Level-Based Analysis of the Population-Based Incremental Learning Algorithm

Per Kristian Lehre and Phan Trung Hai Nguyen

University of Birmingham, Birmingham, United Kingdom
{p.k.lehre, p.nguyen}@cs.bham.ac.uk

Abstract. The Population-Based Incremental Learning (PBIL) algorithm uses a convex combination of the current model and the empirical model to construct the next model, which is then sampled to generate offspring. The Univariate Marginal Distribution Algorithm (UMDA) is a special case of the PBIL, where the current model is ignored. Dang and Lehre (GECCO 2015) showed that UMDA can optimise LeadingOnes efficiently. The question still remained open if the PBIL performs equally well. Here, by applying the level-based theorem in addition to Dvoretzky–Kiefer–Wolfowitz inequality, we show that the PBIL optimises function LeadingOnes in expected time $\mathcal{O}\left(n\lambda\log\lambda + n^2\right)$ for a population size $\lambda = \Omega(\log n)$, which matches the bound of the UMDA. Finally, we show that the result carries over to BinVal, giving the fist runtime result for the PBIL on the BinVal problem.

Keywords: Population-based incremental learning · LeadingOnes · BinVal · Running time analysis · Level-based analysis · Theory

1 Introduction

Estimation of distribution algorithms (EDAs) are a class of randomised search heuristics that optimise objective functions by constructing probabilistic models and then sample the models to generate offspring for the next generation. Various variants of EDA have been proposed over the last decades; they differ from each other in the way their models are represented, updated as well as sampled over generations. In general, EDAs are usually categorised into two main classes: *univariate* and *multivariate*. Univariate EDAs take advantage of first-order statistics (i.e. mean) to build a univariate model, whereas multivariate EDAs apply higher-order statistics to model the correlations between the decision variables.

There are only a few runtime results available for EDAs. Recently, there has been a growing interest in the optimisation time of the UMDA, introduced by Mühlenbein and Paaß [11], on standard benchmark functions [4, 13, 8, 7, 14]. Recall that the optimisation time of an algorithm is the number of fitness evaluations the algorithm needs before a global optimum is sampled for the first time. Dang and Lehre [4] analysed a variant of the UMDA using truncation selection and derived the first upper bounds of $\mathcal{O}(n\lambda\log\lambda)$ and $\mathcal{O}\left(n\lambda\log\lambda + n^2\right)$ on the expected optimisation times of the UMDA on OneMax and LeadingOnes, respectively, where the population size is $\lambda = \Omega(\log n)$. These results were obtained using a relatively new technique called *level-based analysis* [3]. Very recently, Witt [13] proved that the UMDA optimises OneMax within $\mathcal{O}(\mu n)$ and $\mathcal{O}(\mu\sqrt{n})$ when $\mu \geq c \log n$ and $\mu \geq c'\sqrt{n}\log n$ for some constants $c,\ c' > 0$, respectively. However, these bounds only hold when $\lambda = (1+\Theta(1))\mu$. This constraint on λ and μ was relaxed by Lehre and Nguyen [8], where the upper bound $\mathcal{O}(\lambda n)$ holds for $\lambda = \Omega(\mu)$ and $c \log n \leq \mu = \mathcal{O}(\sqrt{n})$ for some constant $c > 0$.

The first rigorous runtime analysis of the PBIL [1], was presented very recently by Wu et al. [14]. In this work, the PBIL was referred to as a cross entropy algorithm. The study proved an upper bound $\mathcal{O}\left(n^{2+\varepsilon}\right)$ of the PBIL with margins $[1/n, 1-1/n]$ on LeadingOnes, where $\lambda = n^{1+\varepsilon}$, $\mu = \mathcal{O}(n^{\varepsilon/2})$, $\eta \in \Omega(1)$ and $\varepsilon \in (0,1)$. Until now, the known runtime bounds for the PBIL were significantly higher than those for the UMDA. Thus, it is of interest to determine whether the PBIL is less efficient than the UMDA, or whether the bounds derived in the early works were too loose.

This paper makes two contributions. First, we address the question above by deriving a tighter bound $\mathcal{O}\left(n\lambda\log\lambda + n^2\right)$ on the expected optimisation time of the PBIL on LeadingOnes. The bound holds for population sizes $\lambda = \Omega(\log n)$, which is a much weaker assumption than $\lambda = \omega(n)$ as required in [14]. Our proof is more straightforward than that in [14] because much of the complexities of the analysis are already handled by the level-based method [3].

The second contribution is the first runtime bound of the PBIL on BinVal. This function was shown to be the hardest among all linear functions for the cGA [5]. The result carries easily over from the level-based analysis of LeadingOnes using an identical partitioning of the search space. This observation further shows that runtime bounds, derived by the level-based method using the canonical partition, of the PBIL or other non-elitist population-based algorithms using truncation selection, on LeadingOnes also hold for BinVal.

The paper is structured as follows. Section 2 introduces the PBIL with margins as well as the level-based theorem, which is the main method employed in the paper. Given all necessary tools, the next two sections then provide upper bounds on the expected optimisation time of the PBIL on LeadingOnes and BinVal. Finally, our concluding remarks are given in Section 5.

