Robustness of Populations in Stochastic Environments

Christian Gießen
Christian-Albrechts-Universität zu Kiel
24118 Kiel, Germany

Timo Kötzing
Fiedrich-Schiller-Universität Jena
07743 Jena, Germany

ABSTRACT

We consider stochastic versions of OneMax and LeadingOnes and analyze the performance of evolutionary algorithms with and without populations on these problems. It is known that the (1+1) EA on OneMax performs well in the presence of very small noise, but poorly for higher noise levels. We extend these results to LeadingOnes and to many different noise models, showing how the application of drift theory can significantly simplify *and generalize* previous analyses.

Most surprisingly, even small populations (of size $\Theta(\log n)$) can make evolutionary algorithms perform well for high noise levels, well outside the abilities of the (1+1) EA! Larger population sizes are even more beneficial; we consider both parent and offspring populations. In this sense, populations are *robust* in these stochastic settings.

Categories and Subject Descriptors

F.2 [**Theory of Computation**]: Analysis of Algorithms and Problem Complexity

General Terms

Theory, Algorithms, Performance

Keywords

Run Time Analysis, Stochastic Fitness Function, Evolutionary Algorithm, Populations, Robustness

1. INTRODUCTION

Evolutionary algorithms (EAs) are general-purpose problem solvers which can be successfully applied to a wide variety of problems with small effort. In particular, EAs are popular where no tailored solutions exist, for example because the structure of the problem is inaccessible (given as a *black box*) or where the structure of the problem is very complicated. In particular, EAs are popular in settings including uncertainties, such as noisy fitness (quality) evaluations; see [BDGG09] for a survey on examples in combinatorial optimization, but also [JB05] for an excellent survey also discussing different sources of uncertainty.

We are interested in formally analyzing the performance of EAs in settings where the fitness function is probabilistic, i.e., a given search point can have different fitness values for different fitness evaluations. One way to deal with such uncertainty is to replace fitness evaluations with an average of a (large) sample of fitness evaluations and then proceed as if there was no noise. In this paper we are interested in a different approach where we accept the noise and try to analyze how much noise can be overcome by EAs without further modifications (note that this research can also be used to decide *how much* resampling is necessary for successful optimization). This was first done in [Dro04], where a noisy variant of the well-known OneMax test function was analyzed for the simplest EA, the (1+1) EA. In essence, it was shown that the (1+1) EA can deal with small noise levels, but not medium noise levels. Recently, there was a sequence of paper discussing ant colony optimization for path finding problems in the presence of uncertainty [ST12, DHK12, FK13], see also [GP96, Gut03] for early work in this area.

For this paper we are exclusively concerned with optimization problems defined on bit strings of fixed length n. In this domain we have the two well-known (static) test functions OneMax and LeadingOnes as follows. For each bit string $x \in \{0,1\}^n$ we let OneMax(x) be the number of 1s in x and LeadingOnes(x) is the number of consecutive 1s counting from the left until the first occurrence of a 0. The performance of various randomized search heuristics on these two *static* problems is known in detail.

We modify these test functions by adding noise. We distinguish between two general noise models: *prior noise* and *posterior noise*. In the first model we assume that the noise comes from not evaluating the search point in questions, but a noisy variant; this corresponds to the noise model used in [Dro04], where, with probability p, a bit of the search point was flipped *before* evaluation. In the second model the fitness value of a search point obtains noise *after* evaluation. For example, one can add a value drawn from a centered normal distribution (or add a value drawn from any other chosen distribution; we call such noise *additive posterior noise*). Posterior noise is essentially the model used in [GP96, ST12, DHK12, FK13].

In each case we consider the noise to be independent for different elements of the search space and for reevaluations

GECCO'14, July 12–16, Vancouver, BC, Canada.

ACM 978-1-4503-2662-9/14/07 ...$15.00.
http://dx.doi.org/10.1145/2576768.2598227.

(note that we assume that all algorithms reevaluate each search point under consideration in every iteration).

In this paper we expand on the work done in [Dro04] in three ways. First, we make the results applicable to many different noise models; second, we analyze the LeadingOnes function in noisy settings; third, we show how the use of populations can make the EA much more robust towards noise.

Regarding the generalization, we reprove the results from [Dro04] as a corollary to more general theorems which can be applied in many different settings. The proofs of these more general theorems rely heavily on *drift theory*, a modern tool which facilitates the formal analysis of randomized search heuristics significantly. Note that this tool was not available for [Dro04]. Another tool suitable for the analysis of populations was recently introduced in [Leh11] in the context of *non-elitism*, i.e. just as in the setting with noise, good solutions can get lost.

Regarding the LeadingOnes test function, we give the first formal analysis of this test function in a noisy setting.

Regarding the use of populations, the paper [PB10] gives a nice overview of several different aspects where populations (and the use of crossover operators) are beneficial for optimization of *static* fitness functions. In contrast to this, we show that populations can also be highly beneficial for the optimization of *stochastic* fitness functions, as they allow a much higher noise.

1.1 Detailed Contribution

The only algorithm we consider is the $(\mu + \lambda)$ EA, for different values of μ and λ (see Section 2 for a detailed description). We consider the (1+1) EA as an EA "without population"; this was the algorithm considered in [Dro04]. Even when we discuss EAs with populations, we only consider cases with $\mu = 1$ or $\lambda = 1$, for simplicity.

We consider optimization successful in the stochastic setting as soon as the algorithm has evaluated the best static solution (in both our cases the all-1s bit string); note that the best static solution is, in all of our models, also the solution with best expected fitness. Whenever we consider the "run time" of an algorithm, this is understood as the *expected* number of *iterations* (or generations) of the EA. In particular, population-based EAs have a higher number of fitness evaluations than iterations (also due to reevaluation of old search points).

