Improved Runtime Bounds for the (1+1) EA on Random 3-CNF Formulas Based on Fitness-Distance Correlation

Benjamin Doerr
École Polytechnique
Université Paris-Saclay
Palaiseau, France

Frank Neumann
School of Computer Science
University of Adelaide
Adelaide, Australia

Andrew M. Sutton
Hasso-Plattner-Institut
Universität Potsdam
Potsdam, Germany

ABSTRACT

With this paper, we contribute to the theoretical understanding of randomized search heuristics by investigating their behavior on random 3-SAT instances. We improve the results for the (1+1) EA obtained by Sutton and Neumann [PPSN 2014, 942–951] in three ways. First, we reduce the upper bound by a linear factor and prove that the (1+1) EA obtains optimal solutions in time $O(n \log n)$ with high probability on asymptotically almost all high-density satisfiable 3-CNF formulas. Second, we extend the range of densities for which this bound holds to satisfiable formulas of at least logarithmic density. Finally, we complement these mathematical results with numerical experiments that summarize the behavior of the (1+1) EA on formulas along the density spectrum, and suggest that the implicit constants hidden in our bounds are low.

Our proofs are based on analyzing the run of the algorithm by establishing a fitness-distance correlation. This approach might be of independent interest and we are optimistic that it is useful for the analysis of randomized search heuristics in various other settings. To our knowledge, this is the first time that fitness-distance correlation is explicitly used to rigorously prove a performance statement for an evolutionary algorithm.

Categories and Subject Descriptors

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

Keywords

Runtime analysis; satisfiability; fitness-distance correlation

1. INTRODUCTION

The analysis of randomized search heuristics has made tremendous progress over the past fifteen years. A wide range of randomized search heuristics such as randomized local search, evolutionary algorithms, and ant colony optimization has been analyzed for specific fitness functions as well as problems from combinatorial optimization. We refer the reader to the timely books [19, 5, 12] for a comprehensive presentation.

Tail bounds on the runtime of the (1+1) EA on randomly generated high-density satisfiable 3-CNF formulas were derived in [23] where a bound of $O(n^2 \log n)$ was shown to hold with probability $1 - o(1)$. The main results of our paper is to improve this bound to $O(n \log n)$ and establish results for lower density formulas. We obtain this result by determining a fitness-distance correlation that allows a partitioning of the set of search points into different layers that correlate with the Hamming distance to an optimal solution.

The notion of fitness distance correlation (FDC) [13] has been widely used in the area of evolutionary computation to explain the difficulty of solving certain problems. All randomized search heuristics are guided by their fitness function and FDC is frequently used to determine the hardness of a problem by considering how the distance of search points to the optimum relates to their fitness values. The intuition is that problems are easy to solve if the fitness improves with decreasing distance to the optimum and hard to solve if the fitness is pointing in the opposite direction. While the intuition sounds sensible, it does not always translate directly into an accurate prediction of algorithm performance. Usually, the FDC is established by sampling search points on relatively small instances and calculating the empirical correlation between fitness and the distance to a known optimum. Different counterexamples have been presented in the literature that show FDC is not always a good predictor of algorithm performance (see e.g. [4, 11, 20]).

A strong FDC is only a reliable predictor if a randomized search heuristic does not encounter any deviations from the assumed usual behavior. In the case that a deviation from the predicted behavior becomes very unlikely, a strong FDC can potentially be used to accurately predict the runtime of randomized search heuristics. This property is explored in this paper, and we show that there is a strong FDC for highly dense 3-CNF formulas. Usually, FDC is determined empirically on a finite sample of points in the search space. In order to make it useful for upper bounds on the runtime of randomized search heuristics, we must be able to make rigorous statements about the properties of the relationship between fitness and distance and show that those properties hold with high probability. Furthermore, we require such statements to explicitly depend on the input size, where usual experimental investigations deal with problems

GECCO '15, July 11 - 15, 2015, Madrid, Spain

ISBN 978-1-4503-3472-3/15/07…$15.00

DOI: http://dx.doi.org/10.1145/2739480.2754659

of relatively small fixed size. We prove rigorous bounds on the FDC for high-density 3-CNF formulas in dependence of the input size n and use it to derive tail bounds on the runtime of the (1+1) EA for this class of problems. We show that the probability that the optimization time exceeds $O(n \log n)$ vanishes with increasing problem size for this problem class. This matches the lower bound on the expected runtime for asymptotically almost all high-density satisfiable 3-CNF formulas as described in [23].