2 Preliminaries

We first introduce the notations used throughout the paper. Let $\mathcal{X} := \{0,1\}^n$ be a finite binary search space with dimension n. The univariate model in generation $t \in \mathbb{N}$ is represented by a vector $p^{(t)} := (p_1^{(t)}, \ldots, p_n^{(t)}) \in [0,1]^n$, where each $p_i^{(t)}$ is called a *marginal*. Let $X_1^{(t)}, \ldots, X_n^{(t)}$ be n independent Bernoulli random variables with success probabilities $p_1^{(t)}, \ldots, p_n^{(t)}$. Furthermore, let $X_{i:j}^{(t)} := \sum_{k=i}^{j} X_k^{(t)}$ be the number of ones sampled from $p_{i:j}^{(t)} := (p_i^{(t)}, \ldots, p_j^{(t)})$ for all $1 \leq i \leq j \leq n$. Each individual (or bitstring) is denoted as $x = (x_1, \ldots, x_n) \in \mathcal{X}$. We aim at maximising an objective function $f : \mathcal{X} \to \mathbb{R}$. We are primarily interested in the optimisation time of these algorithms, so tools to analyse runtime are of importance. We will make use of the level-based theorem [3].

2.1 Two problems

We consider the two pseudo-Boolean functions: LeadingOnes and BinVal, which are widely used theoretical benchmark problems in runtime analyses of EDAs [5, 4, 14]. The former aims at maximising the number of leading ones, while the latter tries to maximise the binary value of the bitstring. The global optimum for both functions are the all-ones bitstring. Furthermore, BinVal is an extreme linear function, where the fitness-contribution of the bits decreases exponentially with the bit-position. Droste [5] showed that among all linear functions, BinVal is difficult for the cGA. Given a bitstring $x = (x_1, \ldots, x_n) \in \mathcal{X}$, the two functions are formally defined as follows:

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

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

2.2 Population-Based Incremental Learning

The PBIL algorithm maintains a univariate model over generations. The probability of a bitstring $x = (x_1, \ldots, x_n)$ sampled from the current model $p^{(t)}$ is given by

$$\Pr\left(x \mid p^{(t)}\right) = \prod_{i=1}^{n} \left(p_i^{(t)}\right)^{x_i} \left(1 - p_i^{(t)}\right)^{1-x_i}. \tag{1}$$

Let $p^{(0)} := (1/2, \ldots, 1/2)$ be the initial model. The algorithm in generation t samples a population of λ individuals, denoted as $P^{(t)} := \{x^{(1)}, x^{(2)}, \ldots, x^{(\lambda)}\}$, which are sorted in descending order according to fitness. The μ fittest individuals are then selected to derive the next model $p^{(t+1)}$ using the component-wise formula $p_i^{(t+1)} := (1-\eta)\, p_i^{(t)} + (\eta/\mu) \sum_{j=1}^{\mu} x_i^{(j)}$ for all $i \in \{1, 2, \ldots, n\}$, where $x_i^{(j)}$ is the i-th bit of the j-th individual in the sorted population, and $\eta \in (0, 1]$ is the smoothing parameter (sometimes known as the learning rate). The ratio $\gamma_0 := \mu/\lambda \in (0, 1)$ is called the selective pressure of the algorithm. Univariate EDAs often employ margins to avoid the marginals to fix at either 0 or 1. In particular, the marginals are usually restricted to the interval $[1/n, 1 - 1/n]$ after being updated, where the quantities $1/n$ and $1 - 1/n$ are called the lower and upper borders, respectively. The algorithm is called the PBIL with margins. Algorithm 1 gives a full description of the PBIL (with margins).

2.3 Level-based analysis

Introduced in [3], the level-based theorem is a general tool that provides upper bounds on the expected optimisation time of many non-elitist population-based

Algorithm 1: PBIL with margins

$t \leftarrow 0$; $p^{(t)} \leftarrow (1/2, 1/2, \ldots, 1/2)$
repeat
  for $j = 1, 2, \ldots, \lambda$ **do**
    sample an offspring $x^{(j)} \sim \Pr(\cdot \mid p^{(t)})$ as defined in (1)
    evaluate the fitness $f(x^{(j)})$
  sort $P^{(t)} \leftarrow \{x^{(1)}, x^{(2)}, \ldots, x^{(\lambda)}\}$ such that $f(x^{(1)}) \geq f(x^{(2)}) \geq \ldots \geq f(x^{(\lambda)})$
  for $i = 1, 2, \ldots, n$ **do**
    $p_i^{(t+1)} \leftarrow \max\left\{1/n, \min\left\{1 - 1/n, (1-\eta)\, p_i^{(t)} + (\eta/\mu) \sum_{j=1}^{\mu} x_i^{(j)}\right\}\right\}$
  $t \leftarrow t + 1$
until *termination condition is fulfilled*

Algorithm 2: Non-elitist population-based algorithm

$t \leftarrow 0$; create initial population $P^{(t)}$
repeat
  for $i = 1, \ldots, \lambda$ **do**
    sample $P_i^{(t+1)} \sim \mathcal{D}(P^{(t)})$
  $t \leftarrow t + 1$
until *termination condition is fulfilled*

algorithms on a wide range of optimisation problems [3, 8, 4]. The theorem assumes that the algorithm to be analysed can be described in the form of Algorithm 2, which maintains a population $P^{(t)} \in \mathcal{X}^\lambda$, where $\mathcal{X}^\lambda$ is the space of all populations with size λ. The theorem is general since it never assumes specific fitness functions, selection mechanisms, or generic operators like mutation and crossover. Furthermore, the theorem 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. The theorem is formally stated in Theorem 1 [3]. We will use the notation $[n] := \{1, 2, \ldots, n\}$ and $A_{\geq j} := \cup_{k=j}^{m} A_k$.

