-\Omega(\mu)}$ that the random variables $Y_{t,\beta+2}, Y_{t,\beta+3}, \ldots, Y_{t,n}$ are pairwise independent for all $t \leq e^{k\mu}$.*

Proof By Lemma 11, we know that $Z_t \leq \beta$ for any $t \leq e^{k\mu}$ w.p. at least $1 - e^{k\mu} \cdot e^{-\Omega(\mu)}$. We also obtain by Lemma 8 that $X_{t+1,j} \sim \mathrm{Bin}\big(\mu, p_{t+1,j}\big)$ for any $j \geq \beta + 2$. In other words, the number of ones sampled at a bit position $j \geq \beta + 2$ among the μ fittest individuals in the next iteration depends only on the marginal $p_{t,j}$. Thus, for any two distinct bit positions $j_1, j_2 \in \{\beta + 2, \ldots, n\}$ sampling a one at bit position j_1 is independent of sampling a one at bit position j_2. □

Now consider an arbitrary bit position $i \geq \beta + 1$. We always get $\mathbb{E}[Y_{t,i} \mid \mathcal{F}_{t-1}] = p_{t,i}$, and by the tower property of conditional expectation we also obtain

$$\mathbb{E}[Y_{t,i}] = \mathbb{E}[\mathbb{E}[Y_{t,i} \mid \mathcal{F}_{t-1}]] = \mathbb{E}[p_{t,i}].$$

For the UMDA without borders, the stochastic process $(p_{t,i} : t \in \mathbb{N})$ is a martingale [15], which results in $\mathbb{E}[p_{t,i}] = p_{0,i} = 1/2$. We will show in the following lemma that for the UMDA with borders the expected value of a marginal at an arbitrary bit position $i \geq \beta + 2$ also stays at 1/2 for any $t \in \mathbb{N}$.

Lemma 13 *Let $\mu \geq c \log n$ for a sufficiently large constant $c > 0$. If there exists a constant $k < n$ such that $Z_t \leq k - 2$ for any $t \in \mathbb{N}$, then for any $i \geq k$ that*

$$\mathbb{E}[p_{t,i}] = \frac{1}{2}.$$

Proof For readability, we omit the index i through out the proof. Recall that $p_t = \max\{1/n, \min\{1 - 1/n, X_{t-1}/\mu\}\}$. By the definition of expectation, we get

$$\mathbb{E}[p_t] = \frac{\Pr\left(X_{t-1} = 0\right)}{n} + \left(1 - \frac{1}{n}\right)\Pr\left(X_{t-1} = \mu\right) + \sum_{k=1}^{\mu-1} \frac{k \cdot \Pr\left(X_{t-1} = k\right)}{\mu}. \quad (14)$$

We note further that

$$\begin{aligned}\mathbb{E}[X_{t-1}] &= \sum_{k=0}^{\mu} k \Pr(X_{t-1} = k) \\ &= \mu \Pr(X_{t-1} = \mu) + \sum_{k=1}^{\mu-1} k \Pr(X_{t-1} = k),\end{aligned}$$

from which we then obtain

$$\sum_{k=1}^{\mu-1} k \cdot \Pr(X_{t-1} = k) = \mathbb{E}[X_{t-1}] - \mu \cdot \Pr(X_{t-1} = \mu). \quad (15)$$

Substituting (15) into (14) yields

$$\mathbb{E}[p_t] = \frac{\mathbb{E}\left[X_{t-1}\right]}{\mu} + \frac{\Pr\left(X_{t-1} = 0\right) - \Pr\left(X_{t-1} = \mu\right)}{n}. \quad (16)$$

We are left to calculate the two probabilities that $X_{t-1} = 0$ and $X_{t-1} = \mu$. Since these are unconditional probabilities, we shall make no assumption (even on p_{t-1}) when calculating them. All we know are that $p_0 = 1/2$ and, by Lemma 8, X_{t-1} is binomially distributed with μ trials and success probability p_{t-1}, which means that there is no bias towards any border in the stochastic process $(X_t : t \in \mathbb{N})$. Due to this symmetry, we get

$$\Pr\left(X_{t-1} = \mu\right) = \Pr\left(X_{t-1} = 0\right). \quad (17)$$

Furthermore, by the tower rule we also have

$$\begin{aligned}\mathbb{E}[X_{t-1}] &= \mathbb{E}[\mathbb{E}[X_{t-1} \mid p_{t-1}]] \\ &= \mathbb{E}\left[\mathbb{E}\left[\mathrm{Bin}\left(\mu, p_{t-1}\right) \mid p_{t-1}\right]\right] \\ &= \mu \cdot \mathbb{E}[p_{t-1}]\end{aligned} \quad (18)$$

Substituting (17) and (18) into (16) yields $\mathbb{E}[p_t] = \mathbb{E}[p_{t-1}]$. Then by induction on time, we obtain

$$\mathbb{E}[p_t] = \mathbb{E}[p_{t-1}] = \mathbb{E}[p_{t-2}] = \ldots = \mathbb{E}[p_0] = \frac{1}{2},$$

which completes the proof. □

Lemma 13 gives us insights into the expected values of the marginals at any time $t \in \mathbb{N}$. One should not confuse the expectation with the actual value of the marginals. Friedrich et al. [15] showed a similar result for the UMDA without border that

even when the expectation stays at 1/2, the actual value of the marginal in iteration t can be close to the trivial lower or upper border due to its large variance. Very recently, Doerr and Zheng [9] obtained a tight bound of $\Theta(\mu)$ on the first hitting time of any trivial border for these marginals. Furthermore, Lehre and Nguyen [28] showed that the variance reaches a value of $\Theta(\mu^2)$ after only $\Omega(\mu)$ iterations.

Lemma 14 *Let $\mu \geq c \log n$ for some sufficiently large constant $c > 0$ and $\lambda \leq \mu e^{1-\varepsilon}/(1+\delta)$ for any constants $\varepsilon \in (0,1)$ and $0 < \delta \leq e^{1-\varepsilon} - 1$. Then, the $n - (\beta + 1) = \Omega(n)$ last bits cannot be sampled as all 1s during any $e^{\Omega(\mu)}$ iterations w.p. $1 - e^{-\Omega(n)}$.*

Proof Given that $\mu\lambda \geq (1+\delta)/e^{1-\varepsilon}$, by Lemma 11 we get $Z_t \leq \beta \leq n(1-\varepsilon)$ for any $t = \text{poly}(n)$ w.o.p. We shall prove the lemma by looking at the $n - (\beta + 2) \geq n - n(1-\varepsilon) - 2 = \varepsilon n - 2 = \Omega(n)$ last bit positions. Let us now consider the total number of zeros sampled at these bit positions in an iteration. We know by Lemma 13 (for $k = \beta + 2$) that their marginals stay at 1/2 in expectation, and we also know by Lemma 12 that the samplings at these bit positions are mutually independent. Therefore, by the linearity of expectations, the expected total number of zeros sampled there is

$$(n - (\beta + 2))\left(1 - \frac{1}{2}\right) = \Omega(n).$$

This means that in order to sample all ones at these bit positions there are still at least $\Omega(n)$ zeros to flip. In other words, we need to deviate a distance of $\Omega(n)$ below the expected value, and by a Chernoff-Hoeffding bound (see Lemma 6) such an event happens w.p. at most

$$2 \cdot \exp\left\{ -\frac{2(\Omega(n))^2}{n} \right\} = e^{-\Omega(n)}.$$

By the union bound, this event happens at least once in a polynomial number of iterations (in n) w.p. still at most $e^{-\Omega(n)}$. The proof is now complete. □

We are ready to show our main result.

Theorem 15 *The UMDA with a parent population size $\mu \geq c \log n$ for some sufficiently large constant $c > 0$ and a selective pressure satisfying*

$$\frac{\mu}{\lambda} \geq \frac{1+\delta}{e^{1-\varepsilon}}$$

for any constants $\varepsilon \in (0,1)$ and $0 < \delta \leq e^{1-\varepsilon} - 1$ has a runtime of $e^{\Omega(\mu)}$ on the LeadingOnes function w.p. $1 - e^{-\Omega(\mu)}$ and also in expectation.

Proof Due to the low selective pressure, we have $\beta \leq n(1-\varepsilon)$. We now consider the two phases as introduced above. During phase 1, the all-ones bitstring cannot be

sampled w.p. at least $1 - e^{-\Omega(n)}$ since by Lemma 14 the $\Omega(n)$ last bit positions cannot be sampled correctly with the same probability. If this phase lasts for $e^{\Omega(\mu)}$ iterations, then we are done, and the theorem holds trivially. Thus, we shall assume that phase 1 lasts for at most poly (n) iterations.

During phase 2, we have observed by Lemma 11 that the random variable Z_t exceeds β in an iteration $t \leq e^{k\mu}$ for some constant $k > 0$ w.p. at most $e^{-\Omega(\mu)}$, while in the same iteration the $\Omega(n)$ last bits are sampled as all ones w.p. at most $e^{-\Omega(n)}$ due to Lemma 14. Thus, the all-ones bitstring can be sampled in that iteration w.p. at most $e^{-\Omega(\mu)}$, and by the union bound the all-ones bitstring is sampled at least once in $e^{k\mu}$ iterations w.p. at most $e^{-\Omega(\mu)}$. Note also that the last statement only holds if the constant c (in $\mu \geq c \log n$) is chosen sufficiently large. Therefore, the algorithm takes at least $e^{\Omega(\mu)}$ iterations to optimise the function w.p. at least $1 - e^{-\Omega(\mu)}$.

By the law of total expectation [36], the expected runtime is at least $e^{\Omega(\mu)}(1 - e^{-\Omega(\mu)}) = e^{\Omega(\mu)}$, which completes the proof. □

4 Runtime Analysis Under High Selective Pressure

4.1 A New Lower Bound for the UMDA

When the selective pressure $\psi = \mu/\lambda$ is set too high such that the value of α, defined in Eq. 11, exceeds the problem instance size n, phase 1 will end when the μ fittest individuals are all-ones bitstrings. By Eq. 11, this case occurs when

$$\frac{\psi}{1-\delta} \leq \left(1 - \frac{1}{n}\right)^n$$

for any constant $\delta \in (0, 1)$. The right-hand side is at least $(1 - \delta)/e$ for any $n \geq (1 + \delta)/\delta$ [30], and the above inequality always holds if we choose the selective pressure $\psi \leq (1 - \delta)^2/e$, We now recall the following result [5, Theorem 4], which in this case yields the first upper bound on the expected runtime for the UMDA on the LeadingOnes function.

Theorem 16 *The UMDA with an offspring population size $\lambda \geq c \log n$ for some sufficiently large constant $c > 0$ and a selective pressure satisfying*

$$\frac{\mu}{\lambda} \leq \frac{1}{e(1+\delta)}$$

for any constant $\delta > 0$ has an expected runtime of $\mathcal{O}\left(n\lambda \log \lambda + n^2\right)$ on the LeadingOnes function.

Until now, we are still missing a lower bound on the expected runtime for the UMDA on the LeadingOnes function, and in this section, we aim at deriving such a lower bound.

Recall that the random variable Z_t, defined in Eq. 5, denotes the number of marginals, counting from the first bit position, which are set to the upper border $1 - 1/n$ in iteration t, and the random variable Z_t^*, defined in Eq. 6, denotes the fitness value of the fittest individual. The following lemma shows the expected difference between these two random variables in an arbitrary iteration $t \in \mathbb{N}$. We pessimistically assume that the Z_t first marginals are all set to one since we are only interested in a lower bound and this will speed up the optimisation process.