We specifically study the *planted* model of random 3-CNF distributions and extend our results to the *filtered* model using a correspondence due to Ben-Sasson et al. [6]. Planted distributions for the maximum clique problem in graphs have also been studied by Storch [22] in the context of randomized search heuristics. In propositional satisfiability, the planted distribution of 3-CNF formulas is known to be easy to solve for classical algorithms [15], and our objective is to advance the theoretical analysis of evolutionary algorithms on random satisfiability models.

The outline of the paper is a follows. We introduce the model and algorithm under investigation in Section 2. We start our analysis by investigating formulas of high (linear) density in Section 3 and prove the $O(n \log n)$ bound. We then extend this analysis to formulas of logarithmic density. Finally, we give a short proof in Section 4 that the runtime of the (1+1) EA is faster than exponential for very low constant densities. Our theoretical results are complemented by experimental investigations in Section 5. We conclude the paper in Section 6.

2. PRELIMINARIES

A k-CNF formula F is constructed from a set of n Boolean variables $\{x_1, x_2, \ldots, x_n\}$ by forming a logical conjunction of exactly m clauses $F = C_1 \wedge C_2 \wedge \ldots \wedge C_m$. Each clause is the logical disjunction of exactly k literals, $C_i = \ell_{i_1} \vee \cdots \vee \ell_{i_k}$ and each literal ℓ_{i_j} is either an occurrence of a variable x or its negation $\neg x$. A k-CNF formula F is *satisfiable* if and only if there is an assignment of variables to truth values so that every clause contains at least one true literal.

The set of all assignments to a set of n Boolean variables is isomorphic to $\{0,1\}^n$ by interpreting the bit at each position i of the string as the state of exactly one Boolean variable x_i. For a length-m formula F on n variables, we define the fitness function $f = f_F \colon \{0,1\}^n \to \{0, \ldots, m\} := x \mapsto |\{C \in F : \ C \text{ is satisfied by } x\}|$. If F is satisfiable, the task of finding a satisfying assignment reduces to the task of maximizing f.

The standard (1+1) EA, illustrated in Algorithm 1, is a basic evolutionary algorithm that maintains a size-one population and produces a single offspring in each step. It can be characterized as a stochastic hill-climbing search that uses the standard bit-wise uniform mutation operator. Given a length-m formula F on n variables we seek an asymptotic bound on the runtime of the (1+1) EA searching for a satisfying assignment to F by optimizing the corresponding pseudo-Boolean function $f = f_F$. We study the infinite stochastic process $\{x^{(t)} : t \in \mathbb{N}_0\}$ on $\{0,1\}^n$ where $x^{(t)}$ is the assignment generated in iteration t of Algorithm 1. The runtime of the (1+1) EA is the random variable $T = \inf\{t \in \mathbb{N}_0 : f(x^{(t)}) = m\}$.

In order to bound the runtime of the (1+1) EA, we will consider the sequence $(x^{(0)}, x^{(1)}, \ldots)$ of assignments generated by the (1+1) EA and study the drift of corresponding stochastic processes that measure fitness values and distance values along this sequence. To make precise statements about the runtime, we rely heavily on the following drift theorem.

Algorithm 1: The (1+1) EA.

```
choose x ∈ {0,1}^n uniformly at random;
repeat forever
    y ← x;
    flip each bit of y independently with prob. 1/n;
    if f(y) ≥ f(x) then x ← y;
```

Theorem 1 (Multiplicative Drift [8, 9]). *Let $\{X_t : t \in \mathbb{N}_0\}$ be a sequence of random variables over $\mathbb{R}_{\geq 0}$. Let T be the random variable that denotes the earliest point in time $t \geq 0$ such that $X_t < 1$. If there exists $\delta > 0$ such that, for all a,*

$$\mathrm{E}(X_t - X_{t+1} \mid T > t, X_t = a) \geq \delta\, a,$$

then, for all a,

$$\mathrm{E}(T \mid X_0 = a) \leq \frac{1 + \ln(a)}{\delta},$$

and

$$\Pr\left(T > \frac{\lambda + \ln(a)}{\delta} \;\middle|\; X_0 = a\right) \leq e^{-\lambda} \text{ for all } \lambda > 0.$$

2.1 Random 3-CNF distributions

We consider distributions of 3-CNF formulas consisting of m clauses of length $k = 3$ over n variables. We also impose the assumption that each clause consists of distinct variables. This assumption is quite natural for 3-CNF formulas since any length-3 clause that contains repeating variables can be immediately reduced to an equivalent length-2 clause (if there are repeating variables with the same polarity) or a tautology (if there are repeating variables with opposite polarity). However, we do allow repeated clauses in F.