In Section 3 we consider OneMax. In particular, we give a general theorem for deriving upper bounds in different noise settings (Theorem 4) and a general Theorem for deriving lower bounds (Theorem 5). As a result we completely reprove the theorems from [Dro04] (Corollary 6). These results concern prior noise where, with probability p, a (uniformly chosen) bit is flipped: The (1+1) EA is successful in this setting for values of p up to $O(\log n/n)$ away from 0 or 1 (i.e., optimizes in polynomial time), otherwise it is unsuccessful. Note that [Dro04] did not cover values of p close to 1.

As a further corollary, we show that the (1+1) EA can optimize in the presence of additive posterior noise with variance of $O(\log n/n)$ efficiently, but not, for example, in the presence of additive noise from an exponential distribution with parameter 1 (Corollary 7). This list of corollaries can easily be extended, for example to cover the case of prior noise based on mutation or similar models.

In Section 3.2 we show that populations can be much more robust towards noise. For example, a linear population is large enough to allow *arbitrary* values of p in the setting of prior bit-flip noise. Furthermore, in the case of constant p in the setting of prior bit-flip noise (a setting far outside the abilities of the (1+1) EA for efficient optimization), already a *logarithmic population size* suffices for efficient optimization. Similarly, we get robustness of small populations for posterior noise models, for example for exponentially distributed noise with constant parameter.

In Section 4, we give our results for LeadingOnes. We show that the (1+1) EA optimizes successfully in the presence of small noise, but we also give an example of higher noise levels where optimization is unsuccessful. Here again populations are helpful, even of logarithmic size.

We conclude the paper with a discussion in Section 5.

2. MATHEMATICAL PRELIMINARIES

In this paper we consider the $(\mu + \lambda)$ EA, an algorithm which bases its progress on mutation (see Algorithm 1 for a detailed description). We consider only the mutation operator which flips each bit independently with probability $1/n$. Ties in the selection of fitter individuals are broken so that individuals from the offspring population are preferred (this allows the (1+1) EA to cross plateaus and is consistent with the definition of, for example, [Dro04]); further ties are broken uniformly at random.

Algorithm 1: $(\mu + \lambda)$ EA

Let P be a set of μ uniformly chosen bit strings;
repeat
  $O \leftarrow \emptyset$;
  for $i = 1$ **to** λ **do**
    pick x u.a.r. from P;
    $O \leftarrow O \cup \{\texttt{mutate}(x)\}$;
  for $x \in P \cup O$ **do** evaluate $f(x)$;
  $P \leftarrow \mu$ f-maximal elements from $P \cup O$;
until *forever*;

Note that all references to the "run time" or the "number of steps" of an algorithm always concern the *expected first hitting time* of the optimum, as mentioned above.

2.1 Drift Theorems

We will use a variety of drift theorems to derive the theorems of this paper. *Drift*, in this context, describes the *expected change* of the best-so-far solution within one iteration with respect to some *potential*. In later proofs we will define potential functions on best-so-far solutions and prove bounds on the drift; these bounds then translate to expected run times with the use of the drift theorems from this section.

The literature knows a large number of drift theorems; this selection is not representative, but merely contains those theorems needed for this paper.

The simplest drift theorem concerns *additive drift*.

Theorem 1 (Additive Drift [HY04]). *Let $(X_t)_{t\geq 0}$ be random variables describing a Markov process over a finite state space $S \subseteq \mathbb{R}$. Let T be the random variable that denotes the earliest point in time $t \geq 0$ such that $X_t = 0$. If* *there exist $c > 0$ such that*

$$E(X_t - X_{t+1} | T > t) \geq c,$$

then

$$E(T|X_0) \leq \frac{X_0}{c}.$$

We will not give the version of the *multiplicative drift theorem* for upper bounds, due to [DJW12], as it is implied by the *Variable Drift Theorem* given in [Joh10, Theorem 4.6] (independently developed in [MRC08, Section 8]); this drift theorem is applicable when the drift is not uniform across the search space; frequently one can find a uniform lower bound and use the additive drift theorem, but using the variable drift theorem will typically give much better bounds. The version of the Variable Drift Theorem that we use is due to [RS12], which removes the restriction of h being differentiable.

THEOREM 2 (VARIABLE DRIFT [RS12]). *Let $(X_t)_{t\geq 0}$ be random variables describing a Markov process over a finite state space $S \subseteq \mathbb{R}_0^+$ and let $x_{\min} := \min\{x \in S \mid x > 0\}$. Furthermore, let T be the random variable that denotes the first point in time $t \in \mathbb{N}$ for which $X_t = 0$. Suppose that there exists a monotone increasing function $h : \mathbb{R}^+ \to \mathbb{R}^+$ such that $1/h$ is integrable and*

$$E(X_t - X_{t+1} \mid X_t) \geq h(X_t)$$

holds for all $t < T$. Then,

$$E(T \mid X_0) \leq \frac{x_{\min}}{h(x_{\min})} + \int_{x_{\min}}^{X_0} \frac{1}{h(x)} dx.$$

Finally, in order to derive *lower* bounds on the run time of EAs, we use the Negative Drift Theorem.