Lemma 17 *It holds for any $t \in \mathbb{N}$ that $\mathbb{E}[Z_t^* - Z_t] < \log(2e\mu)$.*

Proof Let $\delta_t := Z_t^* - Z_t$. Consider the bit positions $Z_t + 2, Z_t + 3, \ldots, n$ among the μ fittest individuals. We shall view this as an abstract population of μ individuals, each of length $n - (Z_t + 1)$, and also let $\delta_t' := Z_t^* - (Z_t + 1) = \delta_t - 1$. In other words, δ_t' is a random variable describing the number of leading 1s of the fittest individual in this abstract population. We first note that if $X_{t,Z_t+1} = 0$, then $Z_t^* = Z_t$ and $\delta_t = 0$. By the law of total expectation, we get

$$\begin{aligned}
\mathbb{E}\left[\delta_t \mid Z_t\right] &= \overbrace{\mathbb{E}\left[\delta_t \cdot \mathbb{1}_{\{X_{t,Z_t+1}=0\}} \mid Z_t\right]}^{=0} + \mathbb{E}\left[\delta_t \cdot \mathbb{1}_{\{X_{t,Z_t+1}>0\}} \mid Z_t\right] \\
&= \mathbb{E}\left[(1+\delta_t') \cdot \mathbb{1}_{\{X_{t,Z_t+1}>0\}} \mid Z_t\right] \\
&\leq 1 + \mathbb{E}\left[\delta_t' \cdot \mathbb{1}_{\{X_{t,Z_t+1}>0\}} \mid Z_t\right].
\end{aligned}$$

We are left to calculate the last conditional expectation. Consider again the abstract population introduced above. The probability of sampling at most k leading 1s in this population is $1 - \prod_{i=Z_t+2}^{(Z_t+2)+k} p_{t,i}$, and the probability that all μ in the abstract population have more than k leading 1s is

$$1 - \left(1 - \prod_{i=Z_t+2}^{(Z_t+2)+k} p_{t,i}\right)^{\mu}.$$

Because $\mathbb{E}[Y] \leq \sum_{i=0}^{\infty} \Pr(Y > i)$ for any bounded integer-valued random variable Y, we then get

$$\mathbb{E}\left[\delta_t' \cdot \mathbb{1}_{\{X_{t,Z_t+1}>0\}} \mid Z_t, p_{t,Z_t+2}, \ldots, p_{t,n}\right] \leq \sum_{k=0}^{\infty} \left(1 - \left(1 - \prod_{i=Z_t+2}^{(Z_t+2)+k} p_{t,i}\right)^{\mu}\right).$$

We know, by Lemma 13, that the values of the marginals $p_{t,i}$ for each $i \geq Z_t + 2$ stay at 1/2 in expectation and also, by Lemma 12, that the samplings at these bit positions are pairwise independent. Note also that $x \mapsto (1-x)^{\mu}$ is a convex function for any $x \in [0,1]$, so by Jensen's inequality for convexity [44, p. 61] we get $\mathbb{E}[(1-x)^{\mu}] \geq (1 - \mathbb{E}[x])^{\mu}$. Thus,

$$
\begin{aligned}
&\mathbb{E}\left[\delta'_t \cdot \mathbb{1}_{\{X_{t,Z_t+1}>0\}} \mid Z_t\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\delta'_t \cdot \mathbb{1}_{\{X_{t,Z_t+1}>0\}} \mid Z_t, p_{t,Z_t+2}, \dots, p_{t,n}\right] \mid Z_t\right] && \text{(by tower rule)} \\
&\leq \sum_{k=0}^{\infty} \left(1 - \mathbb{E}\left[\left(1 - \prod_{i=Z_t+2}^{(Z_t+2+k)} p_{t,i}\right)^{\mu} \mid Z_t\right]\right) && \text{(by linearity of expectation)} \\
&\leq \sum_{k=0}^{\infty} \left(1 - \left(1 - \mathbb{E}\left[\prod_{i=Z_t+2}^{(Z_t+2+k)} p_{t,i} \mid Z_t\right]\right)^{\mu}\right) && \text{(by Jensen's inequality)} \\
&\leq \sum_{k=0}^{\infty} \left(1 - \left(1 - \prod_{i=Z_t+2}^{(Z_t+2+k)} \mathbb{E}\left[p_{t,i} \mid Z_t\right]\right)^{\mu}\right) && \text{(by independence, Lemma 12)} \\
&= \sum_{k=0}^{\infty} (1 - (1 - 2^{-(k+1)})^{\mu}) && \text{(by Lemma 13)} \\
&< \log \mu + 2, && \text{(by the proof of Lemma 7)}
\end{aligned}
$$

which completes the proof. □

Lemma 17 gives an important insight that the two random variables Z_t and Z_t^* only differ by a logarithmic additive term at any point in time in expectation. The global optimum is found when the random variable Z_t^* reaches the value of n. We can therefore alternatively analyse the random variable Z_t instead of Z_t^*. In other words, the random variable Z_t, starting from an initial value Z_0 given in Lemma 7, has to travel an expected distance of $n - \mathcal{O}(\log \mu) - Z_0$ (at bit positions) before the global optimum is found. We shall apply the additive drift theorem (for a lower bound) [21] for a potential function $g(x) = n - x$ on the stochastic process $(Z_t : t \in \mathbb{N})$. The single-step change (also called drift) is

$$\Delta_t := g(Z_t) - g(Z_{t+1}) = Z_{t+1} - Z_t.$$

We are ready to show a lower bound on the expected runtime of the UMDA on the LeadingOnes function.

Theorem 18 *The UMDA with a parent population size $\mu \geq c \log n$ for some sufficiently large constant $c > 0$ and a selective pressure satisfying*

$$\frac{\mu}{\lambda} \leq \frac{(1-\delta)^2}{e}$$

for any constant $\delta \in (0, 1)$ has an expected runtime of $\Omega(n\lambda/\log \lambda)$ on the LeadingOnes function.

Proof Let $i := Z_t + 1$. By definition, $Z_{t+1} = Z_t$ and $\Delta_t = 0$ if there are less than μ individuals with at least i leading 1s sampled in the next iteration (i.e., $C_{t+1,i} < \mu$). Thus, the drift is maximised when $C_{t+1,i} \geq \mu$. By the law of total expectation, we then get

$$\begin{aligned}\mathbb{E}\left[\Delta_t \mid Z_t\right] &= \overbrace{\mathbb{E}\left[\Delta_t \cdot \mathbb{1}_{\{C_{t+1,i}<\mu\}} \mid Z_t\right]}^{=0} + \mathbb{E}\left[\Delta_t \cdot \mathbb{1}_{\{C_{t+1,i}\geq\mu\}} \mid Z_t\right] \\ &= \mathbb{E}\left[\Delta_t \mid C_{t+1,i} \geq \mu, Z_t\right] \cdot \Pr(C_{t+1,i} \geq \mu \mid Z_t) \\ &\leq \mathbb{E}\left[\Delta_t \mid C_{t+1,i} \geq \mu, Z_t\right] \\ &= \mathbb{E}\left[Z_{t+1} \mid C_{t+1,i} \geq \mu, Z_t\right] - Z_t\end{aligned}$$

We are left to bound the expectation. Given Z_t, we know by Lemma 13 that the marginals of bit positions from $Z_t + 2$ to n stay at 1/2 in expectation, and also by Lemma 8 the samplings at these bit positions are pairwise independent. By following the proof of Lemma 17, we can quickly upper bound the required expectation as follows.

$$\mathbb{E}\left[Z_{t+1} \mid C_{t+1,i} \geq \mu, Z_t\right] \leq Z_t + 1 + \mathcal{O}(\log \lambda).$$

Then, the expected drift is

$$\mathbb{E}[\Delta_t \mid Z_t] = \mathcal{O}(\log \lambda).$$

Because the random variable Z_t has to travel an expected distance of $n - \mathcal{O}(\log \lambda) - Z_0$ before the global optimum is found, by the additive drift theorem [21] the expected number of iterations, conditional on Z_0, until the optimum is found for the first time is upper bounded by $(n - \mathcal{O}(\log \lambda) - Z_0)/\mathcal{O}(\log \lambda)$. Note that $\mathbb{E}[Z_0] = \mathcal{O}(\log \lambda)$, there are λ function evaluations performed in each iteration, and by the tower rule, we then obtain an overall expected runtime of

$$\lambda \cdot \frac{n - \mathcal{O}(\log \lambda) - \mathbb{E}[Z_0]}{\mathcal{O}(\log \lambda)} = \Omega\left(\frac{n\lambda}{\log \lambda}\right),$$

which completes the proof. □

4.2 A Tighter Upper Bound for the PBIL

In this section, we aim at showing a tighter upper bound than the upper bound of $\mathcal{O}\left(n^{2+c}\right)$ in [47] for the PBIL on the LeadingOnes function. We shall apply the level-based theorem. To begin with, we first remark that Algorithm 2 assumes a mapping $\mathcal{D}$ from the space of populations $\mathcal{X}^\lambda$ to the space of probability distributions over the search space. The mapping $\mathcal{D}$ is often said to depend on the current population only [3]; however, it is not always necessary, especially for the PBIL with a sufficiently large offspring population size λ. The rationale behind this is that in each iteration the PBIL draws λ samples from the product distribution, specified in Eq. 2, that correspond to λ individuals in the current offspring population. If the number of samples λ is sufficiently large, it is very unlikely that the many empirical frequencies of ones deviate far from the true marginals. We will make this intuition more rigorous via the DKW inequality (see Theorem 2).

We shall use a canonical partition of the search space, where each subset A_j contains bitstrings with j leading 1s.

$$A_j := \{x \in \mathcal{X} : \text{LeadingOnes}(x) = j\}. \tag{19}$$

Thus, there are $n+1$ levels, ranging from A_0 to A_n. We then need to verify three conditions in Theorem 1. Recall that $A_{\geq j} = \cup_{i=j}^{n} A_i$. For conditions (G1) and (G2), we assume that there are at least $\gamma_0 \lambda$ individuals in levels $A_{\geq j}$ in iteration t. Following [5], we choose $\gamma_0 = \mu/\lambda$. This implies that the μ fittest individuals have at least j leading 1s. We define

$$\widetilde{p}_{t,i} := \frac{1}{\lambda} \sum_{j=1}^{\lambda} x_i^{(j)}$$

to be the frequency of ones at bit position i in the entire population of λ individuals. We now show under the assumption of the condition (G1) of the level-based theorem that if the population size is $\lambda = \Omega(\log n)$, the first j marginals cannot be too close to the lower border $1/n$ with high probability.

Lemma 19 *Assume that* $|P_t \cap A_{\geq j}| \geq \gamma_0 \lambda$ *and* $\lambda \geq c((1+1/\varepsilon)/\gamma_0)^2 \ln(n)$ *for any constants* c, $\varepsilon > 0$ *and* $\gamma_0 \in (0,1)$, *then*

(a) $\prod_{i=1}^{j} p_{t,i} \geq \gamma_0/(1+\varepsilon)$ *w.p. at least* $1 - 2n^{-2c}$, *and*
(b) $p_{t,i} \geq \gamma_0/(1+\varepsilon)$ *w.p. at least* $1 - 2n^{-2c}$ *for an arbitrary* $i \in [j]$.