Definition 1. *Let $\Omega_{n,m}$ be the finite set of all 3-CNF formulas over n variables and m clauses.*

We associate random 3-CNF distributions with categorical distributions over the sample space $\Omega_{n,m}$. In particular, the well-known uniform distribution $\mathcal{U}_{n,m}$ is defined by

$$\Pr(F \mid F \sim \mathcal{U}_{n,m}) = |\Omega_{n,m}|^{-1}.$$

The filtered distribution is the uniform distribution conditioned on satisfiability.

$$\Pr(F \mid F \sim \mathcal{U}^{\mathrm{SAT}}_{n,m}) = |\{F \in \Omega_{n,m} : F \text{ is satisfiable}\}|^{-1}.$$

The planted distribution $\mathcal{P}_{n,m}$ is the uniform distribution conditioned on satisfiability by a planted assignment $x^\star$.

$$\Pr(F \mid F \sim \mathcal{P}_{n,m}) = |\{F \in \Omega_{n,m} : F \text{ is satisfied by } x^\star\}|^{-1}.$$

When considering a formula F constructed from $\mathcal{P}_{n,m}$, without loss of generality we will hereafter assume that the planted solution $x^\star = (1, 1, \ldots, 1)$ since the behavior of the (1+1) EA is invariant to negation operations on literals of F. We define the function $d \colon \{0,1\}^n \to \{0, \ldots, n\} := x \mapsto |\{i : x_i = 0\}|$ that measures the Hamming distance to the planted solution.

Definition 2. *Fix a constant $\epsilon > 0$. We define $\mathcal{H} = \mathcal{H}_\epsilon \subseteq \{0,1\}^n \times \{0,1\}^n$ such that $(x,y) \in \mathcal{H}$ if and only if*

1. $|\{i : x_i \neq y_i\}| = 1$,
2. $d(y) = d(x) - 1$, *and*
3. $d(x) \leq (1/2 + \epsilon)n$

The following lemma introduces a two-sided bound on the expected difference in fitness between pairs in $\mathcal{H}$, provided that F is drawn from the $\mathcal{P}_{n,m}$ distribution.

Lemma 1. *Let $(x,y) \in \mathcal{H}$, then*

$$\frac{3m}{7n}(1-\gamma(n)) \leq \mathrm{E}(f(y) - f(x) \mid F \sim \mathcal{P}_{n,m}) \leq \frac{3m}{7n}$$

where $\gamma(n) = (1+o(1))(3/4 + \epsilon - \epsilon^2)$.

Proof. Let A be the set of all 3-CNF clauses on n variables with at least one positive literal that are not satisfied by x but are satisfied by y. Similarly, let B be the set of all 3-CNF clauses on n variables with at least one positive literal that are satisfied by x but not satisfied by y. Let $C \supset A \cup B$ be the set of all 3-CNF clauses on n variables with at least one positive literal.

Suppose $F \sim \mathcal{P}_{n,m}$. Let Z_A be the random variable that counts the occurrences of clauses from A in F and Z_B be the random variable that counts the occurrences of clauses from B in F. Since F contains exactly m clauses chosen from C independently with replacement, $E(Z_A) = m|A|/|C|$, and $E(Z_B) = m|B|/|C|$. Hence,

$$\mathrm{E}(f(y) - f(x)) = \mathrm{E}(Z_A - Z_B) = m\left(\frac{|A| - |B|}{|C|}\right),$$

and the bounds follow from the fact that $|C| = 7\binom{n}{3}$ and Lemma 1 of [23], which states $|A| = \binom{n-1}{2}$ and $0 \leq |B| \leq \gamma(n)\binom{n-1}{2}$, where $\gamma(n)$ is as claimed. □

Taking the random variable $Z := Z_A - Z_B$ as the sum of m independent random variables, the following lemma follows from 1 by Chernoff bounds on Z.