THEOREM 3 (NEGATIVE DRIFT [OW11, OW12]). *Let $(X_t)_{t\geq 0}$ be real-valued random variables describing a stochastic process over some state space. Suppose there is an interval $[a, b] \subseteq \mathbb{R}$, two constants $\delta, \varepsilon > 0$ and, possibly depending on $\ell = b - a$, a function $r(\ell)$ satisfying $1 \leq r(\ell) = o(\ell/\log \ell)$ such that, for all $t \geq 0$, the following conditions hold.*

1. *$E(X_{t+1} - X_t \mid a < X_t < b) \geq \varepsilon$;*
2. *For all $j \geq 0$, $P(|X_{t+1} - X_t| \geq j \mid a < X_t) \leq \frac{r(\ell)}{(1+\delta)^j}$.*

Then there is a constant c such that, for $T = \min\{t \geq 0 : X_t \leq a \mid X_0 \geq b\}$, we have

$$P(T \leq 2^{c\ell/r(\ell)}) = 2^{-\Omega(\ell/r(\ell))}.$$

3. ONEMAX

In this section we present our results regarding OneMax. We fix a (stochastic) OneMax function f according to one of our models. It is easy to verify that, in each of the stochastic models we consider, there is a sequence of independent random variables $(X_k)_{k\leq n}$ such that the following holds.

- For each evaluation of f on a bit string with exactly k 1s, the return value is drawn at random $\sim X_k$ (recall that all evaluations of fitness functions are independent).
- $\forall j \leq k < n : P(X_j < X_{k+1}) \geq P(X_k < X_{k+1})$; intuitively, the larger the true difference in OneMax-value, the more likely this is reflected in a random OneMax evaluation. This simplifies some conditions.

Note that these two properties capture two important properties of the OneMax function: symmetry in the positions (only the number of 1s decides on the fitness, not the position), and monotonicity (the more 1s a bit string has, the higher its fitness; we need this comparison-based version, given the comparison-based definition of the (1+1) EA).

We start by giving an upper and a lower bound for the (1+1) EA in Section 3.1; in Section 3.2 we give upper bounds for population-based EAs, showing that populations are efficient for the stochastic versions of OneMax we consider.

3.1 (1+1) EA

Our first theorem gives an upper bound for the (1+1) EA on OneMax, generalizing a theorem from [Dro04].

THEOREM 4. *Suppose there is a positive constant $c < 1/9$ such that*

$$\forall k < n : P(X_k < X_{k+1}) \geq 1 - c\frac{n-k}{n}. \qquad (1)$$

Then the (1+1) EA optimizes f in $\Theta(n \log n)$ steps. Furthermore, if Equation (1) holds for all $k < n - \ell$ for some $\ell > 2$, and we have

$$\forall k < n : P(X_k < X_{k+1}) \geq 1 - \frac{\ell}{n},$$

then (1+1) EA optimizes f in $n^2 + n2^{O(\ell)}$ steps.

PROOF. Let $p_k = c(n - k)/n$.

We show that there is a positive drift on the number of 1s. Let k be the number of 1s of the current search point.

Let E_0 be the event that the new search point has at least one more 1 and the comparison of old and new search point indicates correctly that the new search point is better. Let E_1 be the event that the new search point has less 1s than the current search point, and that this new search point is nonetheless accepted. Clearly, the *expected* number of 1s conditional on E_1 is at least $k - 2$. Note that $P(E_0) \geq (1-p_k)(n-k)/(en)$ and $P(E_1) \leq p_k k/n$. Thus, the expected *increase* in the number of 1s is at least

$$\begin{aligned} P(E_0) - 2P(E_1) &\geq (1 - p_k)\frac{n-k}{en} - 2p_k\frac{k}{n} \\ &\geq \frac{n-k}{en} - p_k\left(2 + \frac{n-k}{en}\right) \\ &\geq \frac{n-k}{en} - 3p_k \\ &= \frac{n-k}{n}(1/e - 3/c). \end{aligned}$$

Using $c < 1/9$ we see that $1/e-3/c$ is a constant > 0, giving a multiplicative drift as desired.

Regarding the "furthermore" clause, we do not have sufficient drift when we use the number of 1 as the potential function. Thus, we change our potential function in a way which could also be used to show an $\exp(O(n))$ bound for optimizing the needle function with the (1+1) EA. Intuitively, this drift function takes care of the plateau of the last ℓ OneMax values.

We define a helper function a so that, for all $i \geq 1$,

$$a(i) = \begin{cases} \frac{i!}{\ell!}(4\ell)^{\ell-i}, & \text{if } i \leq \ell; \\ 1, & \text{otherwise.} \end{cases}$$

We now define the potential in terms of a as follows. A given search point with exactly k 0s has potential

$$g(k) = \sum_{i=1}^{k} a(i).$$

This potential function makes it particularly easy to compute the differences of the potentials of similar search points (it is a sum of successive $a(i)$). Also note that the sequence a is falling quickly for arguments $\leq \ell$, so that the sum of such elements can be bounded from above by twice its largest element. We only sketch the remaining argument. From our observations and the assumptions of the theorem one can directly compute a drift of $1/n$ towards the optimum, using almost the same arguments as in the first part of this proof (the biggest exception is that, for jumps away from the optimum, we notice that larger and larger jumps hurt us less and less, so that an expected jump of 2 away is sufficient to consider). It is easy to see that $g(n) = n + \exp(O(\ell))$, which gives the desired result. □

Now we come to our second theorem, a lower bound for the (1+1) EA on OneMax.

Theorem 5. *Suppose there is $\ell \leq n/4$ and a constant $c > 16$ such that*

$$\forall k, n - \ell \leq k < n : P(X_k < X_{k+1}) \leq 1 - c\frac{n-k}{n}.$$

Then the (1+1) EA optimizes f in $2^{\Omega(\ell)}$ many steps.

Proof. We want to show that there is a constant negative drift on the number of ones in the interval between $n-\ell$ and n; however, we will count iterations of the process only when an actual change in the number of 1s occurs (i.e., we condition on the change), as otherwise the drift would be too small (this will only yield a smaller bound than when counting all other steps as well). See also [RS12] regarding an explicit negative drift theorem in the presence of self-loops. Let $k \geq n - \ell$ be the number of 1s of the current search point. Since we do not count steps without change, it suffices to show that the drift conditional on lowering the number of 1s is at least twice the drift conditional on increasing the number of 1s (i.e., the negative drift is twice the size of the positive drift). This would yield a negative drift of at least 2/3 (this uses that we have a minimal step width of 1).