Proof We only show the first statement as the second follows from the first statement. Let Q_i be the number of ones sampled among the j first bit positions in the i-th individual in the current population P_t. By the assumption $|P_t \cap A_{\geq j}| \geq \gamma_0 \lambda$ on the current population, the empirical distribution function of Q_i must satisfy

$$\hat{F}_\lambda(j-1) = \frac{1}{\lambda} \sum_{i=1}^{\lambda} \mathbb{1}_{\{Q_i \leq j-1\}} \leq 1 - \hat{q}_t,$$

where $\hat{q}_t \geq \gamma_0$ is the fraction of individuals in the current population with j leading ones, while the true distribution function satisfies $F(j-1) = 1 - q_t$, where $q_t := \prod_{i=1}^{j} p_{t,i}$ is the probability of sampling at least j leading ones in an individual. The DKW inequality yields that

$$\Pr(\hat{q}_t - q_t > \phi) \leq \Pr(|\hat{q}_t - q_t| > \phi) \leq 2e^{-2\lambda\phi^2}$$

for all $\phi > 0$. Therefore, with probability at least $1 - 2e^{-2\lambda\phi^2}$ it holds $\hat{q}_t - q_t \leq \phi$ and, thus, $q_t \geq \hat{q}_t - \phi \geq \gamma_0 - \phi$. Choosing $\phi := \varepsilon\gamma_0/(1+\varepsilon)$, we get $q_t = \prod_{i=1}^{j} p_{t,i} \geq \gamma_0(1 - \varepsilon/(1+\varepsilon)) = \gamma_0/(1+\varepsilon)$ with probability at least $1 - 2e^{-2\phi^2\lambda} \geq 1 - 2n^{-2c}$. □

Lemma 19 tells us that if the current level of the population is j, then all marginals $p_{t,1}, p_{t,2}, \ldots, p_{t,j}$ are at least $\gamma_0/(1+\varepsilon)$ in an iteration $t \in \mathbb{N}$ with probability polynomially close to one. To show an upper bound on the expected runtime for the PBIL on the LeadingOnes function, we first apply the level-based theorem to obtain an upper bound conditional on the event that for all iterations $t \leq t_*$, and $1 \leq i \leq j$, where j is the current level in iteration t, satisfy $p_{t,i} \geq \gamma_0/(1+\varepsilon)$ where t_* is a sufficiently long time interval which will be specified later. In the end, we follow the line of argumentations put forward in [10, Theorem 8] to derive an overall unconditional expected runtime.

We first introduce the AM-GM inequality [34].

Lemma 20 (AM-GM Inequality) *Let $a_1, \ldots, a_n$ be n non-negative real numbers. It holds that*

$$\sum_{i=1}^{n} \frac{a_i}{n} \geq \prod_{i=1}^{n} a_i^{\frac{1}{n}}.$$

Equality occurs if and only if $a_1 = a_2 = \cdots = a_n$.

We are ready to establish an improved upper bound on the expected runtime of the PBIL on the LeadingOnes function. Surprisingly, the proof is straightforward and not very technically demanding compared to the proof in [47].

Theorem 21 *The PBIL with an offspring population size λ with $c \log n \leq \lambda = \text{poly}(n)$ for a sufficiently large constant $c > 0$, a constant smoothing parameter $\rho \in (1/e, 1]$ and a constant selective pressure satisfying*

$$\frac{\mu}{\lambda} \leq \left(\frac{\rho^{2+\ln(1+\varepsilon)}}{e(1+\varepsilon)} \right)^{1/\ln(e\rho)} \tag{20}$$

for any constant $\varepsilon > 0$, has an expected runtime of $\mathcal{O}\left(n\lambda \log \lambda + n^2\right)$ on the LeadingOnes function.

Proof First, we partition the search space into "levels" using the canonical partition defined in (19), in which each subset A_j contains individuals with exactly j leading 1s.There are a total of $n+1$ levels ranging from A_0 to A_n.

Let $\tau := T/\lambda$ denote the runtime of the algorithm in terms of number of iterations. We say that failure event F_t occurs in iteration $t \in \mathbb{N}$ if there exist two indices $i, j \in \mathbb{N}$ satisfying $1 \leq i \leq j \leq n$ such that $|P_t \cap A_{\geq j}| \geq \gamma_0 \lambda$ and $p_{t,i} < \gamma_0/(1+\varepsilon)$, where ε and γ_0 are parameters which will be specified later. Furthermore, for any $t \geq 0$, we let $G_t := \bigwedge_{i=0}^{t} \neg F_i$ denote the event that there is no failure in the first t iterations. We will first estimate the expected runtime of the algorithm starting from any initial state, conditional on the event $G_{\tau \vee s}$, i.e., that no failure occurs before the optimum has been found for the first time or before iteration s, whichever is the larger. Here, s is a parameter we will define later, and $x \vee y := \max(x, y)$. Note that $\Pr\left(G_{\tau \vee s}\right) > 0$ because any new individual in any iteration is optimal with probability

at least $n^{-n} > 0$. Afterwards, we will estimate the overall expected runtime of the algorithm on the function.

To obtain an upper bound on the expected runtime conditional on the event $G_{\tau \vee s}$, we apply the level-based theorem with respect to the partition $A_0, \dots, A_n$ described above.

For the two conditions (G1) and (G2), assuming that $|P_t \cap A_{\geq j}| \geq \gamma_0 \lambda = \mu$, we are required to show that the probability of sampling an offspring in levels $A_{\geq j+1}$ in iteration $t+1$ is lower bounded by $(1+\delta)\gamma$ for some constant $\delta > 0$. We note by Lemma 4 that this probability can be bounded from below as follows:

$$\prod_{i=1}^{j+1} p_{t+1,i} \geq \left(\prod_{i=1}^{j} q_i \right) \cdot p_{t+1,j+1},$$

that holds for any vector $q := (q_1, \dots, q_j)$, which majorises the vector $p^*_{t+1} := (p_{t+1,1}, \dots, p_{t+1,j})$. In the remainder of this proof, we shall construct such a vector q from vector p^*_{t+1}.

In order to construct vector q, we will shift the weight $\sum_{i=1}^{j} p_{t+1,i}$ as far as possible to the marginals with smaller indices. The trivial upper bound on each component q_i is the upper border $1 - 1/n$. For the lower bound, we note from the assumption $|P_t \cap A_{\geq j}| \geq \mu$ that the μ fittest individuals have at least j leading 1s, meaning that when updating the model we always have $p_{t+1,i} = (1-\rho)p_{t,i} + \rho \geq \rho$ for each $i \in [j]$. Therefore, a trivial lower bound on each component q_i is the smoothing parameter ρ. We define a vector $q = (q_1, \dots, q_j)$ as follows:

$$q_i = \begin{cases} 1 - 1/n, & \text{if } 0 \leq i \leq m, \\ \rho, & \text{if } m+2 \leq i \leq j, \\ \sum_{i=1}^{j} p_{t+1,i} - m(1-1/n) - (j-m-1)\rho, & \text{if } i = m+1, \end{cases} \tag{21}$$

for an integer $m = \lfloor g(j) \rfloor$, where

$$g(j) = \frac{\sum_{i=1}^{j} p_{t+1,i} - j\rho}{1 - 1/n - \rho} \geq \frac{\sum_{i=1}^{j} p_{t+1,i} - j\rho}{1 - \rho}. \tag{22}$$

Because of the floor function, we always get $g(j) - 1 < m \leq g(j)$, and thus $\rho \leq q_{m+1} \leq 1 - 1/n$, meaning that the defined value of the component q_{m+1} is indeed a probability. By the definition of the vector q in (21), we have for any $k \in [j-1]$ that

$$\sum_{i=1}^{k} q_i \geq \sum_{i=1}^{k} p_{t+1,i}$$

and

$$\sum_{i=1}^{j} q_i = \sum_{i=1}^{j} p_{t+1,i}.$$

Therefore, according to Definition 3 the vector q majorises the vector p^*_{t+1}. By Lemma 4, the probability of sampling the first j bits correctly is

$$\prod_{i=1}^{j} p_{t+1,i} \geq \prod_{i=1}^{j} q_i \geq \left(1 - \frac{1}{n}\right)^m \cdot \rho^{j-m} \geq \frac{\rho^{j-m}}{e}, \tag{23}$$

which holds because $(1 - 1/n)^m \geq 1/e$ for any integer $m < n$. Recall that we aim at showing that the above probability is at least a constant, so we are done if we can show that $j - m = \mathcal{O}(1)$. We are going to show that this is indeed the case.

Let $p_0 := \gamma_0/(1+\varepsilon)$. We get by Lemmas 19 and 20 that the weight

$$\sum_{i=1}^{j} p_{t,i} \geq j \cdot \left(\prod_{i=1}^{j} p_{t,i}\right)^{1/j} \geq j \cdot p_0^{1/j}; \tag{24}$$

thus,

$$\sum_{i=1}^{j} p_{t+1,i} = (1-\rho) \sum_{i=1}^{j} p_{t,i} + j\rho \geq (1-\rho) j p_0^{1/j} + \rho j.$$

We also have the following.

$$\begin{aligned}
j - m &< j - (g(j) - 1) && \text{(since } m > g(j) - 1\text{)} \\
&\leq j + 1 - \frac{(1-\rho) j p_0^{1/j} + \rho j - j\rho}{1-\rho} && \text{(by Eq. 22)} \\
&= 1 + j(1 - p_0^{1/j}) \\
&\leq 1 + j \cdot (-\ln(p_0)/j) \\
&= 1 - \ln(p_0),
\end{aligned}$$

where the last inequality follows the fact that $\ln(x) \leq n(x^{1/n} - 1)$ for all $n > 0$ and $x > 0$ [39]. Since the value γ_0 is assumed constant, and so is the value $p_0 = \gamma_0/(1+\varepsilon)$ for any constant $\varepsilon > 0$; thus,

$$j - m \leq 1 - \ln(p_0) = \mathcal{O}(1), \tag{25}$$

meaning the the probability of sampling the first j bits correctly in iteration $t+1$ is at least a constant. In the remainder of the proof, we will use this result to verify conditions (G1) and (G2) of the level-based theorem.

For condition (G1), we are interested in a lower bound z_j on the probability of sampling an offspring in levels $A_{\geq j+1}$ in iteration $t+1$. Because the marginal $p_{t+1,j+1} \geq 1/n$, this probability is

$$\left(\prod_{i=1}^{j} p_{t+1,i}\right) \cdot p_{t+1,j+1} \geq \frac{\rho^{j-m}}{e} \cdot \frac{1}{n} \quad \text{(by Eq. 23)}$$
$$\geq \frac{\rho^{\mathcal{O}(1)}}{en} \quad \text{(by Eq. 25)}$$
$$= \Omega\left(\frac{1}{n}\right).$$

Thus, condition (G1) is satisfied with the lower bound $z_* = z_j = \Omega(1/n)$.