Lemma 2. *For n sufficiently large, let $(x,y) \in \mathcal{H}$ be chosen arbitrarily. Then*

$$\Pr\left(\frac{c_1 m}{n} < f(y) - f(x) < \frac{c_2 m}{n} \;\middle|\; F \sim \mathcal{P}_{n,m}\right) = 1 - e^{-\Omega(m/n)}$$

for specific constants $c_1 < 1 - \gamma(n) < c_2$.

2.2 Constraint density

The *constraint density* of a formula is the ratio of clauses to variables m/n. The constraint density quantifies the average number of constraints that are imposed on a variable. Boolean formulas with low constraint density are expected to be easy to satisfy, since each variable has, on average, few constraints. On the other hand, formulas with high constraint density are, on average, easy to refute because backtracking search algorithms can quickly derive a contradiction. The study of a threshold phenomenon in the uniform random 3-CNF distribution $\mathcal{U}_{n,m}$ has been the focus of intense study in the last two decades. The *satisfiability threshold conjecture* [2] asserts that for all $k \geq 3$ if is a formula drawn uniformly at random from the set of all k-CNF formulas with n variables and m clauses, there exists a real number r_k such that

$$\lim_{n\to\infty} \Pr\{F \text{ is satisfiable}\} = \begin{cases} 1 & m/n < r_k; \\ 0 & m/n > r_k. \end{cases}$$

Experimental studies on 3-CNF formulas suggest a threshold around $r_3 \approx 4.26$. There are currently no exact results for the location of this threshold (if it exists), and only upper and lower bounds are known. For a more detailed treatment of random satisfiability along with an exposition of recent developments, see the chapter by Achlioptas [1].

3. HIGH-DENSITY REGIME

We now study the runtime of the (1+1) EA on high-density planted formulas. We begin with linear densities in Section 3.1, namely, length-m formulas on n variables where $m/n \geq cn$ for a specific constant c. In this regime we prove that for asymptotically almost all formulas, the (1+1) EA finds a satisfying assignment in $O(n \log n)$ time with high probability. This improves by a linear factor the previous known tail bounds for the (1+1) EA at these densities [23].

In Section 3.2, we consider sparser formulas where $m/n \geq c \log n$ for a particular c. We treat these densities separately because the randomness of both the algorithm and the formula sampling process must be handled more carefully.

3.1 Linear density

Definition 3. *We say a formula F has* strong FDC *if the following two properties hold.*

Property A. $\forall (x,y) \in \mathcal{H}$, $c_1 m/n < f(y) - f(x) < c_2 m/n$,

Property B. $\forall x, y \in \{0,1\}^n$ *with* $n/2 + \epsilon n \geq d(x) \geq n/2 + 3\epsilon n/4$ *and* $y \leq n/2 + \epsilon n/2$, $f(x) < f(y)$.

Here $c_1 < c_2$ are the constants introduced in Lemma 2.

Lemma 3. *Let $F \sim \mathcal{P}_{n,m}$ where $m/n \geq cn$ for a sufficiently large positive constant c. The probability (taken on $\Omega_{n,m}$) that F has strong FDC is at least $1 - e^{-\Omega(n)}$.*

Proof. By Lemma 2 together with a union bound over the elements of $\mathcal{H}$, Property A of Definition 3 holds with probability $1 - e^{-\Omega(n)}$. For Property B, let $z, z' \in \{0,1\}^n$ such that $d(z) = d(z') = i$. Note that $E_i = \mathrm{E}(f(z)) = \mathrm{E}(f(z'))$ is independent of z, where we take the expectation over $\mathcal{P}_{n,m}$. Let $u, v \in \{0,1\}^n$ where $n/2 + \epsilon n \geq d(u) \geq n/2 + 3\epsilon n/4$ and $d(v) \leq n/2 + \epsilon n/2$. Denote $a := d(u)$ and $b := d(v)$. Note that u can be transformed to v by changing $a - b \geq \epsilon n/4$ zeros to ones. We argue that $E_b \geq E_a + \Theta(m)$. By the above stated independence, $\mathrm{E}(f(v) - f(u)) = E_b - E_a$. Furthermore, by a repeated application of Lemma 1, we have $E_b \geq E_a + (3/7)(1 - \gamma(n))(m/n)(\epsilon n/4) = E_a + \Theta(m)$.