Let $p_k = c(n-k)/n$. Clearly, the negative drift is at least $p_k k/(en)$.

Let E_0 be the event that the new search point has at least one more 1 and the comparison of old and new search point indicates correctly that the new search point is better. Clearly, the *expected* number of 1s conditional on E_0 is at most $k+2$ and $P(E_0) \leq (n-k)/n$. This gives a positive drift of at most $2(n-k)/n$. Dividing the lower bound for the negative drift by the upper bound for the positive drift, we get a ratio of at least

$$\frac{p_k k}{en} \frac{n}{2(n-k)} = \frac{ck}{2en}.$$

From $k \geq 3n/4$ and $c > 16$ we get the desired bound of 2 on the ratio. As the (1+1) EA makes long jumps with sufficiently small probability, an application of the Negative Drift Theorem (Theorem 3) concludes the proof. □

Our two theorems can be used for easy corollaries, showing the optimization time of the (1+1) EA given different noise models. We first consider the noise model given in [Dro04].

Corollary 6 ([Dro04]). *Suppose prior noise which, with probability p, flips a bit uniformly at random. Then we have that the (1+1) EA optimizes* OneMax *in time*

$$\begin{cases} \Theta(n \log n), & \text{if } p \leq 1/(10n); \\ \text{polynomial}, & \text{if } p = O(\log n/n); \\ \text{superpolynomial}, & \text{if } p = \omega(\log n/n) \cap 1 - \omega(\log n/n); \\ \text{polynomial}, & \text{if } p = 1 - O(\log n/n). \end{cases}$$

Proof. Suppose first $p \leq c/n$ for some $c \leq 1/10$ and let $k < n$. We estimate $P(X_k < X_{k+1})$ by observing that that the event $X_k \geq X_{k+1}$ requires the individual with k 1s to be evaluated to $k+1$ or the other to k. The first option has probability $\leq p(n-k)/n$, the second of $p \leq c/n$. Thus, we get the desired bound from Theorem 4.

In the case of $p \leq c \log n/n$ we similarly get the bound $P(X_k \geq X_{k+1}) \leq c \log n/n$; this is sufficient up to a distance of $c \log n$ from the optimum, which gives a polynomial bound.

Suppose $p = \omega(\log n/n) \cap 1 - \omega(\log n/n)$. Then, for all k, we estimate $P(X_k = X_{k+1})$ as $\geq p(1-p)$ (either both X_k and X_{k+1} evaluate to k or both to $k+1$). Theorem 5 gives a superpolynomial run time.

Suppose now $p = 1 - O(\log n/n)$. If X_k evaluates to $k-1$, then $X_k < X_{k+1}$. However, $P(X_k = k-1) \geq \log n/(2n)$ so that Theorem 4 gives the desired result. □

Regarding posterior noise we give the following corollary.

Corollary 7. *Suppose posterior noise, sampling from some distribution D with variance σ^2. Then we have that the (1+1) EA optimizes* OneMax *in polynomial time if $\sigma^2 = O(\log n/n)$. On the other hand, if, for example, D is exponentially distributed with parameter 1, then the (1+1) EA optimizes* OneMax *in superpolynomial time only.*

Proof. We have $X_k \sim k + D$ and $X_{k+1} \sim k + 1 + D$; let D' be the difference of two independent copies of D. We have

$$P(X_k < X_{k+1}) = P(0 < X_{k+1} - X_k) = P(-1 < D').$$

The variance of the difference of two i.i.d. centered variables both with variance σ^2 is $2\sigma^2$; let D' be this difference. Now we apply Chebyshev's Inequality to see that

$$P(|D'| \geq 1) \leq 2\sigma^2.$$

Thus, for $\sigma^2 = O(\log n/n)$, the probability of being at least 1 from the mean while having a variance of $O(\log n/n)$ is $O(\log n/n)$, as sufficient for polynomial run time (see Theorem 4).

In the case of D an exponential distribution we have a constant chance of $X_k \geq k+2$ and $X_{k+1} \leq k+2$, which leads to the claimed result using Theorem 5. □

3.2 Population-Based EA

In this section we give upper bounds for a population-based EA. In particular, we consider parent populations (i.e., the $(\mu+1)$ EA). From [Wit06] we know that the $(\mu+1)$ EA needs $\Theta(\mu n + n \log n)$ iterations to optimize the *static* version of OneMax. We conjecture a similar run time for sufficiently benevolent noisy versions, but, for simplicity also for the requirements on the noise model, we only give a slightly weaker bound.

Theorem 8. *Let μ be given and, for each $k < n$, let Y_k denote the minimum over μ independent copies of X_k. If there is a positive constant $c < 1/9$ such that*

$$\forall k, n/4 < k < n : P(Y_k < X_{k+1}) \geq 1 - c\frac{n-k}{n\mu}, \quad (2)$$

then the $(\mu+1)$ EA optimizes f in $O(\mu n \log n)$ iterations.

Note that the requirement of Equation (2) might seem to get more restrictive with growing μ; however, it only gets linearly more restrictive in the fraction on the right-hand-side, while the impact of growing μ on the random variable Y_k is typically much stronger.

Proof. Let $p_k = c(n-k)/(n\mu)$.

We show that there is a positive drift on the number of ones in the current best search point. Let k be the number of 1s of the current best search point.

Let E_0 be the event that the new search point has at least one more 1 than the best one and the comparison of old and new search point indicates correctly that the new search point is better. We have $P(E_0) \geq (1-p_k)(n-k)/(e\mu n)$.