For condition (G2), assuming further that $|P_t \cap A_{\geq j+1}| \geq \gamma\lambda$, meaning that the marginal at bit position $j+1$ will be set to $p_{t+1,j+1} \geq (1-\rho)p_{t,j+1} + \rho(\gamma\lambda)/\mu \geq \rho\gamma/\gamma_0$. In this case, the probability of sampling the first $j+1$ bits correctly is

$$\left(\prod_{i=1}^{j} p_{t+1,i}\right) \cdot p_{t+1,j+1} \geq \frac{\rho^{1-\ln p_0}}{e} \cdot \frac{\rho\gamma}{\gamma_0} \geq \frac{\rho^{2-\ln(p_0)}\gamma}{e\gamma_0} = \frac{\rho^2\gamma}{\rho^{\ln(p_0)}\gamma_0 e}. \tag{26}$$

For $\rho = 1$, the lower bound in (26) becomes $\gamma/(e\gamma_0)$, and the condition (G2) can be easily confirmed by setting $\gamma_0 \leq 1/(e(1+\varepsilon))$ for any constant $\varepsilon > 0$. We note that this is already obtained in Theorem 16 for the UMDA on the LeadingOnes function. Otherwise, if the smoothing parameter $\rho < 1$, we can rewrite

$$\rho^{\ln(p_0)} = p_0^{\ln\rho} = \frac{\gamma_0^{\ln\rho}}{(1+\varepsilon)^{\ln\rho}}$$

for any constant $\varepsilon > 0$, then (26) is equivalent to

$$\geq \frac{\rho^2\gamma}{\gamma_0 e} \cdot \frac{(1+\varepsilon)^{\ln\rho}}{\gamma_0^{\ln\rho}} = \frac{\rho^2\gamma(1+\varepsilon)^{\ln\rho}}{e\gamma_0^{1+\ln\rho}} \geq (1+\varepsilon)\gamma,$$

which always holds if we choose the value γ_0 such that

$$\gamma_0^{1+\ln\rho} \leq \frac{\rho^2}{e(1+\varepsilon)^{1-\ln\rho}} = \frac{\rho^{2+\ln(1+\varepsilon)}}{e(1+\varepsilon)}. \tag{27}$$

For any $\rho \in (0,1]$, the right-hand side of Eq. 27 is always less than one as $1 - \ln\rho \geq 1$, so in the left-hand side we require $1 + \ln\rho > 0$, which is equivalent to $\rho > 1/e$. We then obtain the following bound on γ_0:

$$\gamma_0 \leq \left(\frac{\rho^{2+\ln(1+\varepsilon)}}{e(1+\varepsilon)}\right)^{1/\ln(e\rho)}. \tag{28}$$

The smoothing parameter ρ is a constant, so is the upper bound on γ_0. In the end, condition (G2) of Theorem 1 is verified.

To satisfy the condition (G3), it suffices to choose $\lambda \geq c \log n$ for a sufficiently large constant $c > 0$.

Having verified the three conditions (G1), (G2) and (G3), and noting that $\ln\left(6\delta\lambda/(4+\delta\lambda z_j)\right) < \ln(3\delta\lambda/2)$, Theorem 1 now guarantees an upper bound, for some constant $c_1 > 0$,

$$\mathbb{E}\left[T \mid G_{\tau \vee s}\right] \leq \left(\frac{8}{\delta^2}\right) \sum_{j=0}^{n-1} \left[\lambda \ln\left(\frac{3\delta\lambda}{2}\right) + \frac{1}{z_j}\right] \leq c_1(n\lambda \log \lambda + n^2) =: \lambda t_*. \quad (29)$$

To obtain an upper bound on the unconditional expected runtime, we divide the run into consecutive phases, each of length $s := 2t_* = \text{poly}(n)$ iterations. Note that for all $i \in \mathbb{N}$, the event $\tau \leq s$ is independent of the failure event F_{s+i}. By Lemma 28[1] and (29), it follows that the probability that the algorithm finds the optimum within one phase is

$$\Pr(\tau \leq s) \geq \Pr\left(G_s\right)(1 - t_*/s) = \Pr\left(G_s\right)/2. \quad (30)$$

We now estimate the probability of the event G_s. By Lemma 19 and a union bound, failure event F_t occurs with probability at most $2n^{-2c+1}$ assuming the population size satisfies $\lambda \geq c((1+1/\varepsilon)/\gamma_0)^2 \ln(n)$ for a constant $c > 0$. By another union bound and assuming that c is chosen sufficiently large, the probability of no failure within $s = \text{poly}(n)$ iterations is

$$\Pr\left(G_s\right) \geq 1 - s2n^{-2c+1} = 1 - o(1). \quad (31)$$

From (30) and (31), it follows that the algorithm, starting from any initial state, finds the optimum within a phase with probability at least $1/2 - o(1)$.

If the algorithm does not find the optimum within a phase or event G_s does not hold, the algorithm enters some unknown state at the end of the current phase. Because our analysis makes no assumption about the state of the algorithm at the beginning of the phase, we can repeat the same analysis for the next phase. Hence, the number of phases until an optimum is found for the first time is stochastically dominated by a geometric random variable [35, Definition 2.8] with success probability $1/2 - o(1)$. By [7, Corollary 8.3] and [35, p. 32], the expected number of iterations until an optimum is found for the first time is at most $1/(1/2 - o(1)) = \mathcal{O}(1)$.

It follows that the overall expected runtime of the PBIL on the LeadingOnes function is $\mathcal{O}(\lambda s) = \mathcal{O}\left(n\lambda \log \lambda + n^2\right)$, which completes the proof. □

We note from Eq. 28 that the threshold on the selective pressure is a function of the smoothing parameter $\rho \in (1/e, 1]$, denoted by $h(\rho)$. When $\rho \to 1$, that is, the PBIL converges to the UMDA, $h(\rho) \to 1/(e(1+\varepsilon))$, which matches the selective pressure considered in Theorem 16. Also, $h(\rho)$ is an increasing function and has a very small value when ρ gets closer to $1/e \approx 0.3679$ (see Fig. 1). In other words, we need to pick an extremely high selective pressure when the smoothing parameter ρ approaches $1/e$ (from above).

[1] In the Appendix.

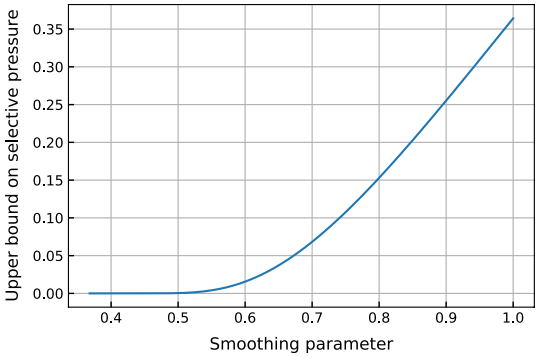

Fig. 1 Threshold on the selective pressure for the PBIL with $\rho \in (1/e, 1]$ on the LeadingOnes function in Eq. 28 with $\varepsilon = 0.01$. Note also that $1/e \approx 0.3679$ and $1/(e(1+\varepsilon)) \approx 0.3642$

4.3 Direct Extensions

The function

$$\text{BinVal}(x) := \sum_{i=1}^{n} 2^{n-i} x_i$$

is another test function also widely used in runtime analyses of EDAs [6, 10, 27, 46]. This is a linear function where the bit weights decrease exponentially with bit positions. Due to some similarity with the LeadingOnes function, we will show that the runtime bound derived in Theorem 21 can be extended to the BinVal function. We first partition the search space into non-empty disjoint subsets $A_0, \ldots, A_n$ as follows.

$$A_j := \left\{ x \in \mathcal{X} \mid \sum_{i=1}^{j} 2^{n-i} \leq \text{BinVal}(x) < \sum_{i=1}^{j+1} 2^{n-i} \right\}$$

for $j \in [n] \cup \{0\}$, where $\sum_{i=1}^{0} 2^{n-i} = 0$. The following lemma formalises the similarity between the two functions.

Lemma 22 *$x \in A_j$ if and only if* $\text{LeadingOnes}(x) = j$.

Proof For the sufficient condition, if $x \in A_j$, meaning that

$$\sum_{i=1}^{j} 2^{n-i} \leq \text{BinVal}(x) < \sum_{i=1}^{j+1} 2^{n-i},$$

then the first j bits must be 1s, followed by a 0 at bit position $j+1$. This is due to the fact that $2^{n-(j+1)} > \sum_{i=j+2}^{n} 2^{n-i}$. For the necessary condition, if $\text{LeadingOnes}(x) = j$, the first j bits are 1s, followed by a 0 at bit position $j+1$. The BinVal-value of the bitstring is at most

$$\sum_{i \neq j+1} 2^{n-i} < \sum_{i=1}^{j+1} 2^{n-i}.$$

Therefore, x must be in the level of A_j. □

We now consider the sorting of individuals after the population is sampled in an arbitrary iteration. For the LeadingOnes function, all that matters to determine the ranking of a bitstring is the number of leading 1s. Alternatively, we can say the ranking of an individual depends on the position of the leftmost zero in the bitstring, and all following bits have no contribution to the overall fitness of the individual. However, this is not the case for the BinVal function, where all individuals are first sorted according to their LeadingOnes-values. Ties are broken not uniformly at random as for the LeadingOnes function but by comparing the number of leading 1s following the leftmost zero among these individuals. However, since the proof of Theorem 21 never takes bits after the leftmost zero into account, the result also holds for the BinVal function. The following corollary yields the first upper bound on the expected runtime of the PBIL on the BinVal function. We note that a similar bound for the UMDA on the BinVal function is shown in [6].

Corollary 23 *The PBIL with an offspring population size $\lambda \geq c \log n$ for a sufficiently large constant $c > 0$, a constant smoothing parameter $\rho \in (1/e, 1]$, and a constant selective pressure satisfying*

$$\frac{\mu}{\lambda} \leq \left(\frac{\rho^{2+\ln(1+\varepsilon)}}{e(1+\varepsilon)} \right)^{1/\ln(e\rho)}$$

for any constant $\varepsilon > 0$, has an expected runtime of $\mathcal{O}\big(n\lambda \log \lambda + n^2\big)$ on BinVal.

Furthermore, due to the similarity between the PBIL and the λ-MMAS [42], we are now able to establish the expected runtime of the λ-MMAS on the LeadingOnes and BinVal functions. For the λ-MMAS, we have $\mu = 1$, substituting this into Eq. 20 and noting also that $\lambda \geq c \log n$, we then obtain

$$\lambda \geq \max \left\{ c \log n, \left(\frac{\rho^{2+\ln(1+\varepsilon)}}{e(1+\varepsilon)} \right)^{1/\ln(e\rho)} \right\} = \Omega(\log n).$$

Corollary 24 *The λ-MMAS with a population size $\lambda \geq c \log n$ for a sufficiently large constant $c > 0$ has an expected runtime of $\mathcal{O}\big(n\lambda \log \lambda + n^2\big)$ on the LeadingOnes and BinVal functions.*

5 Runtime Analyses on Noisy LeadingOnes

We also consider a prior noise model and formally define the problem for any constant $0 < p < 1$ as follows.

$$F(x_1, \dots, x_n) = \begin{cases} f(x_1, \dots, x_n), & \text{w.p. } 1-p, \text{ and} \\ f(\dots, 1-x_i, \dots), & \text{w.p. } p, \text{ where } i \sim \text{Unif}([n]). \end{cases}$$

We denote F as the noisy fitness and f as the actual fitness. For simplicity, we also denote P_t as the population prior to noise. The same noise model is studied in [4, 11, 17, 41, 43] for population-based EAs on the OneMax and LeadingOnes functions.

We shall make use of the level-based theorem and first partition the search space $\mathcal{X}$ into $n+1$ disjoint subsets $A_0, \dots, A_n$ as in Eq. 19. Recall that $A_{\geq j} = \cup_{i=j}^{n} A_i$. We then need to verify three conditions (G1), (G2) and (G3) of the level-based theorem, where due to the presence of noise we choose the parameter $\gamma_0 = \psi/((1-\delta)(1-p))$ for any constant $\delta \in (0,1)$ and the selective pressure $\psi = \mu/\lambda$ to leverage the impact of noise in our analysis. The following lemma tells us the number of individuals in the population in iteration t which have fitness $F(x) = f(x) \geq j$.

Lemma 25 *Assume that $|P_t \cap A_{\geq j}| \geq \gamma_0 \lambda$, where $\gamma_0 := \psi/((1-p)(1-\delta))$ for some constant $\delta \in (0,1)$, and $\psi = \mu/\lambda$ is assumed constant. Then, there are at least μ individuals with the fitness $F(x) = f(x) \geq j$ in the noisy population w.p. $1 - e^{-\Omega(\mu)}$.*

Furthermore, there are at most $\varepsilon\mu$ individuals with actual fitness $f(x) \leq j-1$ and noisy fitness $F(x) \geq j$ for some small constant $\varepsilon \in (0,1)$ w.p. $1 - e^{-\Omega(\mu)}$.