Let $q := (E_a + E_b)/2$ and let y be any search point with $d(y) = b$. Note that $f(y)$ is a random variable that can be written as sum of m independent 0/1 random variables. Consequently, the additive Chernoff bound shows that

$$\Pr(f(y) \leq q) = \Pr(f(y) \leq E_b - (E_b - E_a)/2) \leq e^{-\Theta(m)}.$$

The same argument shows that any x with $d(x) = a$ has a fitness greater than q with probability $e^{-\Theta(m)}$ only. Applying a union bound over the applicable pairs $x, y \in \{0,1\}^n$, we conclude that Property B of Definition 3 holds with probability at least $1 - e^{-\Omega(n)}$. A final union bound over both properties concludes the proof. □

Theorem 2. *Let $m/n \geq cn$ for a sufficiently large positive constant c. The runtime of the (1+1) EA is bounded by $O(n \log n)$ with probability $1-o(1)$ on every planted formula with n variables and m clauses except for a set of measure tending to zero exponentially fast with n.*

Proof. By Lemma 3, every planted formula at density at least cn has strong FDC except for an $e^{-\Omega(n)}$-fraction, so we will assume for the remainder of the proof that we are working with a formula that has the strong FDC property.

If F has strong FDC, then for states that are not too far away from the planted assignment, the fitness and distance are tightly correlated in the following sense. $\forall\, (x, y) \in \mathcal{H}$,

$$f(x) + c_2 d(x) m/n \geq m, \text{ and } f(x) + c_1 d(x) m/n \leq m \quad (1)$$

since each step on a path of length $d(x)$ to 1^n increases the fitness by at least (respectively at most) a constant number multiplied by m/n and $f(1^n) = m$.

We consider the drift of the stochastic process $\{X_t : t \in \mathbb{N}_0\}$ where $X_t = m - f(x^{(t)})$. Assume at iteration t that $0 < d(x^{(t)}) \leq (1/2 + \epsilon)n$ (we will later show this holds with high probability over the run). There must be a set $S \subseteq \{0,1\}^n$ consisting of $d(x^{(t)})$ elements such that for each $y \in S$, $(x^{(t)}, y) \in \mathcal{H}$. For each such y, since $f(y) > f(x^{(t)}) + c_1 m/n > f(x^{(t)})$, a mutation from $x^{(t)}$ to y is clearly accepted by selection. Furthermore, selection does not accept mutation to lower fitness values so $X_t - X_{t+1} \geq 0$. Let $\mathcal{E}$ denote the event that mutation produces some $y \in S$ from $x^{(t)}$. By the law of total expectation,

$$\begin{aligned}\mathrm{E}(X_t - X_{t+1} \mid X_t) &\geq \mathrm{E}(X_t - X_{t+1} \mid X_t, \mathcal{E}) \Pr(\mathcal{E}) \\ &\geq \mathrm{E}(X_t - X_{t+1} \mid X_t, \mathcal{E}) \frac{d(x^{(t)})}{en}.\end{aligned}$$

By the inequality in (1),

$$\frac{m - f(x^{(t)})}{c_2 m/n} = \frac{X_t}{c_2 m/n} \leq d(x^{(t)})$$

so we can bound the drift as

$$\begin{aligned}\mathrm{E}(X_t - X_{t+1} \mid X_t) &\geq \mathrm{E}(X_t - X_{t+1} \mid X_t, \mathcal{E}) \frac{X_t}{en(c_2 m/n)} \\ &= (f(y) - f(x^{(t)})) \frac{X_t}{en(c_2 m/n)},\end{aligned}$$

and, since F has strong FDC and $(x^{(t)}, y) \in \mathcal{H}$,

$$\geq X_t \frac{c_1 m/n}{en(c_2 m/n)} = X_t \frac{c_1/c_2}{en}. \quad (2)$$

We only need to show that with high probability, the process never leaves $\mathcal{H}$. Using the multiplicative Chernoff bound, the initial search point generated uniformly at random has $d(x^{(0)}) \leq n/2 + \epsilon n/2$ with high probability. In this case, by Property B of Definition 3, the EA can never reach a search point with distance b or worse in $\mathcal{H}$. Since $\mathcal{H}$ by definition contains points at distance at most $(1/2 + \epsilon)n$, in order for the process to leave $\mathcal{H}$, it must jump over the gap between $n/2 + 3\epsilon n/4$ and $n/2 + \epsilon n$. This can only occur after mutating at least $\epsilon n/4$ bits: an event that occurs with probability at most $e^{-\Omega(n \log n)}$ under uniform mutation.