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 generation t. Denote $y \sim \mathcal{D}(P^{(t)})$. 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 $\Pr(y \in A_{\geq j+1}) \geq z_j$.*

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

(G3) *and the population size* $\lambda \in \mathbb{N}$ *satisfies* $\lambda \geq \left(\frac{4}{\gamma_0 \delta^2}\right) \ln\left(\frac{128m}{z_* \delta^2}\right)$ *where* $z_* := \min_{j \in [m-1]}\{z_j\}$*, then*

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

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 unnecessarily always the case, especially for the PBIL with a sufficiently large offspring population size λ. The rationale behind this is that in each generation the PBIL draws λ samples from the current model $p^{(t)}$, that correspond to λ individuals in the current population, and if the number of samples λ is sufficiently large, it is highly likely that the empirical distributions for all positions among the entire population cannot deviate too far from the true distributions, i.e. marginals $p_i^{(t)}$. Moreover, the theorem relies on three conditions (G1), (G2) and (G3); thus, as long as these three can be fully verified, the PBIL, whose model is constructed from the current population $P^{(t)}$ in addition to the current model $p^{(t)}$, is still eligible to the level-based analysis.

2.4 Other tools

In addition to the level-based theorem, we also make use of some other mathematical results. First of all is the Dvoretzky–Kiefer–Wolfowitz inequality [9], which provides an estimate on how close an empirical distribution function will be to the true distribution from which the samples are drawn. The following theorem follows by replacing $\varepsilon = \varepsilon' \sqrt{\lambda}$ into [9, 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* $\hat{F}_\lambda$ *be the empirical distribution function which is defined by* $\hat{F}_\lambda(x) := (1/\lambda) \sum_{i=1}^{\lambda} \mathbb{1}_{\{X_i \,\leq\, x\}}$*. For any* $\lambda \in \mathbb{N}$ *and* $\varepsilon > 0$*, we always have*

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

Furthermore, properties of majorisation between two vectors are also exploited. The concept is formally defined in Definition 3 [6], followed by its important property (in Lemma 1) that we use intensively throughout the paper.

Definition 3. *Given vectors* $p^{(1)} := (p_1^{(1)}, \ldots, p_n^{(1)})$ *and* $p^{(2)} := (p_1^{(2)}, \ldots, p_n^{(2)})$*, where* $p_1^{(1)} \geq p_2^{(1)} \geq \ldots \geq p_n^{(1)}$ *and similarly for the* $p_i^{(2)}$*s. Vector* $p^{(1)}$ *is said to majorise vector* $p^{(2)}$*, in symbols* $p^{(1)} \succ p^{(2)}$*, if* $p_1^{(1)} \geq p_1^{(2)}, \ldots, \sum_{i=1}^{n-1} p_i^{(1)} \geq \sum_{i=1}^{n-1} p_i^{(2)}$ *and* $\sum_{i=1}^{n} p_i^{(1)} = \sum_{i=1}^{n} p_i^{(2)}$.

Lemma 1 ([2]). *Let $X_1, \dots, X_n$ be n independent Bernoulli random variables with success probabilities $p_1, \dots, p_n$, respectively. Denote $p := (p_1, p_2, \dots, p_n)$; let $S(p) := \sum_{i=1}^n X_i$ and $D_\lambda := \{p : p_i \in [0,1],\ i \in [n],\ \sum_{i=1}^n p_i = \lambda\}$. For two vectors $p^{(1)}, p^{(2)} \in D_\lambda$, if $p^{(1)} \prec p^{(2)}$ then $\Pr\left(S(p^{(1)}) = n\right) \geq \Pr\left(S(p^{(2)}) = n\right)$.*

Lemma 2 (Main lemma). *Let $p^{(1)}$ and $p^{(2)} \in D_\lambda$ be two vectors as defined in Lemma 1, where all components in $p^{(\cdot)}$ are arranged in descending order. Let $z^{(1)} := (z_1^{(1)}, \dots, z_n^{(1)})$ where each $z_i^{(1)} := (1-\eta)\, p_i^{(1)} + \eta$, and $z^{(2)} := (z_1^{(2)}, \dots, z_n^{(2)})$, where each $z_i^{(2)} := (1-\eta)\, p_i^{(2)} + \eta$ for any constant $\eta \in (0,1]$. If $p^{(2)} \succ p^{(1)}$, then $z^{(2)} \succ z^{(1)}$.*

Proof. For all $j \in [n-1]$, it holds that $\sum_{i=1}^j z_i^{(2)} \geq \sum_{i=1}^j z_i^{(1)}$ since $\sum_{i=1}^j p_i^{(2)} \geq \sum_{i=1}^j p_i^{(1)}$. Furthermore, if $j = n$, then $\sum_{i=1}^n z_i^{(2)} = \sum_{i=1}^n z_i^{(1)}$ due to $\sum_{i=1}^n p_i^{(2)} = \sum_{i=1}^n p_i^{(1)}$. By Definition 3, $z^{(2)} \succ z^{(1)}$. □

3 Runtime Analysis of the PBIL on LeadingOnes

We now show how to apply the level-based theorem to analyse the runtime of the PBIL. We use a *canonical partition* of the search space, where each subset A_j contains bitstrings with exactly j leading ones.

$$A_j := \{x \in \{0,1\}^n \mid \textsc{LeadingOnes}(x) = j\}. \tag{2}$$

Conditions (G1) and (G2) of Theorem 1 assume that there are at least $\gamma_0\lambda$ individuals in levels $A_{\geq j}$ in generation t. Recall $\gamma_0 := \mu/\lambda$. This implies that the first j bits among the μ fittest individuals are all ones. Denote $\hat{p}_i^{(t)} := (1/\lambda) \sum_{j=1}^\lambda x_i^{(j)}$ as the frequencies of ones at position i in the current population. We first show that under the assumption of the two conditions of Theorem 1 and with a population size $\lambda = \Omega(\log n)$, the first j marginals cannot be too close to the lower border $1/n$ with probability at least $1 - n^{-\Omega(1)}$.