Proof We take an alternative view on the sampling of the population and the application of noise. More specifically, we first sample the population, sort it in descending order according to the true fitness, and then noise occurs at any individual w.p. p. Because noise does not occur at an individual w.p. $1-p$, amongst the $\gamma_0\lambda$ individuals in levels $A_{\geq j}$, in expectation there are

$$(1-p)\gamma_0\lambda = \frac{\psi\lambda}{1-\delta} = \frac{\mu}{1-\delta}$$

individuals unaffected by noise. Furthermore, by a Chernoff bound [36], there are at least $(1-\delta)\cdot\mu/(1-\delta) = \mu$ such individuals for some constant $0 < \delta < 1$ w.p. at least $1 - e^{-(\delta^2/2)\cdot\mu/(1-\delta)} = 1 - e^{-\Omega(\mu)}$, which proves the first statement.

For the second statement, we only consider individuals with actual fitness $f(x) < j$ and noisy fitness $F(x) \geq j$ in the population. If such an individual is selected when updating the model, it will introduce a 0 to the total number of 0s among the μ fittest individuals for the first j bits. Let B denote the number of such individuals. There are at most $(1-\gamma_0)\lambda$ individuals with actual fitness $f(x) < j$, so the probability that their noisy fitness values are at least $F(x) \geq j$ is at most p/n because a specific bit must be flipped in the prior noise model. Hence the expected number of these individuals is upper bounded by

$$\mathbb{E}[B] \leq \frac{(1-\gamma_0)\lambda p}{n} < \frac{\lambda p}{n}. \tag{32}$$

We now show by a Chernoff bound that the event $B \geq \varepsilon\mu$ for a small constant $\varepsilon \in (0,1)$ occurs w.p. at most $e^{-\Omega(\mu)}$. We shall rely on the fact that $\lambda p/n \leq \mu\varepsilon/2$ for sufficiently large n, which follows from the assumption $\mu/\lambda = \mathcal{O}(1)$. We use the parameter $\delta := \varepsilon\mu/\mathbb{E}[B] - 1$, which by (32) and the assumption $\lambda p/n \leq \varepsilon\mu/2$ satisfies $\delta \geq \varepsilon\mu n/(p\lambda) - 1 \geq 1$. We also have the lower bound

$$\delta \cdot \mathbb{E}[B] = \varepsilon\mu - \mathbb{E}[B] \geq \varepsilon\mu - \frac{\lambda p}{n} \geq \frac{\varepsilon\mu}{2}.$$

A Chernoff bound [36] now gives the desired result

$$\Pr(B \geq \varepsilon\mu) = \Pr(B \geq (1+\delta)\mathbb{E}[B]) \leq e^{-\delta\mathbb{E}[B]/3} = e^{-\varepsilon\mu/6}, \tag{33}$$

which completes the proof. □

We now derive upper bounds on the expected runtime of the UMDA on the LeadingOnes function in the noisy environment.

Theorem 26 *Consider a prior noise model with constant parameter $p \in (0,1)$. The UMDA with a parent population size $\mu \geq c \log n$ for some sufficiently large constant $c > 0$ and a constant selective pressure satisfying*

$$\frac{\mu}{\lambda} \leq \frac{1-\varepsilon}{e(1+\delta)}$$

for some constants $\varepsilon, \delta \in (0,1)$ has an expected runtime of $\mathcal{O}\big(n\lambda \log \lambda + n^2\big)$ on the LeadingOnes function.

Proof We will apply the level-based theorem. Each level A_j for $j \in [n] \cup \{0\}$ is formally defined as in (19), and there are a total of $m := n+1$ levels.

For the condition (G1), we assume that $|P_t \cap A_{\geq j}| \geq \gamma_0\lambda$, and we are required to show that the probability of sampling an offspring in levels $A_{\geq j+1}$ in iteration $t+1$ is lower bounded by a value z_j. We choose the parameter $\gamma_0 = \psi/((1-\delta)(1-p))$ for any constant $\delta \in (0,1)$ and the constant selective pressure $\psi = \mu/\lambda$. For convenience, we also partition the noisy population into four groups:

1. Individuals with fitness $f(x) \geq j$ and $F(x) \geq j$.
2. Individuals with fitness $f(x) \geq j$ and $F(x) < j$.
3. Individuals with fitness $f(x) < j$ and $F(x) \geq j$.
4. Individuals with fitness $f(x) < j$ and $F(x) < j$.

By Lemma 25, there are at least μ individuals in group 1 w.p. $1 - e^{-\Omega(\mu)}$. The algorithm selects the μ fittest individuals according to the noisy fitness values to update the probabilistic model. Hence, unless the mentioned event does not happen, no individuals from group 2 or group 4 will be included when updating the model.

We are now going to analyse how individuals from group 3 impact the marginal probabilities. Let B denote the number of individuals in group 3. We pessimistically assume that the algorithm uses all of the B individuals in group 3 and $\mu - B$ individuals chosen from group 1 when updating the model. For all $i \in [j]$, let Q_i be the number of individuals in group 3 which have 1s at bit positions 1 through j, except for one position i where they have a 0. By definition, we then have $\sum_{i=1}^{j} Q_i = B$. The marginal probabilities after updating the model are

$$p_{t,i} = \begin{cases} 1 - Q_i/\mu, & \text{if } Q_i > 0, \\ 1 - Q_i/\mu - 1/n, & \text{if } Q_i = 0. \end{cases} \tag{34}$$

Again by Lemma 4, we can lower bound the probability of sampling an offspring x with actual fitness $f(x) \geq j$, by

$$\prod_{i=1}^{j} p_{t,i} \geq \prod_{i=1}^{j} q_i, \tag{35}$$

which holds for any vector $q := (q_1, \ldots, q_j)$ which majorises the vector $(p_{t,1}, \ldots, p_{t,j})$. By Definition 3, we construct such a vector q which by the definition majorises the vector $(p_{t,1}, \ldots, p_{t,j})$ as follows.

$$q_i = \begin{cases} 1 - 1/n, & \text{if } i < j, \\ \sum_{k=1}^{j} p_{t,k} - (1 - 1/n)(j-1), & \text{if } i = j. \end{cases}$$

We now show that with high probability, the vector element q_j stays within the interval $[1 - 1/n - \varepsilon, 1 - 1/n]$, i.e., q_j is indeed a probability. Since $p_{t,i} \leq 1 - 1/n$ for all $i \leq j$, we have the upper bound $q_j \leq (1 - 1/n)j - (1 - 1/n)(j-1) = 1 - 1/n$. For the lower bound, we note from (34) that $p_{t,i} \geq 1 - Q_i/\mu - 1/n$ for all $i \leq j$ and any $Q_i \geq 0$, so we also obtain

$$\begin{aligned} q_j &\geq \sum_{k=1}^{j} \left(1 - \frac{Q_i}{\mu} - \frac{1}{n}\right) - \left(1 - \frac{1}{n}\right)(j-1) \\ &= 1 - \frac{1}{n} - \sum_{k=1}^{j} \frac{Q_i}{\mu} \\ &= 1 - \frac{1}{n} - \frac{B}{\mu}. \end{aligned}$$

By Lemma 25, we have $B \leq \varepsilon\mu$ for some small constant $\varepsilon \in (0, 1)$ w.p. $1 - e^{-\Omega(\mu)}$. Assume that this high-probability event actually happens, we therefore have $q_j \geq 1 - 1/n - \varepsilon$. From this result, the definition of the vector q and (35), we can conclude that the probability of sampling in iteration $t + 1$ an offspring x with actual fitness $f(x) \geq j$ is

$$\prod_{i=1}^{j} p_{t,i} \geq \prod_{i=1}^{j} q_i \geq \left(1 - \frac{1}{n}\right)^{j-1} \left(1 - \frac{1}{n} - \varepsilon\right) \geq \frac{1 - \varepsilon - o(1)}{e} = \Omega(1)$$

since $(1-1/n)^{j-1} \geq 1/e$ for any $n > 0$. Because we also have $p_{t,j+1} \geq 1/n$, the probability of sampling an offspring in levels $A_{\geq j+1}$ is at least $\Omega(1) \cdot (1/n) = \Omega(1/n)$. Thus, the condition (G1) holds with a value of $z_j = \Omega(1/n)$.

For the condition (G2), we assume further that $|P_t \cap A_{\geq j+1}| \geq \gamma\lambda$ for some value $\gamma \in (0, \gamma_0)$, and we are also required to show that the probability of sampling an offspring in levels $A_{\geq j+1}$ is at least $(1+\delta)\gamma$ for some small constant $\delta \in (0,1)$. Because the marginal $p_{t,j+1}$ can be lower bounded by $\gamma\lambda/\mu$, the above probability can be written as follows.

$$\prod_{i=1}^{j+1} p_{t,i} \geq p_{t,j+1} \cdot \prod_{i=1}^{j} p_{t,i} \geq \frac{\gamma\lambda}{\mu} \cdot \frac{1-\varepsilon-o(1)}{e} \geq (1+\delta)\gamma,$$

where by choosing

$$\frac{\mu}{\lambda} \leq \frac{1-\varepsilon-o(1)}{e(1+\delta)} = \frac{1-\varepsilon'}{e(1+\delta)}$$

for some constants $\delta, \varepsilon \in (0,1)$ and some other constant $\varepsilon' \in (0,1)$. Thus, the condition (G2) of the level-based theorem is verified.

The condition (G3) requires the offspring population size to satisfy

$$\lambda \geq \frac{4}{\gamma_0 \delta^2} \ln\left(\frac{128m}{\delta^2 \cdot \min_j\{z_j\}}\right),$$

which, by noting that $\gamma_0 = (\mu/\lambda)/((1-\delta)(1-p))$, is equivalent to

$$\mu \geq \frac{4(1-\delta)(1-p)}{\delta^2} \ln\left(\frac{128m}{\delta^2 \cdot \min_j\{z_j\}}\right),$$

which can be easily satisfied by choosing a sufficiently large constant c in $\mu \geq c \log n$.

Having verified the three conditions (G1), (G2) and (G3), and noting that $\ln(\delta\lambda/(4+\delta z_j)) < \ln(3\delta\lambda/2)$, the level-based theorem now guarantees an upper bound of

$$\mathcal{O}\big(n\lambda \log \lambda + n^2\big).$$

Note that, throughout the proof, we always assume the occurrence of the following two events in each iteration (see Lemma 25):

(1) The number of individuals in group 1 is at least μ w.p. $1 - e^{-\Omega(\mu)}$,
(2) The number of individuals in group 3 is $B \leq \varepsilon\mu$ for some small constant $\varepsilon \in (0,1)$ w.p. $1 - e^{-\Omega(\mu)}$.

By the union bound, either or all of these events happen in an iteration $t \in \mathbb{N}$ with probability at most $2n^{-2c+1} + e^{-\Omega(\mu)} + e^{-\Omega(\mu)} = n^{-c_2}$ for some constant $c_2 > 0$. The complementary event occurs with probability at least $1 - n^{-c_2}$. Following the same

line of argumentation as in [10, Theorem 8] (which has already been applied in the proof of Theorem 21), the overall expected runtime is $\mathcal{O}\left(n\lambda \log \lambda + n^2\right)$. □

We remark here that the exponential lower bound in Theorem 15 for the LeadingOnes function without noise also holds for the noisy LeadingOnes function. We are also interested in the runtime of the PBIL on the noisy LeadingOnes. The following theorem derives such a result.