We thus assume that the process does not leave $\mathcal{H}$, and so the inequality of (2) is valid for all times t. Finally, we apply Theorem 1 using inequality (2) by setting $\delta = c_1/(c_2 en)$ and $\lambda = \log n$ to obtain the tail bound. □

Let F^* be any formula on n variables and m clauses with exactly one satisfying assignment. Ben-Sasson et al. [6] proved that for densities above $m/n > c \ln n$ for a particular constant c, the probability of generating F^* from $\mathcal{P}_{n,m}$ is asymptotically equal to the probability of generating F^* from the filtered distribution $\mathcal{U}^{\mathrm{SAT}}_{n,m}$. They also prove that with high probability in this regime, a formula from either distribution has a unique satisfying assignment. Thus we can extend our result to the uniform distribution conditioned on satisfiability to obtain the following corollary that covers all satisfiable formulas in the high-density regime.

Corollary 1. *Let $m/n > cn$ for a sufficiently large positive constant c. The runtime of the (1+1) EA is bounded by $O(n \log n)$ with probability $1 - o(1)$ on all satisfiable 3-CNF formulas on n variables and m clauses except for a set of measure tending to zero exponentially fast with n.*

We conclude the typical runtime of the (1+1) EA very rarely deviates above $O(n \log n)$ for asymptotically almost all satisfiable formulas of sufficiently high linear density. This complements the $\Omega(n \log n)$ lower bound on the *expected* runtime derived for the same class of formulas in [23, Theorem 6]. A corresponding upper bound on the expected runtime is trickier since there is still a very low probability that the (1+1) EA can escape $\mathcal{H}$ and become trapped for a long time at a local optimum. Our results yield a simple solution to such a heavy-tailed runtime: perform sufficiently frequent restarts and apply Theorem 2 or Corollary 1 to obtain a matching upper bound on the expectation.

3.2 Logarithmic density

For smaller densities we can also obtain a similar tail bound on the runtime, but we have to take a slightly different approach. At high densities, Theorem 2 makes a statement about the runtime over all but a vanishing fraction of formulas. At densities asymptotically lower, we can make a statement about the likelihood of a runtime of $O(n \log n)$, but the probability is taken over both the dynamics of the (1+1) EA process and the sampling of the random formula.

For any $x \in \{0,1\}^n$ with $d(x) = k$ we define $P(x) := (x = x_1, x_2, \ldots, x_k = 1^n)$ to be the unique path where x_{i+1} is constructed from x_i by flipping the leftmost zero bit of x.

Lemma 4. *Let c be a sufficiently large positive constant and let $F \sim \mathcal{P}_{n,m}$ with $m/n > c \ln n$. If $d(x) \leq (1/2 + \epsilon)n$, then for every $x_i, x_{i+1} \in P(x)$, $c_1 m/n \leq f(x_{i+1}) - f(x_i) \leq c_2 m/n$ with probability at least $1 - n^{-3}$.*

Proof. Define the indicator random variable $\chi\colon \{0,1\}^n \times \{0,1\}^n \times \Omega_{n,m} \to \{0,1\}$ where

$$\chi(x, y; F) = \begin{cases} 1 & \text{if } \frac{c_1 m}{n} < f(y) - f(x) < \frac{c_2 m}{n}, \\ 0 & \text{otherwise.} \end{cases}$$

For all $x_i, x_{i+1} \in P(x)$, by the union bound

$$\begin{aligned}&\Pr\left(\bigcup_{x_i, x_{i+1} \in P(x^{(t)})} \chi(x_i, x_{i+1}; F) = 0 \;\middle|\; F \sim \mathcal{P}_{n,m} \right) \\ &\leq \sum_{x_i, x_{i+1} \in P(x^{(t)})} \Pr\left(\chi(x_i, x_{i+1}; F) = 0 \mid F \sim \mathcal{P}_{n,m}\right) \\ &\leq n e^{-\Omega(m/n)} \leq n^{-3},\end{aligned}$$

where we have applied Lemma 2 and used the fact that $m > cn \ln n$, c sufficiently large. □

Lemma 4 bounds the probability that the fitness and distance are sufficiently correlated along a path to the planted solution on a random formula of density at least $c \ln n$ for a specific constant c. We now apply this result to derive a bound on the runtime of the (1+1) EA over such a formula.