Lemma 3. *If $|P^{(t)} \cap A_{\geq j}| \geq \gamma_0\lambda$ 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 it holds with probability at least $1 - 2n^{-2c}$ that $p_i^{(t)} \geq \gamma_0/(1+\varepsilon)$ for all $i \in [j]$.*

Proof. Consider an arbitrary bit $i \in [j]$. Let Q_i be the number of ones sampled at position i in the current population, and the corresponding empirical distribution function of the number of zeros is $F_\lambda(0) = (1/\lambda) \sum_{j=1}^\lambda \mathbb{1}_{\{x_i^{(j)} \leq 0\}} = (\lambda - Q_i)/\lambda = 1 - \hat{p}_i^{(t)}$, and the true distribution function is $F(0) = 1 - p_i^{(t)}$. The DKW inequality (see Theorem 2) yields that $\Pr(\hat{p}_i^{(t)} - p_i^{(t)} > \phi) \leq \Pr(|\hat{p}_i^{(t)} - p_i^{(t)}| > \phi) \leq 2e^{-2\lambda\phi^2}$ for all $\phi > 0$. Therefore, with probability at least $1 - 2e^{-2\lambda\phi^2}$ we have $\hat{p}_i^{(t)} - p_i^{(t)} \leq \phi$ and, thus, $p_i^{(t)} \geq \hat{p}_i^{(t)} - \phi \geq \gamma_0 - \phi$ since $\hat{p}_i^{(t)} \geq \gamma_0\lambda/\lambda = \gamma_0$ due to $|P^{(t)} \cap A_{\geq j}| \geq \gamma_0\lambda$. We then choose $\phi \leq \varepsilon\gamma_0/(1+\varepsilon)$ for some constant $\varepsilon > 0$ and $\lambda \geq c((1+1/\varepsilon)/\gamma_0)^2 \ln(n)$. Putting everything together, it holds that $p_i^{(t)} \geq \gamma_0(1 - \varepsilon/(1+\varepsilon)) = \gamma_0/(1+\varepsilon)$ with probability at least $1 - 2n^{-2c}$. □

Given the μ top individuals having at least j leading ones, we now estimate the probability of sampling j leading ones from the current model $p^{(t)}$.

Lemma 4. *For any non-empty subset $I \subseteq [n]$, define $C_I := \{x \in \{0,1\}^n \mid \prod_{i\in I} x_i = 1\}$. If $|P^{(t)} \cap C_I| \geq \gamma_0\lambda$ and $\lambda \geq c((1+1/\varepsilon)/\gamma_0)^2 \ln(n)$ for any constants $\varepsilon > 0$, $\gamma_0 \in (0,1)$, then it holds with probability at least $1-2n^{-2c}$ that $q^{(t)} := \prod_{i\in I} p_i^{(t)} \geq \gamma_0/(1+\varepsilon)$.*

Proof. We prove the statement using the DKW inequality (see Theorem 2). Let $m = |I|$. Given an offspring sample $Y \sim p^{(t)}$ from the current model, let $Y_I := \sum_{i\in I} Y_i$ be the number of one-bits in bit-positions I. By the assumption $|P^{(t)} \cap C_I| \geq \gamma_0\lambda$ on the current population, the empirical distribution function of Y_I must satisfy $\hat{F}_\lambda(m-1) = \frac{1}{\lambda}\sum_{i=1}^{\lambda} \mathbb{1}_{\{Y_{I,i} \leq m-1\}} \leq 1-\hat{q}^{(t)}$, where $\hat{q}^{(t)} \geq \gamma_0$ is the fraction of individuals in the current population with j leading ones, and the true distribution function satisfies $F(m-1) = 1-q^{(t)}$. 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 $q^{(t)} \geq \gamma_0(1-\varepsilon/(1+\varepsilon)) = \gamma_0/(1+\varepsilon)$ with probability at least $1-2e^{-2\phi^2\lambda} \geq 1-2n^{-2c}$. □

Given the current level is j, we speak of a *success* if the first j marginals never drop below $\gamma_0/(1+\varepsilon)$; otherwise, we speak of a *failure*. If there are no failures at all, let us assume that $\mathcal{O}\left(n\log\lambda + n^2/\lambda\right)$ is an upper bound on the expected number of generations of the PBIL on LeadingOnes. The following lemma shows that this is also the the expected optimisation time of the PBIL on LeadingOnes.

Lemma 5. *If the expected number of generations required by the* PBIL *to optimise* LeadingOnes *in case of no failure is at most $t^* \in \mathcal{O}\left(n\log\lambda + n^2/\lambda\right)$ regardless of the initial probability vector of the* PBIL*, the expected number of generations of the* PBIL *on* LeadingOnes *is at most $4(1+o(1))t^*$.*