Let E_1 be the event that the new search point has less 1s than the current best search point, the best search point is unique, and that this unique search point is discarded (if the best search point is not unique, E_1 is the empty event). Clearly, the *expected* number of 1s conditional on E_1 (if $E_1 \neq \emptyset$) is at least $k-2$. We have $P(E_1) \leq p_k$. Thus, the expected *increase* in the number of 1s is at least

$$\begin{aligned} P(E_0) - 2P(E_1) &\geq (1-p_k)\frac{n-k}{e\mu n} - 2p_k \\ &= \frac{n-k}{e\mu n} - p_k\left(2 + \frac{n-k}{e\mu n}\right) \\ &\geq \frac{n-k}{e\mu n} - 3p_k. \end{aligned}$$

Using the choice of c and the definition of p_k, we see that we have sufficient multiplicative drift as desired. □

From these theorems we can again derive many corollaries regarding concrete noise models. These includes corollaries implying an exponential speedups of populations of logarithmic size, when compared with the performance of the (1+1) EA!

Corollary 9. *Suppose prior noise which, with probability p, flips a bit uniformly at random. Let $\mu \geq 5\log n/p$. Then we have that the $(\mu+1)$ EA optimizes* OneMax *in time $O(\mu n \log n)$. In particular, for $p = 1/2$, we have that a population size of $\mu \geq 10 \log n$ suffices for an optimization time of $O(\mu n \log n)$.*

Proof. For $k \geq n/4$, the probability that none of μ individuals with k 1s is evaluated to $k-1$ is $\leq 1/(\mu n)$, as a simple computation shows (we omit the details of the application of the Chernoff Bound). Thus, Theorem 8 gives the desired result. □

The intuitive reason behind the previous corollary was that, regardless of the noisy evaluation of the best-so-far individual, all worse individuals have a constant chance *each* to be strictly worse.

Corollary 10. *Let any non-negative additive posterior noise be given which has a non-zero constant probability of evaluating to < 1. Then there is a constant c such that, for $\mu \geq c \log n$, the $(\mu+1)$ EA optimizes* OneMax *in time $O(\mu n \log n)$.*

Proof. Let D be the posterior noise and $p = P(D < 1)$ a constant $\neq 0$. The case of $p = 1$ is trivial. Let $c = -2/\log(1-p)$ and $\mu = \lceil c \log n \rceil$. We have

$$P(Y_k \geq k+1) = P(D \geq 1)^\mu \leq (1-p)^{-2\log n/\log(1-p)} = n^{-2}.$$

This implies the claimed result. □

Note that the last corollary applies, for example, to additive posterior noise taken from an exponential distribution with parameter 1.

4. LEADINGONES

We follow up on the section about OneMax with results for LeadingOnes. For this purpose we now fix a stochastic LeadingOnes function f according to one of our models. For each k, we let x_k^{opt} be the bit string which has only 1s, except for position $k+1$; let x_k^{pes} be the bit string with k leading ones and otherwise only 0s. In a sense, x_k^{opt} is *optimal* for a bit string with a leading ones value of k, while x_k^{pes} is pessimal. We let $(X_k^{\mathrm{opt}})_{k \leq n}$ and $(X_k^{\mathrm{pes}})_{k \leq n}$ be two sequences of independent random variables such that, for all $k \leq n$, $X_k^{\mathrm{opt}} \sim f(x_k^{\mathrm{opt}})$ and $X_k^{\mathrm{pes}} \sim f(x_k^{\mathrm{pes}})$. We will assume the following about f.

- For each evaluation of f on a bit string with the leftmost zero at position $k+1$, the return value is drawn according to a distribution which is in between X_k^{pes} and X_k^{opt} with respect to stochastic dominance.
- $\forall j \leq k < n : P(X_j^{\mathrm{opt}} < X_{k+1}^{\mathrm{opt}}) \geq P(X_k^{\mathrm{opt}} < X_{k+1}^{\mathrm{opt}})$.
- $\forall j \leq k < n : P(X_j^{\mathrm{pes}} < X_{k+1}^{\mathrm{pes}}) \geq P(X_k^{\mathrm{pes}} < X_{k+1}^{\mathrm{pes}})$.

We show that, despite the more drastic consequences of noise, we still find sufficient conditions for efficient optimization similar to the ones we have already seen in Section 3.

We begin by giving upper and lower bounds for the (1+1) EA in Section 4.1. In Section 4.2 and Section 4.3 we show the effectiveness of parent and offspring populations, respectively, for the stochastic LeadingOnes problem by giving upper bounds.

4.1 (1+1) EA

Theorem 11. *Suppose there is a positive constant $c < 1/12$ such that*

$$\forall k < n : P(X_k^{\mathrm{opt}} < X_{k+1}^{\mathrm{pes}}) \geq 1 - \frac{c}{kn} \quad (3)$$

Then the (1+1) EA optimizes f in $O(n^2)$ steps.

PROOF. Let $p_k = c/(kn)$. We show that there is a positive drift on the number of leading 1 bits. Let k be the length of the prefix of the current search point consisting of 1s.