Theorem 3. *Let $F \sim \mathcal{P}_{n,m}$, $m/n > c \ln n$ for a sufficiently large positive constant c. The runtime of the (1+1) EA is bounded by $O(n \log n)$ with probability polynomially close to one, where the probability is taken over both the random sampling of the formula and the random optimization time.*

PROOF. Let r be a sufficiently large constant. We consider the first $rn \ln n$ steps of the (1+1) EA on a formula F drawn from $\mathcal{P}_{n,m}$ uniformly at random and show that, with probability polynomially close to one, the (1+1) EA has found a satisfying assignment.

We say a bitstring $x \in \{0,1\}^n$ has an *FDC path* if the following two properties hold at x.

Property A′. $\forall x_i, x_{i+1} \in P(x)$, $c_1 m/n \leq f(x_{i+1}) - f(x_i) \leq c_2 m/n$.

Property B′. $d(x) \leq n/2 + 3\epsilon n/4$,

We say the process has *failed* at time t if there is any $0 \leq s \leq t$ such that $x^{(s)}$ does not have an FDC path. We argue by induction that the probability that the process fails at time t conditioned on the event that it has not failed at time $t-1$ is sufficiently high. Specifically, in each step we show that if process has not yet failed, Property B′ holds with probability $1 - e^{-\Omega(n)}$, and the probability of Property A′ conditioned on B′ is polynomially close to one.

For the initial point, we assume a slightly stronger condition than Property B′: that $d(x^{(0)}) \leq n/2 + \epsilon n/2$, which by Chernoff bounds holds with probability exponentially close to one. Let $h = f(x^{(0)})$. If $t > 0$ and the process has not yet failed at time $t-1$, then Property B′ can be violated only if the (1+1) EA accepts a point sufficiently further away from the planted solution. Let y be the offspring generated in iteration t of the (1+1) EA. We make a case distinction on three disjoint events occurring in the mutation step.

Case $d(y) > n/2 + \epsilon n$. This event occurs only with probability $e^{-\Omega(n \log n)}$ since Property B′ holds at $x^{(t-1)}$, and so $\epsilon n/4$ bits must change during mutation to produce y.

Case $n/2 + \epsilon n \geq d(y) > n/2 + 3\epsilon n/4$. By an argument similar to the one in the proof of Lemma 3 that Property B of Definition 3 holds at high densities, we have $f(y) \geq h$ only with $e^{-\Theta(m)}$ probability. Since the fitness of points can only monotonically increase during a run of the (1+1) EA, under this event $f(y) < h \leq f(x^{(t-1)})$ and so $x^{(t)} = x^{(t-1)}$ since y would not be accepted.

Case $d(y) \leq n/2 + 3\epsilon n/4$. In this case, Property B′ will also not be violated by $x^{(t)}$ because, $x^{(t)} = y$ or $x^{(t)} = x^{(t-1)}$, both of which satisfy Property B′.

Since the first two events occur with exponentially small probability (and the events partition the probability space), we can conclude that, as long as the process has not failed by iteration $t-1$, then with probability $1 - e^{-\Omega(n)}$, Property B′ holds at $x^{(t)}$.

Assuming Property B′ holds at $x^{(t)}$, we can apply Lemma 4 and conclude the conditional probability that Property A′ also holds at $x^{(t)}$ is at least $1 - n^{-3}$. The joint probability that both properties hold is therefore at least $1 - O(n^{-3})$.

Now by induction on t, the process does not fail for $rn \ln n$ steps with probability at least $1 - O(n^{-1})$. Conditioning on the event that the process has not failed for a phase of $rn \ln n$ steps, the bound on the drift during this phase is the same as with inequality (2). We can again apply Theorem 1 with $\lambda = \log n$ to obtain the tail bound. □

4. LOW-DENSITY REGIME

On the uniform distribution, 3-CNF formulas that seem to be difficult for complete search algorithms lie near the proposed critical threshold $r_3 \approx 4.26$. However, at very low densities, random formulas become easy to solve again, even by very simple backtracking-free heuristics. The *pure literal heuristic* operates by iteratively finding a *pure literal* in the formula (i.e., one whose negation does not appear), setting it to true, and then removing all clauses that contains it. This approach succeeds with high probability for uniform random formulas at constraint densities $m/n < 1.637$ [17, 18]. A similar backtracking-free heuristic called the *generalized unit clause heuristic* succeeds with asymptotically positive probability on uniform random formulas with $m/n < 3.003$ [10]. Alekhnovich and Ben-Sasson [3] discovered a deep connection between constraint-directed random walk (iteratively flipping a random variable in a random unsatisfied clause) and the pure literal heuristic. They proved that the random walk finds a satisfying assignment with high probability in linear time for constraint densities at most 1.637.