Theorem 27 *Consider the prior noise model with constant parameter $p \in (0, 1)$. The PBIL with a parent population size $\mu \geq c \log n$ for some sufficiently large constant $c > 0$, a constant smoothing parameter $\rho \in (1/e, 1]$, and also a constant selective pressure satisfying*

$$\frac{\mu}{\lambda} \leq \left(\frac{\rho^{2+\frac{\rho e}{1-\rho}+\ln((1-p)(1+\varepsilon)^2)}}{e(1+\varepsilon)} \right)^{1/\ln(e\rho)} \tag{36}$$

for some constant $\varepsilon \in (0, 1)$ has an expected runtime of $\mathcal{O}\left(n^2 + n\lambda \log \lambda\right)$ on the LeadingOnes function.

Proof We assume that $\gamma_0 = \psi/((1-\delta)(1-p))$ for any constant $\delta \in (0, 1)$ and the selective pressure $\psi = \mu/\lambda$. We also partition the noisy population into four groups as in the proof of Theorem 26 and pessimistically assume that the PBIL uses all of the B individuals in group 3 and $\mu - B$ individuals chosen from group 1 when updating the model. For all $i \in [j]$, let Q_i be the number of individuals in group 3 which has 1s at bit positions 1 through j, except for one position i where it has a 0. By definition, we then have

$$\sum_{i=1}^{j} Q_i = B. \tag{37}$$

Similarly to the proof of Theorem 21, we shall show that the probability of sampling the first j bits correctly is at least a constant using a majorisation argument. Because noise only impacts the weight $\sum_{i=1}^{j} p_{t+1,i}$, we still define the vector q as in (21) and an integer $m = \lfloor g(j) \rfloor$ as in (22). We are left to show a constant upper bound on the difference $j - m$ used in Eq. 23. We notice that in this case the weight becomes

$$\begin{aligned}\sum_{i=1}^{j} p_{t+1,i} &= (1-\rho) \sum_{i=1}^{j} p_{t,i} + \frac{\rho}{\mu} \sum_{i=1}^{j} X_{i,t} \\ &\geq (1-\rho)(jp_0^{1/j}) + \frac{\rho}{\mu} \sum_{i=1}^{j} X_{i,t}, \quad \text{(by Eq. 24)}\end{aligned}$$

which by noting that $\sum_{i=1}^{j} X_{i,t} = j\mu - \sum_{i=1}^{j} Q_i = j\mu - B$ satisfies

$$\geq (1-\rho)jp_0^{1/j} + \frac{\rho}{\mu}(j\mu - B) = (1-\rho)jp_0^{1/j} + \rho j - \rho\frac{B}{\mu}. \tag{38}$$

Putting everything together, we then obtain

$$\begin{aligned} & j - mj + 1 - \frac{(1-\rho)jp_0^{1/j} - \rho B/\mu}{1-\rho} \quad \text{(by Eq. 22 \& Eq. 38)} \\ &= 1 + \frac{\rho B}{(1-\rho)\mu} + j - jp_0^{1/j} \\ &\leq 1 + \frac{\rho B}{(1-\rho)\mu} - \ln(p_0), \qquad \text{(by Eq. 25)} \end{aligned}$$

which by (33) that $B \leq \varepsilon\mu$ for some small constant $\varepsilon > 0$ w.p. $1 - e^{-\Omega(\mu)}$ satisfies

$$\leq 1 + \frac{\rho\varepsilon\mu}{(1-\rho)\mu} - \ln(p_0) = 1 + \frac{\rho\varepsilon}{1-\rho} - \ln(p_0) = \mathcal{O}(1). \tag{39}$$

Thus, the probability of sampling an offspring in levels $A_{\geq j}$ is at least a constant, which immediately results in a lower bound of $\Omega(1/n)$ on the probability of sampling the first $j+1$ bits correctly, confirming the condition (G1) of the level-based theorem.

For condition (G2), we use the lower bound $p_{t+1,j+1} \geq \rho\gamma\lambda/\mu = \rho\gamma/\psi$. Then, the probability of sampling an offspring in levels $A_{\geq j+1}$ is

$$\begin{aligned} \left(\prod_{i=1}^{j} p_{t+1,i}\right) \cdot p_{t+1,j+1} &\geq \frac{\gamma}{e\psi} \cdot \rho^{2+\frac{\rho\varepsilon}{1-\rho}-\ln(p_0)} \qquad \text{(by Eq. 26 \& Eq. 39)} \\ &\geq \frac{\gamma\rho^{2+\frac{\rho\varepsilon}{1-\rho}}}{e\psi} \cdot \frac{(1+\varepsilon)^{\ln\rho}}{\gamma_0^{\ln\rho}} \\ &\geq \frac{\gamma\rho^{2+\frac{\rho\varepsilon}{1-\rho}}}{e} \cdot \frac{(1-p)^{\ln\rho}(1+\varepsilon)^{2\ln\rho}}{(\psi)^{1+\ln\rho}} \\ &\geq (1+\varepsilon)\gamma, \end{aligned}$$

which always holds if we choose the selective pressure $\psi = \mu/\lambda$ such that

$$\frac{\mu}{\lambda} \leq \left(\frac{\rho^{2+\frac{\rho\varepsilon}{1-\rho}+\ln((1-p)(1+\varepsilon)^2)}}{e(1+\varepsilon)}\right)^{1/\ln(e\rho)}.$$

Similar to Eq. 27, we also require $\rho \in (1/e, 1]$. The condition (G2) is now verified.

For condition (G3), it suffices to use a population size $\lambda \geq c\log n$, for a sufficiently large constant $c > 0$. Having verified three conditions, Theorem 1 now guarantees an upper bound of $\mathcal{O}\big(n^2 + n\lambda\log\lambda\big)$. Note that throughout the proof we always assume the occurrence of the following three events:

(1) Each of the first j marginals is at least $p_0 \geq \gamma_0/(1+\varepsilon)$ w.p. $1 - 2n^{-2c}$ for any constants $c > 0$ and $\varepsilon > 0$, which requires a population of $\lambda \geq c((1+1/\varepsilon)/\gamma_0)^2\ln(n) = \Omega(\log n)$ (see Lemma 19),

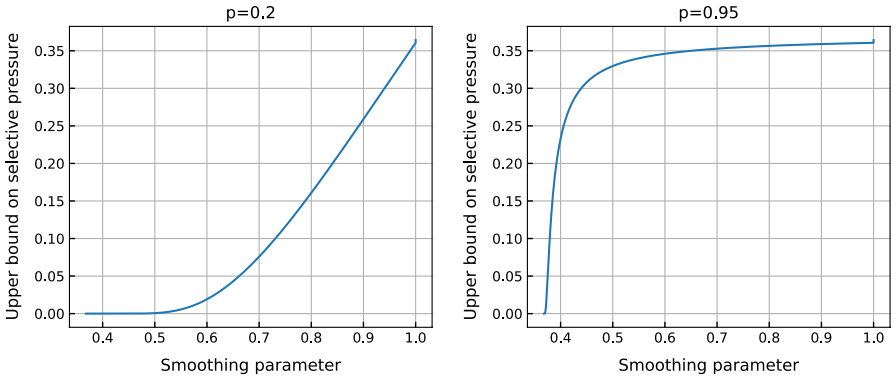

Fig. 2 Threshold on the selective pressure for the PBIL with $\rho \in (1/e, 1]$ on the noisy LeadingOnes function in Eq. 36 with $\varepsilon = 0.01$ for two noise probabilities $p = 0.2$ (left) and $p = 0.95$ (right). Note also that $1/e \approx 0.3679$ and $1/(e(1+\varepsilon)) \approx 0.3642$

(2) The number of individuals in group 1 (with actual fitness $f(x) \geq j$ and noisy fitness $F(x) \geq j$) is at least μ w.p. $1 - e^{-\Omega(\mu)}$ (see Lemma 25),
(3) The number of individuals in group 3 is $B \leq \varepsilon\mu$ for some small constant $\varepsilon > 0$ w.p. $1 - e^{-\Omega(\mu)}$.

By the union bound, either or all of these events happen in an iteration $t \in \mathbb{N}$ with probability at most $2n^{-2c+1} + e^{-\Omega(\mu)} + e^{-\Omega(\mu)} = n^{-c_2}$ for some constant $c_2 > 0$. The complementary event occurs with probability at least $1 - n^{-c_2}$. Following the same line of argumentation as in [10, Theorem 8] (which has already been applied in the proofs of Theorems 21 and 26), the overall expected runtime is $\mathcal{O}\big(n\lambda \log \lambda + n^2\big)$. □

Figure 2 plotted the threshold on the selective pressure in Eq. 36 for two noise probabilities $p = 0.2$ and $p = 0.95$.

6 Experiments

In this section, we provide an empirical study to see how closely the theoretical results match the experimental results for reasonable problem instance sizes, and to investigate a broader range of parameters. Our analysis is focused on different regimes on the selective pressure in the noise-free setting.

6.1 Under Low Selective Pressure

We have shown in Theorem 15 that when the selective pressure $\psi \geq (1+\delta)/e^{1-\varepsilon}$ for any constants $\delta > 0$ and $\varepsilon \in (0, 1)$, the UMDA requires an expected runtime of $e^{\Omega(\mu)}$ to optimise the LeadingOnes function. We now choose $\delta = 0.2$ and $\varepsilon = 0.1$, we then get $\psi \geq (1+0.2)/e^{1-0.1} \approx 0.4879$. Thus, the choice $\psi = 0.5$ should be sufficient to yield an exponential runtime. For the population size, we experiment with two

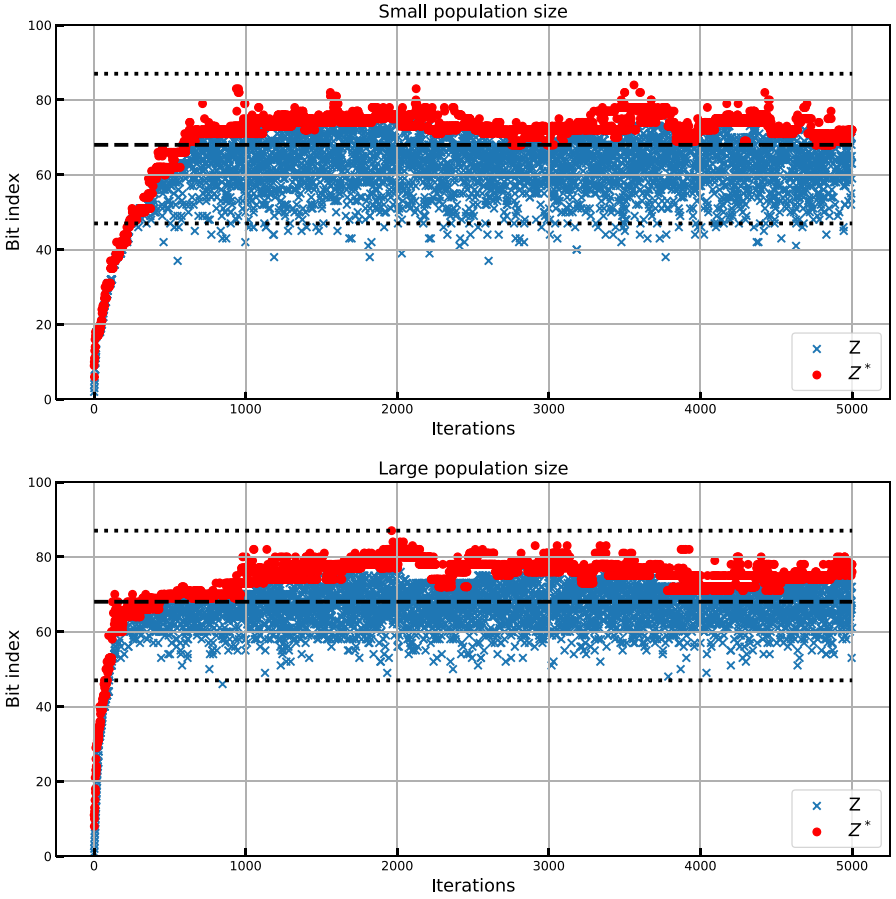

Fig. 3 The LeadingOnes-values of the fittest and the μ-th individuals, i.e., Z^* and Z, respectively, for the UMDA for $n = 100$ and $\mu/\lambda = 0.5$ over 5000 iterations. (Top) Small population size $\mu = 5 \log n$. (Bottom) Large population size $\mu = n$. The upper and lower dotted lines denote the bounds $\beta \approx 87$ in Eq. 12 and $\alpha \approx 47$ in Eq. 11, respectively, while the dashed line in the middle represents the value $\kappa \approx 69$ in Eq. 13