Proof. From the point when the algorithm starts, we divide the time into identical phases, each lasting t^* generations. Let $\mathcal{E}_i$ denote the event that the i-th interval is a failure for $i \in \mathbb{N}$. According to Lemma 3, $\Pr(\mathcal{E}_i) \leq 2n^{-2c}\, \mathcal{O}(n\log\lambda + n^2/\lambda) = \mathcal{O}(n^{-c'+2})$ by union bound for another constant $c' > 0$ when the population is of at most exponential size, that is $\lambda \leq 2^{\alpha n}$ where $\alpha > 0$ is a constant with respect to n, and the constant c large enough such that $c' > 2$, and $\Pr\left(\overline{\mathcal{E}}_1 \wedge \overline{\mathcal{E}}_2\right) \geq 1 - \Pr(\mathcal{E}_1) - \Pr(\mathcal{E}_2) \geq 1 - \mathcal{O}(n^{-c'+2})$ by union bound. Let T be the number of generations performed by the algorithm until a global optimum is found for the first time. We know that $\mathbb{E}\left[T \mid \wedge_{i\in\mathbb{N}} \overline{\mathcal{E}}_i\right] \leq t^*$, and $\Pr\left(T \leq 2t^* \mid \wedge_{i\in\mathbb{N}} \overline{\mathcal{E}}_i\right) \geq 1/2$ since $\Pr\left(T \geq 2t^* \mid \wedge_{i\in\mathbb{N}} \overline{\mathcal{E}}_i\right) \leq 1/2$ by Markov's inequality [10]. We now consider each pair of two consecutive phases. If there is a failure in a pair of phases, we wait until that pair has passed by and then repeat the arguments above as if no failure has ever happened. It holds that

$$\begin{aligned}\mathbb{E}\left[T \mid \overline{\mathcal{E}}_1 \wedge \overline{\mathcal{E}}_2\right] &\leq 2t^* \Pr\left(T \leq 2t^* \mid \overline{\mathcal{E}}_1 \wedge \overline{\mathcal{E}}_2\right) + (2t^* + \mathbb{E}[T]) \Pr\left(T \geq 2t^* \mid \overline{\mathcal{E}}_1 \wedge \overline{\mathcal{E}}_2\right) \\ &= 2t^* + \Pr\left(T \geq 2t^* \mid \overline{\mathcal{E}}_1 \wedge \overline{\mathcal{E}}_2\right) \mathbb{E}[T] \\ &\leq 2t^* + (1/2)\mathbb{E}[T]\end{aligned}$$

since $\Pr\left(T \leq 2t^* \mid \overline{\mathcal{E}}_1 \wedge \overline{\mathcal{E}}_2\right) \geq \Pr\left(T \leq 2t^* \mid \wedge_{i\in\mathbb{N}} \overline{\mathcal{E}}_i\right) \geq 1/2$. Substituting the result into the following yields

$$\begin{aligned}\mathbb{E}[T] &= \Pr\left(\overline{\mathcal{E}}_1 \wedge \overline{\mathcal{E}}_2\right) \mathbb{E}\left[T \mid \overline{\mathcal{E}}_1 \wedge \overline{\mathcal{E}}_2\right] + \Pr\left(\mathcal{E}_1 \vee \mathcal{E}_2\right)\left(2t^* + \mathbb{E}[T]\right) \\ &\leq \Pr\left(\overline{\mathcal{E}}_1 \wedge \overline{\mathcal{E}}_2\right)\left(2t^* + (1/2)\mathbb{E}[T]\right) + \Pr\left(\mathcal{E}_1 \vee \mathcal{E}_2\right)\left(2t^* + \mathbb{E}[T]\right) \\ &= 2t^* + ((1/2)\Pr\left(\overline{\mathcal{E}}_1 \wedge \overline{\mathcal{E}}_2\right) + \Pr\left(\mathcal{E}_1 \vee \mathcal{E}_2\right))\mathbb{E}[T] \\ &= 2t^* + \mathbb{E}[T] - (1/2)\Pr\left(\overline{\mathcal{E}}_1 \wedge \overline{\mathcal{E}}_2\right)\mathbb{E}[T].\end{aligned}$$

Thus, $\mathbb{E}[T] \leq 4t^*/\Pr\left(\overline{\mathcal{E}}_1 \wedge \overline{\mathcal{E}}_2\right) = 4t^*\left(1 + o(1)\right)$. □

By the result of Lemma 5, the phase-based analysis that is exploited until there is a pair with no failure only leads to a multiplicative constant in the expectation. We need to calculate the value of t^* that will also asymptotically be the overall expected number of generations of the PBIL on LeadingOnes. We now give our runtime bound for the PBIL on LeadingOnes with sufficiently large population λ. The proof is very straightforward compared to that in [14]. The floor and ceiling functions of $x \in \mathbb{R}$ are $\lfloor x \rfloor$ and $\lceil x \rceil$, respectively.

Theorem 3. *The* PBIL *with margins and offspring population size $\lambda \geq c \log n$ for a sufficiently large constant $c > 0$, parent population size $\mu = \gamma_0 \lambda$ for any constant γ_0 satisfying $\gamma_0 \leq \eta^{\lceil \xi \rceil + 1}/((1+\delta)e)$ where $\xi = \ln(p_0)/(p_0 - 1)$ and $p_0 := \gamma_0/(1+\varepsilon)$ for any positive constants δ, ε and smoothing parameter $\eta \in (0, 1]$, has expected optimisation time $\mathcal{O}\left(n\lambda \log \lambda + n^2\right)$ on* LeadingOnes.

Proof. We strictly follow the procedure recommended in [3].

Step 1: Recall that we use the canonical partition, defined in (2), in which each subset A_j contains individuals with exactly j leading ones. There are a total of $m = n + 1$ levels ranging from A_0 to A_n.