We stick to our previous notation and denote by E_0 the event that the new search point has a longer prefix consisting of 1s and the comparison of old and new search point indicates correctly that the new search point is better. Let E_1 be the event that the new search point has a smaller number of leading ones than the current search point, and that it is accepted. Conditioning on E_1 we can trivially bound the expected number of leading ones below by 0. We have $P(E_0) \geq (1-p_k)/(en)$ and $P(E_1) \leq p_k(1-e^{-1}) \leq p_k$. Therefore, the expected increase in the number of leading 1s is

$$\begin{aligned} P(E_0) - kP(E_1) &\geq \frac{(1-p_k)}{en} - kp_k \\ &\geq \frac{1}{en} - p_k\left(\frac{1}{en} + k\right) \\ &\geq \frac{1}{en} - p_k \cdot (2k) \ . \end{aligned}$$

Due to our choice of $c < 1/6$ and our definition of p_k we have a positive additive drift of $1/(2en)$ leading to an upper bound of $O(n^2)$ for the expected run time of the algorithm by applying the Additive Drift Theorem (see Theorem 1). □

It is only to be expected that noise disrupts the optimization of LEADINGONES immensely. Consequently, our following corollaries to Theorem 11 are rather weak with respect to the noise allowed (basically, the algorithm will only experience constantly many incorrect decisions during optimization, in expectation).

COROLLARY 12. *Suppose prior noise which, with probability p, flips a bit uniformly at random. Then we have that the (1+1) EA optimizes* LEADINGONES *in time $O(n^2)$ if $p \leq 1/(3n^2)$.*

PROOF. Suppose first $p \leq 1/(3n^2)$ and let $k < n$. We estimate $P(X_k^{\text{opt}} < X_{k+1}^{\text{pes}})$ by observing that the event $X_k^{\text{opt}} \geq X_{k+1}^{\text{pes}}$ requires the individual with k 1s to be evaluated to $\geq k+1$ or the other to $\leq k$. The first option has a probability of at most p/n, the second of pk/n. Hence, $P(X_k^{\text{opt}} < X_{k+1}^{\text{pes}}) \geq 1-(k+1)/(3n^3) \geq 1-1/(3n^2)$ and the desired bound follows from Theorem 11. □

Regarding posterior noise we give the following corollary.

COROLLARY 13. *Suppose posterior noise, sampling from a centered distribution D with variance σ^2. Then we have that the (1+1) EA optimizes* LEADINGONES *in $O(n^2)$ if $\sigma^2 \leq 1/(12n^2)$.*

PROOF. Note that, in this case, $X_k^{\text{opt}} \sim X_k^{\text{pes}}$, for all $k \leq n$. With the same argument as in Corollary 7 we have $P(X_k^{\text{opt}} < X_{k+1}^{\text{pes}}) \geq 1-2\sigma^2$. Thus, for $\sigma^2 = 1/(12n^2)$ the claim follows from Theorem 11. □

Next we give a lower bound for the (1+1) EA for the prior noise model. We will not give a general lower bound that holds for both of our models because it is very easy for the (1+1) EA to detect an inferior noisy offspring by selection if LEADINGONES is subjected to posterior noise.

THEOREM 14. *Suppose prior noise which, with probability 1/2, flips a bit uniformly at random. Then we have that the (1+1) EA optimizes* LEADINGONES *in $2^{\Omega(n)}$ steps.*

PROOF. We show that there is a constant negative drift on the number of ones in the interval between $99n/100$ and n. Let $k \geq 99n/100$ be the number of 1s of the current search point.

Let E_0 be the event that the new search point has at least one more 1 than the current search point and the comparison of old and new search point indicates that the new search point is to be accepted. The expected number of 1s conditional on E_0 is at most $k+2$ and we have a trivial bound of $P(E_0) \leq (n-k)/n \leq 1/100$.

Let E_1 be the event that the new search point differs from the old by flipping exactly one 1 in the right half of positions, and that it is accepted. We want to estimate $P(E_1)$. There are at least $49n/100$ 1s in the right half of positions, so the probability of flipping exactly one of them and no other is at least $49/(100e)$. In order to estimate the probability of accepting such an offspring, we consider two cases. First, assume that the parent has a leading ones value of at least $n/2$. Then the probability of the noisy evaluation evaluating the parent to a value $< n/2$ is at least 1/4 (by choosing to flip a bit in the left half in the evaluation), while evaluating the offspring to its true value $\geq n/2$ has a probability of at least 1/2. In total we have $P(E_1) \geq 49/(800e)$ in this case.

Second, assume that the parent has a leading ones value of $< n/2$. Then both parent and offspring have the same leading ones value; with probability 1/4 they both evaluate to their true value, which favors the offspring. Thus, in this case, we get $P(E_1) \geq 49/(400e)$. Overall we have now $P(E_1) \geq 49/(800e) > 1/50$

Thus, we have that the total (negative) drift of

$$P(E_1) - 2P(E_0) \geq 49/(800e) - 2/100$$

which gives us a constant negative drift. Since long jumps are sufficiently small (due to our choice of using the number of 1s as potential), we can apply Theorem 3 which yields our result. □

4.2 Parent Populations

THEOREM 15. *Let μ be given and, for each $k < n$, let Y_k denote the minimum over μ observed values of X_k^{opt}. If there is a positive constant $c < 1/12$ such that*

$$\forall k : n/4 < k < n \Rightarrow P(Y_k < X_{k+1}^{\text{pes}}) \geq 1 - \frac{c}{\mu kn} \ , \qquad (4)$$

then the $(\mu+1)$ EA optimizes f in $O(\mu n^2)$ steps.

Note that, due to the dependence of Y_k on μ, Equation (4) typically gets less restrictive with growing μ, just as in Theorem 8.

PROOF. Let $p_k = c/(\mu kn)$. We show that there is a positive drift on the number of leading 1 bits of a current best individual. Let k be the length of the prefix of the current search point consisting of 1s. Let E_0 be the event that a best individual is improved by at least 1 and accepted by mutation. Let E_1 be the event that the new individual has a smaller prefix consisting of 1s and that the unique best individual is dropped from the parent population. We have that $P(E_0) \geq (1-p_k)/(e\mu n)$, assuming pessimistically that there is only one best individual. On the other hand $P(E_1) \leq p_k$.