different settings: $\mu = 5 \log n$ (small) and $\mu = n$ (large) for a problem instance size $n = 100$. Substituting everything into (11) and (12), we then get $\alpha \approx 47$ and $\beta \approx 87$. The numbers of leading 1s of the fittest individual and the μ-th individual in the sorted population (denoted by random variables Z_t^* and Z_t respectively) are shown in Fig. 3 over an epoch of 5000 iterations. The dotted blue lines denote the constant functions of $\alpha = 47$ and $\beta = 87$. One can see that the Z_t-values keep increasing until it reaches the value of α during the early stage and always stays well under value β afterwards. Furthermore, Z_t^*-values do not deviate too far from Z_t that matches our analysis since the chance of sampling all ones from the $n - \beta$ last bits is exponentially small.

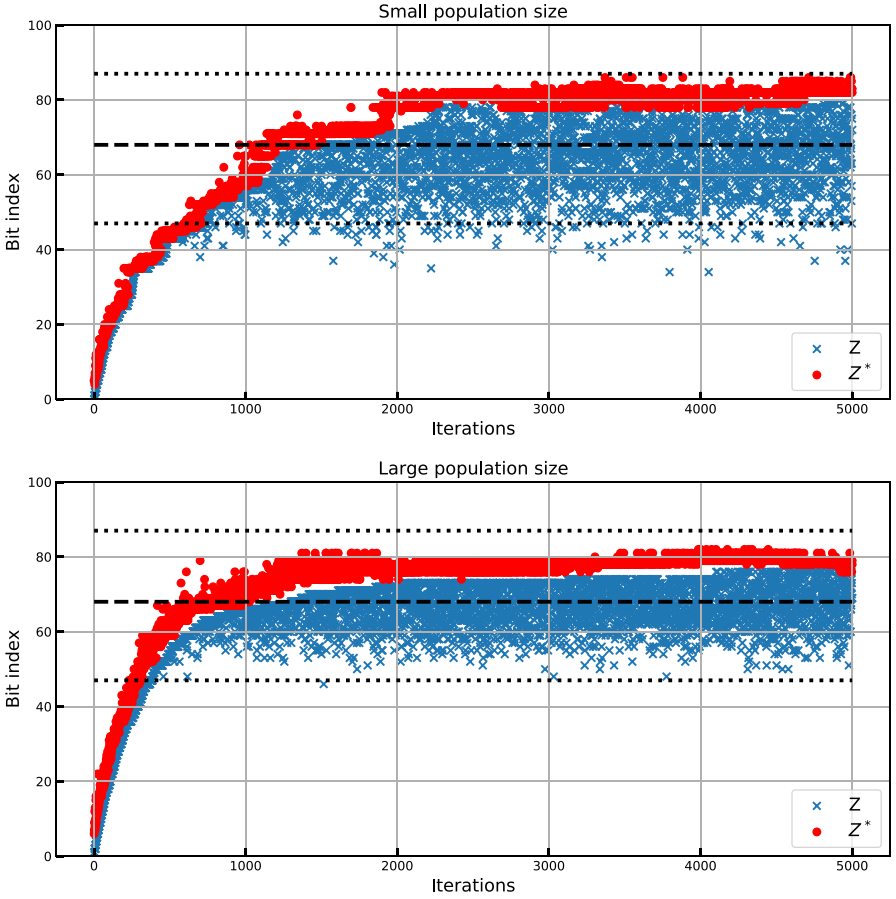

Fig. 4 The LeadingOnes-values of the fittest and the μth individuals, i.e., Z^* and Z, respectively, for the PBIL with $\rho = 0.5$ for $n = 100$ and $\mu/\lambda = 0.5$ over 5000 iterations. (Top) Small population size $\mu = 5 \log n$. (Bottom) Large population size $\mu = n$. The upper and lower dotted lines denote the values $\beta \approx 87$ in Eq. 12 and $\alpha \approx 47$ in Eq. 11, respectively, while the dashed line in the middle represents the value $\kappa \approx 69$ in Eq. 13

We also run the same experiments for the PBIL when we further choose a smoothing parameter of $\rho = 0.5 \in (1/e, 1]$. As predicted, one can see that two random variables Z_t and Z_t^* stay well below the threshold β (Fig. 4).

6.2 Under High Selective Pressure

When the selective pressure is sufficiently high, that is, $\psi \leq (1 - o(1))(1 - \delta)/e$ for any constant $\delta \in (0, 1)$, there is an upper bound $\mathcal{O}\left(n^2 + n\lambda \log \lambda\right)$ on the expected runtime [5]. Theorem 18 yields a lower bound of $\Omega(n\lambda/\log n)$. We start by looking at how the values of random variable Z_t and Z_t^* change over time. Our analysis shows that it never decreases during the whole optimisation course with overwhelming

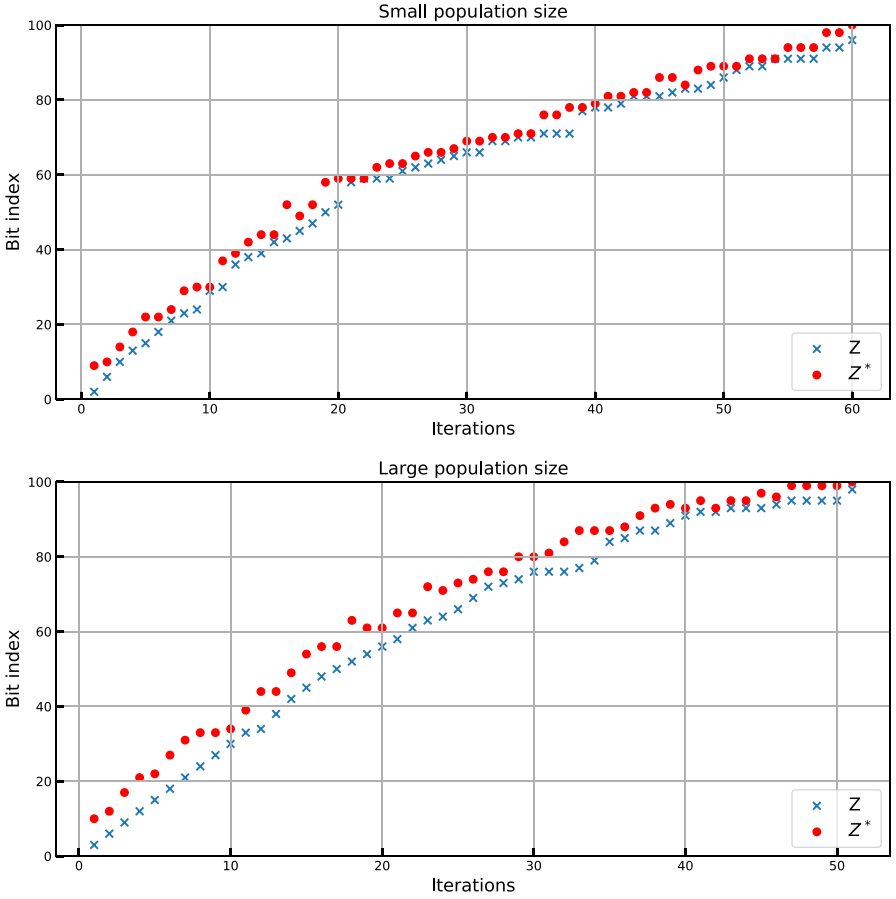

Fig. 5 The LeadingOnes-values of the fittest and the μth individuals, i.e., Z^* and Z, respectively, in one run of the UMDA with $n = 100$ and $\mu/\lambda = 0.1$. (Top) Small population size $\mu = 5 \log n$. (Bottom) Large population size $\mu = n$

probability and eventually reaches the value of n. Similarly, we consider the two different settings for population size and also note that our result holds for a parent population size $\mu \geq c \log n$, when the constant $c > 0$ must be tuned carefully; in this experiment, we set $c = 5$ (an integer larger than 3 should be sufficient). We then get $\psi \leq (1 - 1/100)(1 - 0.1)/e \approx 0.3278$. Therefore, the choice of $\psi = 0.1$ should be sufficient and we then get $\alpha \approx 160 > n = 100$. The experiment outcomes are shown in Fig. 5. The empirical result shows that both the Z- and Z^*-values keep increasing over the whole course of optimisation, matching our findings in Sect. 4.1. Furthermore, the difference between the Z- and Z^*-values in each iteration is relatively small, which again matches the result of Lemma 17.

7 Conclusion

In this paper, we have derived runtime results for population-based univariate EDAs (i.e., the UMDA and the PBIL) on the LeadingOnes function—a well-known test problem in the theory of evolutionary computation. For the UMDA, we have found that the algorithm under a low selective pressure requires an exponential expected runtime in the population size. More specifically, the algorithm takes an expected runtime of $2^{\Omega(\mu)}$ when $\mu \geq c \log n$ for a sufficiently large constant $c > 0$ and $\mu/\lambda \geq (1+\delta)/e^{1-\varepsilon}$ for any constant $\delta > 0$ and $\varepsilon \in (0,1)$. The analyses reveal the limitations of the probabilistic model based on probability vectors as the algorithm hardly stays at promising states for long enough to make progress. This leads the algorithm into a non-optimal equilibrium state from which the global optimum is exponentially unlikely to be sampled. On the other hand, when the selective pressure is high we obtain a lower bound of $\Omega(n\lambda/\log\lambda)$ on the expected runtime for the algorithm.

We then moved on to consider the PBIL on the LeadingOnes function. The algorithm is shown to optimise the function within an expected runtime of $\mathcal{O}(n^2)$ for appropriate parameter settings. Our findings here improve the currently best-known upper bound of $\mathcal{O}(n^{2+c})$ in [47] by a significant factor of $\Theta(n^c)$ for some constant $c \in (0,1)$.

Furthermore, we for the first time study the performances of the UMDA and the PBIL on the LeadingOnes function under a prior noise model, where a uniformly chosen bit is flipped with a constant probability $p \in (0,1)$ before invoking the fitness function. We show that an $\mathcal{O}(n^2)$ expected runtime still holds in this case for both algorithms under an offspring population size $\lambda = \Omega(\log n) \cap \mathcal{O}(n/\log n)$. Despite the simplicity of the noise model, this can be viewed as the first step towards broadening our understanding of the two algorithms' behaviours in a noisy environment.