Step 2: Given $|P^{(t)} \cap A_{\geq j}| \geq \gamma_0 \lambda = \mu$ and $|P^{(t)} \cap A_{\geq j+1}| \geq \gamma \lambda$, we prove that the probability of sampling an offspring in $A_{\geq j+1}$ in generation $t+1$ is lower bounded by $(1+\delta)\gamma$ for some constant $\delta > 0$.

Lemma 1 asserts that if we can find a vector $z^{(t)} = (z_1^{(t)}, \ldots, z_j^{(t)})$ that majorises $p_{1:j}^{(t)}$, then the probability of obtaining j successes from a Poisson-binomial distribution with parameters j and $p_{1:j}^{(t)}$ is lower bounded by the same distribution with parameters j and $z^{(t)}$. Following [14], we compare $X_1^{(t)}, \ldots, X_j^{(t)}$ with another sequence of independent Bernoulli random variables $Z_1^{(t)}, \ldots, Z_j^{(t)}$ with success probabilities $z_1^{(t)}, \ldots, z_j^{(t)}$. Note that $Z^{(t)} := \sum_{k=1}^{j} Z_k^{(t)}$. Define $m := \lfloor (\sum_{i=1}^{j} p_i^{(t)} - jp_0)/(1 - \frac{1}{n} - p_0) \rfloor$ where $p_0 := \frac{\gamma_0}{1+\varepsilon}$, and let $Z_1^{(t)}, \ldots, Z_m^{(t)}$

all have success probability $z_1^{(t)} = \ldots = z_m^{(t)} = 1 - \frac{1}{n}$, $Z_{m+2}^{(t)}, \ldots, Z_j^{(t)}$ get p_0 and possibly a random variable $Z_{m+1}^{(t)}$ takes intermediate value $[p_0, 1 - \frac{1}{n}]$ to guarantee $\sum_{i=1}^{j} p_i^{(t)} = \sum_{i=1}^{j} z_i^{(t)}$.

Since $\sum_{i=1}^{j} p_i^{(t)} \geq j \cdot (\prod_{i=1}^{j} p_i^{(t)})^{1/j} \geq j \cdot p_0^{1/j}$ by the Arithmetic Mean-Geometric Mean inequality (see Lemma 7 in the Appendix) and Lemma 4, we get $m \geq \lfloor j(p_0^{1/j} - p_0)/\left(1 - \frac{1}{n} - p_0\right) \rfloor$. Let us consider the following function:

$$g(j) = j \cdot \frac{p_0^{1/j} - p_0}{1 - p_0} - j = j \cdot \frac{p_0^{1/j} - 1}{1 - p_0}.$$

This function has a horizontal asymptote at $y = -\xi$, where $\xi := \frac{\ln p_0}{p_0 - 1}$ (see calculation in the Appendix). Thus, $m \geq j - \lceil \xi \rceil$ for all $j \geq 0$.

Note that we have just performed all calculations on the current model in generation t. The PBIL then updates the current model $p^{(t)}$ to obtain $p^{(t+1)}$ using the component-wise formula $p_i^{(t+1)} = (1-\eta)p_i^{(t)} + \frac{\eta}{\mu} \sum_{k=1}^{\mu} x_i^{(k)}$. For all $i \in [j]$, we know that $\sum_{k=1}^{\mu} x_i^{(k)} = \mu$ due to the assumption of condition (G2). After the model is updated, we obtain

- $z_i^{(t+1)} = 1 - \frac{1}{n}$ for all $i \leq j - \lceil \xi \rceil$,
- $z_i^{(t+1)} \geq (1-\eta)\, p_0 + \eta \geq \eta$ for all $j - \lceil \xi \rceil < i \leq j$, and
- $p_{j+1}^{(t+1)} \geq (1-\eta)\, p_{j+1}^{(t)} + \eta \frac{\gamma}{\gamma_0} \geq \eta \frac{\gamma}{\gamma_0}$ due to $\sum_{k=1}^{\mu} x_{j+1}^{(k)} = \gamma\lambda$.

Let us denote $z_i^{(t+1)} = (1-\eta)z_i^{(t)} + \eta$. Lemmas 1 and 2 assert that $z^{(t+1)}$ majorises $p_{i:j}^{(t+1)}$, and $\Pr(X_{1:j}^{(t+1)} = j) \geq \Pr(Z^{(t+1)} = j)$. In words, the probability of sampling an offspring in $A_{\geq j}$ in generation $t+1$ is lower bounded by the probability of obtaining j successes from a Poisson-binomial distribution with parameters j and $z^{(t+1)}$. More precisely, at generation $t+1$,

$$\begin{aligned}
\Pr(X_{1:j+1}^{(t+1)} = j+1) &\geq \Pr(X_{1:j}^{(t+1)} = j) \cdot \Pr(X_{j+1}^{(t+1)} = 1) \\
&\geq \Pr(Z^{(t+1)} = j) \cdot p_{j+1}^{(t+1)} \geq (1 - 1/n)^{j - \lceil \xi \rceil} \eta^{\lceil \xi \rceil + 1} \gamma/\gamma_0 \geq (1+\delta)\gamma,
\end{aligned}$$

where $\left(1 - \frac{1}{n}\right)^{j - \lceil \xi \rceil} \geq \frac{1}{e}$ and $\gamma_0 \leq \frac{\eta^{\lceil \xi \rceil + 1}}{e(1+\delta)}$ for any constant $\delta > 0$. Thus, condition (G2) of Theorem 1 is verified.