Since, conditioned on E_1, the expected number of leading ones of a best individual is trivially at least 0 we can bound the expected increase in the number of leading 1s of a best individual below by

$$\begin{aligned} P(E_0) - kP(E_1) &\geq (1-p_k)\frac{1}{e\mu n} - kp_k \\ &\geq \frac{1}{e\mu n} - p_k\left(\frac{1}{e\mu n} + k\right) \\ &\geq \frac{1}{e\mu n} - p_k(2k) \ , \end{aligned}$$

and the last term can be bounded below by $1/(2e\mu n)$ due to our choice of $c < 1/12$ and the definition of p_k. Applying Theorem 1 yields an upper bound of $2e\mu n^2$ for the expected number of generations until the optimum is found. Taking the cost of initialization into account we have an expected run time of $\mu + 2e\mu n^2$, proving our claim. □

Theorem 15 is not strong enough to derive an upper bound for the prior noise model where a bit flip is performed with certain probability. This is because the probability that a best individual is dropped from the population is too high. Regarding posterior noise, we can still derive the following corollary.

COROLLARY 16. *Let any non-negative additive posterior noise be given which has a non-zero constant probability of evaluating to < 1. Then there is a constant c such that, for $\mu \geq c \log n$, the (μ+1) EA optimizes* LeadingOnes *in time $O(\mu n^2)$.*

PROOF. Let D be the posterior noise and $p = P(D < 1)$ a non-zero constant. Let $c = -3/\log(1-p)$ and $\mu = \lceil c \log n \rceil$. We have

$$P(Y_k < X_{k+1}) = 1 - P(D \geq 1)^{\mu} \geq 1-(1-p)^{-3\log n/\log(1-p)} \ ,$$

and the last term equals $1 - n^{-3}$ which yields the result by applying Theorem 15. □

4.3 Offspring Populations

In this section we consider the $(1 + \lambda)$ EA. The next theorem gives conditions for efficient optimization of stochastic LeadingOnes problems. The bound we give is the same as is shown for static LeadingOnes in [JDJW05].

THEOREM 17. *Let $\lambda \geq 72 \log n$ and, for each $k < n$, let Y_k denote the maximum over λ observed values of X_k^{opt} (belonging to inferior individuals) and let Z_k denote the maximum over at least $\lambda/6$ observed values of X_k^{pes} (belonging to better individuals). Suppose there are positive, non-zero constants $q < 1$ and $c < q/2$ such that*

$$\forall k < n : P(Y_k < X_{k+1}^{\mathrm{pes}}) \geq q \ , \tag{5}$$

and

$$\forall k < n : P(Y_{k-1} < Z_k) \geq 1 - \frac{c}{k(1+\frac{en}{\lambda})} \ . \tag{6}$$

Then the $(1+\lambda)$ EA optimizes f in $O(n+n^2/\lambda)$ generations and needs $O(n^2 + \lambda n)$ fitness evaluations.

PROOF. Let $p_k = c/(k(1 + \frac{en}{\lambda}))$. We show that there is a positive drift on the number of leading 1 bits. Let k be the length of the prefix of the current search point consisting of 1s.

Let E_0 be the event that at least one offspring is improved by at least 1 and is correctly accepted. Let E_1 be the event that at least one offspring has a smaller prefix consisting of 1s than the current search point, and that it is still accepted. We have that $P(E_0)$ can be bounded below by

$$q\left(1-\left(1-\frac{1}{en}\right)^{\lambda}\right) \geq q\frac{\lambda}{en+\lambda} \ .$$

The above inequality can be shown by induction on λ, see also [He10, Theorem 1].

In order to estimate $P(E_1)$ let E_2 be the event that more than $\lambda/6$ copied offspring are created. We have

$$P(E_1) \leq P(E_1|E_2)P(E_2) + P(E_1|\overline{E_2})P(\overline{E_2}) \ .$$

Equation (6) gives us $P(E_1|E_2) \leq p_k$. By using a Chernoff bound we further get $P(\overline{E_2}) \leq e^{-\lambda/24}$. Bounding the other probabilities by 1 we obtain $P(E_1) \leq p_k + e^{-\lambda/24}$. Conditioning on E_1, the expected number of leading ones is trivially at least 0. Therefore, the expected increase in the number of leading 1s is

$$\begin{aligned} P(E_0) - kP(E_1) &\geq q\frac{\lambda}{en+\lambda} - k\left(p_k + e^{-\frac{\lambda}{24}}\right) \\ &\geq q\frac{\lambda}{en+\lambda} - k\left(p_k + \frac{1}{n^3}\right) \\ &\geq q\frac{\lambda}{en+\lambda} - 2kp_k \\ &= (q-2c)\frac{\lambda}{en+\lambda} \ . \end{aligned}$$

Due to our choice of q and $c < q/2$ we have that $q - 2c$ is a positive constant. Applying Theorem 1 yields an upper bound of $O(n + en^2/\lambda)$ for the expected number of generations. Since λ fitness evaluations are performed in each generation the expected number of fitness evaluations until an optimum is found is therefore $O(\lambda n + n^2)$. □

We give the following corollary, showing the superiority of EAs with offspring populations over the (1+1) EA, even with only small populations.

COROLLARY 18. *Let p be bounded away from 0 and 1 by a constant. Suppose prior noise which, with probability p, flips a bit uniformly at random. Then there is a constant c such that, for λ with $c \log n \leq \lambda \leq 2c \log n$, we have that the $(1 + \lambda)$ EA optimizes* LeadingOnes *in time $O(\lambda n + n^2)$.*

PROOF. We give a rough sketch of the argument. We want to apply Theorem 17. It is easy to see that Equation (5) holds as, most likely, all the inferior individuals are not evaluated better than they are, and with constant probability the good individual is evaluated to its true value. Equation (6) follows since, given sufficiently many good individuals (in this case $\lambda/6$), there is a high probability that at least one of them evaluates to its true value, while all inferior individuals do not improve just as in our considerations regarding Equation (5). □

5. DISCUSSION

In this paper we consider the optimization of noisy versions of OneMax and LeadingOnes. The summary of the results is that populations are *necessary* for successful optimization for any substantial noise levels. The surprising result is that even very small populations (of size logarithmic in the problem size) already lead to very high robustness to noise (see Corollaries 9 and 18).