The UMDA with an offspring population size $\lambda = \Omega(\log n) \cap \mathcal{O}(n/\log n)$ needs an $\mathcal{O}(n^2)$ expected time on the LeadingOnes function [5]. In this case, Theorem 18 yields a lower bound $\Omega(n^2/\log^2 n)$. Thus, it remains open whether this gap of $\Theta(\log^2 n)$ could be closed to achieve a tight bound on the runtime. Note that our result in Theorem 15, together with Theorem 16, provide upper bounds on the expected runtime of the UMDA on the LeadingOnes function when the selective pressure is low and high (around the threshold value of $1/e$). Although we could choose the constant small/large enough such that the selective pressure becomes arbitrarily close to $1/e$, it is still unknown whether the UMDA will take a polynomial or exponential expected runtime when the selective pressure is exactly $1/e$. Another avenue for future work would be to investigate the PBIL with a smoothing parameter $\rho \in (0, 1/e)$. Our analysis does not cover this regime of the smoothing parameter.

Additional Results

In the following variant of Markov's inequality, we use the notation $x \vee y := \max(x, y)$.

Lemma 28 *Assume any random variable $\tau \in \mathbb{N}$ and a sequence of events $F_0, F_1, \ldots$ such that for all $s, i \in \mathbb{N}$, the event $\tau \leq s$ is independent of the event F_{s+i}. Define for all $t \in \mathbb{N}$ the event $G_t := \bigwedge_{i=0}^{t}(\neg F_i)$. For any $t \in \mathbb{R}$, $s \in \mathbb{N}$ with $s \geq t$, if $\Pr\left(G_{\tau \vee s}\right) > 0$ and*

$$\mathbb{E}\left[\tau \mid G_{\tau \vee s}\right] \leq t \tag{40}$$

then $\Pr(\tau > s) \leq 1 - \Pr\left(G_s\right)(1 - t/s)$.

Proof By the law of total probability, we have

$$\begin{aligned}
\Pr(\tau \leq s) &\geq \Pr\left(G_s\right) \Pr\left(\tau \leq s \mid G_s\right) \\
\text{by independence of } \tau \leq s \text{ and } F_{s+i} \text{ for all } i \in \mathbb{N} &= \Pr\left(G_s\right) \Pr\left(\tau \leq s \mid G_{\tau \vee s}\right) \\
&= \Pr\left(G_s\right)\left(1 - \Pr\left(\tau > s \mid G_{\tau \vee s}\right)\right) \\
\text{by Markov's inequality and 40)} &\geq \Pr\left(G_s\right)\left(1 - \frac{\mathbb{E}\left[\tau \mid G_{\tau \vee s}\right]}{s}\right) \\
&\geq \Pr\left(G_s\right)\left(1 - \frac{t}{s}\right).
\end{aligned}$$

□

Acknowledgements Lehre was supported by a Turing AI Fellowship (EPSRC Grant ref EP/V025562/1).

References

1. Baluja, S.: Population-based incremental learning: a method for integrating genetic search based function optimization and competitive learning. Technical report, Carnegie Mellon University (1994)
2. Bosman, P.A.N., Thierens, D.: The balance between proximity and diversity in multiobjective evolutionary algorithms. IEEE Trans. Evol. Comput. **7**(2), 174–188 (2003)
3. Corus, D., Dang, D.C., Eremeev, A.V., Lehre, P.K.: Level-based analysis of genetic algorithms and other search processes. IEEE Trans. Evol. Comput. **22**(5), 707–719 (2018)
4. Dang, D.C., Lehre, P.K.: Efficient optimisation of noisy fitness functions with population-based evolutionary algorithms. In: Proceedings of the Conference on Foundations of Genetic Algorithms, FOGA '15, pp. 62–68 (2015)
5. Dang, D.C., Lehre, P.K.: Simplified runtime analysis of estimation of distribution algorithms. In: Proceedings of the Genetic and Evolutionary Computation Conference, GECCO '15, pp. 513–518 (2015)

6. Dang, D.C., Lehre, P.K., Nguyen, P.T.H.: Level-based analysis of the univariate marginal distribution algorithm. Algorithmica **81**, 668–702 (2018)
7. Doerr, B. Probabilistic tools for the analysis of randomized optimization heuristics. CoRR. abs/1801.06733 (2018)
8. Doerr , B., Kötzing, T. Multiplicative up-drift. In: Proceedings of the Genetic and Evolutionary Computation Conference, GECCO '19, pp. 1470–1478 (2019)
9. Doerr, B., Zheng, W.: Sharp bounds for genetic drift in estimation of distribution algorithms. IEEE Trans. Evol. Comput. **24**, 1 (2020)
10. Droste, S.: A rigorous analysis of the compact genetic algorithm for linear functions. Nat. Comput. **5**(3), 257–283 (2006)
11. Droste, S., Jansen, T., Wegener, I.: On the analysis of the (1 + 1) evolutionary algorithm. Theor. Comput. Sci. **276**(1–2), 51–81 (2002)
12. Dubhashi, D., Panconesi, A.: Concentration of Measure for the Analysis of Randomized Algorithms, 1st edn. Cambridge University Press, Cambridge (2009)
13. Eiben, A.E., Smith, J.E.: Introduction to Evolutionary Computing. Springer, Berlin (2003)
14. Feller, W.: An Introduction to Probability Theory and Its Applications, vol. 1, 3rd edn. Wiley, New York (1968)
15. Friedrich, T., Kötzing, T., Krejca, M.S.: EDAs cannot be balanced and stable. In: Proceedings of the Genetic and Evolutionary Computation Conference, GECCO '16, pp. 1139–1146 (2016)
16. Friedrich, T., Kötzing, T., Krejca, M.S., Sutton, A.M.: The compact genetic algorithm is efficient under extreme gaussian noise. IEEE Trans. Evol. Comput. **21**(3), 477–490 (2017)
17. Gießen, C., Kötzing, T.: Robustness of populations in stochastic environments. Algorithmica **75**(3), 462–489 (2016)
18. Gleser, L.J.: On the distribution of the number of successes in independent trials. Ann. Probab. **3**(1), 182–188 (1975)
19. Harik, G.R., Lobo, F.G., Goldberg, D.E.: The compact genetic algorithm. IlliGAL report No. 97006. University of Illinois at Urbana-Champaign (1997)
20. Hauschild, M., Pelikan, M.: An introduction and survey of estimation of distribution algorithms. Swarm Evol. Comput. **1**(3), 111–128 (2011)
21. He, J.: Towards an analytic framework for analysing the computation time of evolutionary algorithms. Artif. Intell. **145**(1), 59–97 (2003)
22. Krejca M.S., Carsten, W.: Theory of estimation-of-distribution algorithms. CoRR, arXiv:1806.05392 (2018)
23. Larrañaga, P., Karshenas, H., Bielza, C., Santana, R.: A review on probabilistic graphical models in evolutionary computation. J. Heuristics **18**(5), 795–819 (2012)
24. Larrañaga, P., Lozano, J.A.: Estimation of Distribution Algorithms: A New Tool for Evolutionary Computation. Genetic Algorithms and Evolutionary Computation, Springer, New York (2001)
25. Lehre, P.K.: Fitness-levels for non-elitist populations. In: Proceedings of the Genetic and Evolutionary Computation Conference, GECCO '11, pp. 2075–2082 (2011)
26. Lehre, P.K., Nguyen, P.T.H.: Improved runtime bounds for the univariate marginal distribution algorithm via anti-concentration. In: Proceedings of the Genetic and Evolutionary Computation Conference, GECCO '17, pp. 1383–1390 (2017)
27. Lehre, P.K., Nguyen, P.T.H.: Level-based analysis of the population-based incremental learning algorithm. In: Proceedings of the Conference on Parallel Problem Solving from Nature. PPSN XV, pp. 105–116 (2018)
28. Lehre, P.K., Nguyen, P.T.H.: On the limitations of the univariate marginal distribution algorithm to deception and where bivariate EDAs might help. In: Proceedings of the Conference on Foundations of Genetic Algorithms, FOGA '19, pp. 154–168 (2019)
29. Lehre, P.K., Nguyen, P.T.H.: Runtime analysis of the univariate marginal distribution algorithm under low selective pressure and prior noise. In: Proceedings of the Genetic and Evolutionary Computation Conference, GECCO '19. pp. 1497–1505 (2019)
30. Lehre, P.K., Oliveto, P.S.: Theoretical Analysis of Stochastic Search Algorithms, pp. 1–36. Springer, Berlin (2018)
31. Lehre, P.K., Witt, C.: Black-box search by unbiased variation. Algorithmica **64**(4), 623–642 (2012)
32. Marshall, A.W., Olkin, I., Arnold, B.C.: Inequalities: Theory of Majorization and Its Applications. Springer Series in Statistics, Springer, New York (2011)
33. Massart, P.: The tight constant in the Dvoretzky–Kiefer–Wolfowitz inequality. Ann. Probab. **18**(3), 1269–1283 (1990)

34. Mitrinović, D.S.: Analytic Inequalities. Springer, Berlin (1970)
35. Mitzenmacher, M., Upfal, E.: Probability and Computing: Randomized Algorithms and Probabilistic Analysis. Cambridge University Press, New York (2005)
36. Motwani, R., Raghavan, P.: Randomised Algorithms. Cambridge University Press, Cambridge (1995)
37. Mühlenbein, H., Mahnig, T.: Convergence theory and applications of the factorized distribution algorithm. CIT J. Comput. Inform. Technol. **7**(1), 19–32 (1999)
38. Mühlenbein, H., Paaß, G.: From recombination of genes to the estimation of distributions I. binary parameters. In: Proceedings of the Conference on Parallel Problem Solving from NaturE. PPSN IV, pp. 178–187 (1996)
39. Natural logarithm: Inequalities—wolfram functions site. https://functions.wolfram.com/ElementaryFunctions/Log/29/. Accessed 09 Nov 2020
40. Pelikan, M., Goldberg, D.E., Lobo, F.G.: A survey of optimization by building and using probabilistic models. Comput. Optim. Appl. **21**(1), 5–20 (2002)
41. Qian, C., Bian, C., Jiang, W., Tang, K.: Running time analysis of the (1+1)-EA for Onemax and Leadingones under bit-wise noise. In: Proceedings of the Conference on Genetic and Evolutionary Computation, GECCO '17, pp. 1399-1406 (2017)
42. Stützle, T., Hoos, H.H.: Max–Min ant system. Fut. Gen. Comput. Syst. **16**(8), 889–914 (2000)
43. Sudholt, D.: On the robustness of evolutionary algorithms to noise: refined results and an example where noise helps. In: Proceedings of the Conference on Genetic and Evolutionary Computation, GECCO '18, pp. 1523-1530 (2018)
44. Williams, D.: Probability with Martingales. Cambridge University Press, Cambridge (1991)
45. Witt, C.: Upper bounds on the runtime of the univariate marginal distribution algorithm on Onemax. In: Proceedings of the Genetic and Evolutionary Computation Conference, GECCO '17, pp. 1415–1422 (2017)
46. Witt, C.: Domino convergence: why one should hill-climb on linear functions. In: Proceedings of the Genetic and Evolutionary Computation Conference, GECCO '18, pp. 1539–1546 (2018)
47. Wu, Z., Kolonko, M., Möhring, R.H.: Stochastic runtime analysis of a cross entropy algorithm. IEEE Trans. Evol. Comput. **21**(4), 616–628 (2017)

Authors and Affiliations

Per Kristian Lehre[1] · Phan Trung Hai Nguyen[1,2]

✉ Per Kristian Lehre
p.k.lehre@cs.bham.ac.uk

✉ Phan Trung Hai Nguyen
p.nguyen@exeter.ac.uk

[1] School of Computer Science, University of Birmingham, Birmingham B15 2TT, UK

[2] Present Address: Department of Computer Science, University of Exeter, Exeter EX4 4QJ, UK