Step 3: Given that $|P^{(t)} \cap A_{\geq j}| \geq \gamma_0 \lambda$, we aim at showing that the probability of sampling an offspring in $A_{\geq j+1}$ in generation $t+1$ is at least z_j. Note in particular that Lemma 4 yields $\Pr(X_{1:j}^{(t+1)} = j) \geq \frac{\gamma_0}{1+\varepsilon}$. The probability of sampling an offspring in $A_{\geq j+1}$ in generation $t+1$ is lower bounded by

$$\Pr(X_{1:j}^{(t+1)} = j) \cdot \Pr(X_{j+1}^{(t+1)} = 1) \geq \frac{\gamma_0}{1+\varepsilon} \cdot \frac{1}{n} =: z_j.$$

where $\Pr(X_{j+1}^{(t+1)} = 1) = p_{j+1}^{(t+1)} \geq \frac{1}{n}$. Therefore, condition (G1) of Theorem 1 is satisfied with $z_j = z_* = \frac{\gamma_0}{(1+\varepsilon)n}$.

Step 4: Condition (G3) of Theorem 1 requires a population size $\lambda = \Omega(\log n)$. This bound matches with the condition on $\lambda \geq c \log n$ for some sufficiently large constant $c > 0$ from the previous lemmas. Overall, $\lambda = \Omega(\log n)$.

Step 5: When $z_j = \frac{\gamma_0}{(1+\varepsilon)n}$ where $\gamma_0 \leq \frac{\eta^{\lceil \xi \rceil + 1}}{(1+\delta)e}$ and $\lambda \geq c \log n$ for some constants $\varepsilon > 0$, $\eta \in (0,1]$ and sufficiently large $c > 0$, all conditions of Theorem 1 are verified. Using that $\ln\left(\frac{6\delta\lambda}{4+\delta\lambda z_j}\right) < \ln\left(\frac{3\delta\lambda}{2}\right)$ an upper bound on the expected optimisation time of the PBIL on LeadingOnes is guaranteed as follows.

$$\left(\frac{8}{\delta^2}\right)\sum_{j=0}^{n-1}\left[\lambda \ln\left(\frac{3\delta\lambda}{2}\right) + \frac{1}{z_j}\right] < \frac{8}{\delta^2} n\lambda \log \lambda + \frac{8(1+\varepsilon)}{\delta^2 \gamma_0} n^2 + o\left(n^2\right) \in \mathcal{O}\left(n\lambda \log \lambda + n^2\right).$$

Hence, the expected number of generations t^* is $\mathcal{O}\left(n \log \lambda + \frac{n^2}{\lambda}\right)$ for a sufficiently large λ in the case of no failure and, thus, meets the assumption in Lemma 5. The expected optimisation time of the PBIL on LeadingOnes is still asymptotically $\mathcal{O}\left(n\lambda \log \lambda + n^2\right)$. This completes the proof. □

Our improved upper bound of $\mathcal{O}\left(n^2\right)$ on the optimisation time of the PBIL with population size $\lambda = \Theta(\log n)$ on LeadingOnes is significantly better than the previous bound $\mathcal{O}\left(n^{2+\varepsilon}\right)$ from [14]. Our result is not only stronger, but the proof is much simpler as most of the complexities of the population dynamics of the algorithm is handled by Theorem 1 [3]. Furthermore, we also provide specific values for the multiplicative constants, i.e. $\frac{32}{\delta^2}$ and $\frac{32(1+\varepsilon)}{\delta^2\gamma_0}$ for the terms $n\lambda \log \lambda$ and n^2, respectively (see Step 5 in Theorem 3). Moreover, the result also matches the runtime bound of the UMDA on LeadingOnes for a small population $\lambda = \Theta(\log n)$ [4].

Note that Theorem 3 requires some condition on the selective pressure, that is $\gamma_0 \leq \frac{\eta^{\lceil \xi \rceil + 1}}{(1+\delta)e}$ where $\xi = \frac{\ln p_0}{p_0 - 1}$ and $p_0 := \frac{\gamma_0}{1+\varepsilon}$ for any positive constants δ, ε and smoothing parameter $\eta \in (0,1]$. Although for practical applications, we have to address these constraints to find a suitable set of values for γ_0, this result here tells us that there exists some settings for the PBIL such that it can optimise LeadingOnes within $\mathcal{O}\left(n\lambda \log \lambda + n^2\right)$ time in expectation.

4 Runtime Analysis of the PBIL on BinVal

We first partition the search space into non-empty disjoint subsets $A_0, \ldots, A_n$.

Lemma 6. *Let us define the levels as* $A_j := \{x \in \{0,1\}^n \mid \sum_{i=1}^{j} 2^{n-i} \leq \text{BinVal}(x) < \sum_{i=1}^{j+1} 2^{n-i}\}$, *for* $j \in [n] \cup \{0\}$, *where* $\sum_{i=1}^{0} 2^{n-i} = 0$. *If a bitstring x has exactly j leading ones, i.e.* LeadingOnes$(x) = j$, *then* $x \in A_j$.

Proof. Consider a bitstring $x = 1^j 0 \{0,1\}^{n-j-1}$. The fitness contribution of the first j leading ones to BinVal(x) is $\sum_{i=1}^{j} 2^{n-i}$. The $(j+1)$-th bit has no contribution, while that of the last $n-j-1$ bits ranges from zero to $\sum_{i=j+2}^{n} 2^{n-i} = \sum_{i=0}^{n-j-2} 2^i = 2^{n-j-1}-1$. So overall, $\sum_{i=1}^{j} 2^{n-i} \le \textsc{BinVal}(x) \le \sum_{i=1}^{j+1} 2^{n-i}-1 < \sum_{i=1}^{j+1} 2^{n-i}$. Hence, the bitstring x belongs to level A_j. $\square$

In both problems, all that matters to determine the level of a bitstring is the position of the first 0-bit when scanning from the most significant to the least significant bits. Now consider two bitstrings in the same level for BinVal, their rankings after the population is sorted are also determined by some other less significant bits; however, the proof of Theorem 3 never takes these bits into account. Thus, the following corollary yields the first upper bound on the expected optimisation time of the PBIL and the UMDA (when $\eta = 1$) for BinVal.