From the formal analysis we see the reason for this robustness: In a (parent) population of size μ, while the best individual might *look* bad in a given generation, there will surely be (objectively) worse individuals which also *look* worse. This holds as long as there are enough individuals in the parent populations to make sure that one of them will evaluate worse than the (objectively) best individual. For example, if a non-best individual will evaluate worse than the best individual with constant probability, a logarithmic number of non-best individuals is large enough to get very high confidence that such a bad individual is dropped. This observation probably extends to the analysis of the $(\mu + \lambda)$ EA with $\mu > 1$ *and* $\lambda > 1$.

As for offspring populations, in the $(1+\lambda)$ EA the current individual is cloned multiple times and thus hedges against bad evaluations (as long as good evaluations are sufficiently likely). This does not extend to the analysis of the $(\mu+\lambda)$ EA with $\mu > 1$ and $\lambda > 1$ in a straightforward way, as clones might be made from sub-optimal individuals.

Acknowledgements

The research leading to these results has received funding from the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement no 618091 (SAGE). The authors would like to thank Benjamin Doerr and Tobias Friedrich for many useful discussions on the topic. Furthermore, the reviewers of this paper gave a lot of helpful feedback, for which we are very grateful.

6. REFERENCES

[BDGG09] L. Bianchi, M. Dorigo, L. Gambardella, and W. Gutjahr. A Survey on Metaheuristics for Stochastic Combinatorial optimization. *Natural Computing*, 8:239–287, 2009.

[DHK12] B. Doerr, A. Hota, and T. Kötzing. Ants Easily Solve Stochastic Shortest Path Problems. In *Proc. of GECCO'12*, pages 17–24, 2012.

[DJW12] B. Doerr, D. Johannsen, and C. Winzen. Multiplicative Drift Analysis. *Algorithmica*, 64:673–697, 2012.

[Dro04] S. Droste. Analysis of the (1+1) EA for a Noisy OneMax. In *Proc. of GECCO'04*, pages 1088–1099, 2004.

[FK13] M. Feldmann and T. Kötzing. Optimizing Expected Path Lengths with Ant Colony Optimization Using Fitness Proportional Update. In *Proc. of FOGA'13*, pages 65–74, 2013.

[GP96] W. Gutjahr and G. Pflug. Simulated Annealing for Noisy Cost Functions. *Journal of Global Optimization*, 8:1–13, 1996.

[Gut03] W. Gutjahr. A Converging ACO Algorithm for Stochastic Combinatorial optimization. In *Proc. of SAGA'03*, pages 10–25, 2003.

[He10] Jun He. A Note on the First Hitting Time of $(1 + \lambda)$ Evolutionary Algorithms for Linear Functions with Boolean Inputs. In *Proc. of CEC'10*. IEEE, 2010.

[HY04] J. He and X. Yao. A Study of Drift Analysis for Estimating Computation Time of Evolutionary Algorithms. *Natural Computing*, 3:21–35, 2004.

[JB05] Y. Jin and J. Branke. Evolutionary Optimization in Uncertain Environments—A Survey. *IEEE Transactions on Evolutionary Computation*, 9:303–317, 2005.

[JDJW05] T. Jansen, K. De Jong, and I. Wegener. On the Choice of the Offspring Population Size in Evolutionary Algorithms. *Evolotionary Computation*, 13:413–440, 2005.

[Joh10] D. Johannsen. *Random Combinatorial Structures and Randomized Search Heuristics.* PhD thesis, Universität des Saarlandes, 2010. Available online at http://scidok.sulb.uni-saarland.de/volltexte/2011/3529/pdf/Dissertation_3166_Joha_Dani_2010.pdf.

[Leh11] P. K. Lehre. Fitness-levels for non-elitist populations. In *Proc. of GECCO'11*, pages 2075–2082, 2011.

[MRC08] B. Mitavskiy, J. E. Rowe, and C. Cannings. Preliminary Theoretical Analysis of a Local Search Algorithm to Optimize Network Communication Subject to Preserving the Total Number of Links. In *Proc. of CEC'08*, pages 1484–1491, 2008.

[OW11] P. Simone Oliveto and C. Witt. Simplified Drift Analysis for Proving Lower Bounds in Evolutionary computation. *Algorithmica*, 59:369–386, 2011.

[OW12] P. Oliveto and C. Witt. Erratum: Simplified Drift Analysis for Proving Lower Bounds in Evolutionary Computation. *arXiv*, abs/1211.7184, 2012.

[PB10] A. Prügel-Bennett. Benefits of a Population: Five Mechanisms That Advantage Population-Based Algorithms. *IEEE Transactions on Evolutionary Computation*, 14:500–517, 2010.

[RS12] J. E. Rowe and D. Sudholt. The Choice of the Offspring Population Size in the (1,λ) EA. In *Proc. of GECCO'12*, pages 1349–1356, 2012.

[ST12] D. Sudholt and C. Thyssen. A Simple Ant Colony Optimizer for Stochastic Shortest Path problems. *Algorithmica*, 64:643–672, 2012.

[Wit06] C. Witt. Runtime Analysis of the $(\mu + 1)$ EA on Simple Pseudo-Boolean Functions. *Evolutionary Computation*, 14:65–86, 2006.