Corollary 1. *The* PBIL *with margins and offspring population size $\lambda \ge c \log n$ for a sufficiently large constant $c > 0$, parent population size $\mu = \gamma_0\lambda$ for any constant γ_0 satisfying $\gamma_0 \le \eta^{\lceil\xi\rceil+1}/((1+\delta)e)$ where $\xi = \ln(p_0)/(p_0-1)$ and $p_0 := \gamma_0/(1+\varepsilon)$ for any positive constants δ, ε and smoothing parameter $\eta \in (0,1]$, has expected optimisation time $\mathcal{O}\left(n\lambda \log \lambda + n^2\right)$ on* BinVal.

5 Conclusions

Runtime analyses of EDAs are scarce. Motivated by this, we have derived an upper bound of $\mathcal{O}\left(n\lambda \log \lambda + n^2\right)$ on the expected optimisation time of the PBIL on both LeadingOnes and BinVal for a population size $\lambda = \Omega(\log n)$. The result improves upon the previously best-known bound $\mathcal{O}\left(n^{2+\varepsilon}\right)$ from [14], and requires a much smaller population size $\lambda = \Omega(\log n)$, and uses relatively straightforward arguments. We also presents the first upper bound on the expected optimisation time of the PBIL on BinVal.

Furthermore, our analysis demonstrates that the level-based theorem can yield runtime bounds for EDAs whose models are updated using information gathered from the current and previous generations. An additional aspect of our analysis is the use of the DKW inequality to bound the true distribution by the empirical population sample when the number of samples is large enough. We expect these arguments will lead to new results in runtime analysis of evolutionary algorithms.

Appendix

Lemma 7 (AM-GM Inequality [12]). *Let $x_1, \ldots, x_n$ be n non-negative real numbers. It always holds that*

$$\frac{x_1 + x_2 + \cdots + x_n}{n} \ge \sqrt[n]{x_1 \cdot x_2 \cdots x_n},$$

and equality holds if and only if $x_1 = x_2 = \cdots = x_n$.

Proof (of horizontal asymptote). The function can be rewritten as $g(j) = \frac{1}{1-p_0} \cdot \frac{p_0^{1/j}-1}{1/j}$. Denote $t := 1/j$, we obtain $g(t) = \frac{1}{1-p_0} \cdot \frac{p_0^t-1}{t}$. Applying L'Hôpital's rule yields:

$$\lim_{j \to +\infty} g(j) = \lim_{t \to 0^+} g(t) = \frac{\lim_{t \to 0^+} (p_0^t \ln p_0)}{1-p_0} = \frac{\ln p_0}{1-p_0} = -\frac{\ln p_0}{p_0 - 1}.$$

□

References

1. Baluja, S.: Population-based incremental learning: A method for integrating genetic search based function optimization and competitive learning. Tech. rep., Carnegie Mellon University (1994)
2. Boland, P.J., Proschan, F.: The reliability of k out of n systems. The Annals of Probability **11**(3), 760–764 (1983)
3. Corus, D., Dang, D.C., Eremeev, A.V., Lehre, P.K.: Level-based analysis of genetic algorithms and other search processes. IEEE Transactions on Evolutionary Computation **PP**(99), 1–1 (2017)
4. Dang, D.C., Lehre, P.K.: Simplified runtime analysis of estimation of distribution algorithms. In: Proceedings of the Genetic and Evolutionary Computation Conference. pp. 513–518. GECCO '15 (2015)
5. Droste, S.: A rigorous analysis of the compact genetic algorithm for linear functions. Natural Computing **5**(3), 257–283 (2006)
6. Gleser, L.J.: On the distribution of the number of successes in independent trials. The Annals of Probability **3**(1), 182–188 (1975)
7. Krejca, M.S., Witt, C.: Lower bounds on the run time of the univariate marginal distribution algorithm on onemax. In: Proceedings of the Foundation of Genetic Algorithms. pp. 65–79. FOGA '17 (2017)
8. 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. pp. 1383–1390. GECCO '17 (2017)
9. Massart, P.: The tight constant in the Dvoretzky-Kiefer-Wolfowitz inequality. The Annals of Probability **18**(3), 1269–1283 (1990)
10. Mitzenmacher, M., Upfal, E.: Probability and Computing: Randomized Algorithms and Probabilistic Analysis. Cambridge University Press (2005)
11. Mühlenbein, H., Paaß, G.: From recombination of genes to the estimation of distributions i. binary parameters. In: Proceedings of the Parallel Problem Solving from Nature. pp. 178–187. PPSN IV (1996)
12. Steele, J.M.: The Cauchy-Schwarz Master Class: An Introduction to the Art of Mathematical Inequalities. Cambridge University Press (2004)
13. Witt, C.: Upper bounds on the runtime of the univariate marginal distribution algorithm on onemax. In: Proceedings of the Genetic and Evolutionary Computation Conference. pp. 1415–1422. GECCO '17 (2017)
14. Wu, Z., Kolonko, M., Möhring, R.H.: Stochastic runtime analysis of the cross-entropy algorithm. IEEE Transactions on Evolutionary Computation **21**(4), 616–628 (2017)