In the interest of a more complete picture, we would also like to understand the behavior of the (1+1) EA at very sparse densities. Such formulas are likely easy to solve because most assignments to a random formula are already satisfying, and we conjecture that the (1+1) EA also runs in polynomial time at these densities. This conjecture is strongly supported by empirical evidence in Section 5, however proving the conjecture is likely to require different techniques than the ones that are useful for high density formulas. In this section, we show that if the density is low enough, the structure of constraints is so sparse that the formula breaks up into small components that the (1+1) EA can solve separately. From this we easily derive a $2^{o(n)}$ subexponential time bound and at least conclude that the runtime for the (1+1) EA is faster than any exponential function at low densities.

Lemma 5. *Let $H = H_d(n, m)$ denote a random d-uniform hypergraph with n vertices and exactly m hyperedges selected uniformly at random with replacement from the family of $\binom{n}{d}$ possible d-sets.*

Let $\alpha = dm/n$ denote the average degree *of H. If $\alpha < (d-1)^{-1}$, then with high probability, the number of vertices in the largest connected component of H is $O(\log n)$.*

PROOF. We consider m rounds of selecting edges uniformly at random with replacement. Let $X_1, X_2, \ldots, X_m$ denote the sequence of random variables where X_i corresponds to the size of the largest connected component in round i. Moreover, let $\widehat{X}_1, \widehat{X}_2, \ldots, \widehat{X}_m$ be the same sequence corresponding to the process of selecting edges uniformly at random *without* replacement. Note that for all $1 \leq i < m$, $\Pr(X_{i+1} > X_i \mid X_i) \leq \Pr(\widehat{X}_{i+1} > \widehat{X}_i \mid \widehat{X}_i)$ since if we use replacement sampling, there are only more chances to add a pre-existing edge, which has no effect on the size of any

component. Therefore X_m is stochastically dominated by $\widehat{X}_m$ and so for every $k \in \mathbb{N}$, $\Pr(X_m < k) \geq \Pr(\widehat{X}_m < k)$. Finally the result that $\Pr(\widehat{X}_m = O(\log n)) = 1 - o(1)$ is due to Schmidt-Pruzan and Shamir [21]. □

The *constraint hypergraph* of a formula is a hypergraph $H = (X, E)$ where X corresponds to the set of variables in F and E is a sequence of m nonempty subsets of X constructed as follows. Each clause C of F corresponds to a unique $S \in E$ that contains exactly the variables that appear as literals in C. Thus every 3-CNF formula on n variables with m clauses has a unique 3-regular constraint hypergraph with m hyperedges (parallel hyperedges are allowed). It is easy to see that at very low constant densities, the constraint structure of Boolean formulas breaks up into small components that the (1+1) EA can solve separately. This is captured by the following theorem.

Theorem 4. *Let F be a 3-CNF formula drawn from $\mathcal{U}_{n,m}$ with density $m/n < 1/6$. Then with high probability the (1+1) EA finds a satisfying assignment for F in subexponential time.*

Proof. We consider the average degree α of the constraint hypergraph H of F. Since F is sampled uniformly at random from $\Omega_{n,m}$, its constraint hypergraph is a random 3-uniform hypergraph with n vertices and m edges sampled uniformly at random with replacement since each of the 2^3 distinct clauses associated with each unique 3-set is also selected uniformly at random. Since $\alpha = 3m/n < 1/2$, by Lemma 5, with high probability the largest connected component in H contains $O(\log n)$ vertices.

In this case, let q be the number of connected components in H. We partition the clause set $\{S_1, S_2, \ldots, S_q\}$ where S_i is the set of clauses that contain only variables from the i-th connected component of H. The fitness function f can be expressed as $f(x) = \sum_{i=1}^{q} f_i(x)$ where $f_i(x)$ counts the number of clauses in S_i that are satisfied by x. Since each f_i depends on at most $O(\log n)$ bits of x, f is decomposable into linearly separable subfunctions of bounded size.

The proof is then completed by a simple fitness levels argument [24]. In particular, let $(A_0, \ldots, A_m)$ be a partition of $\{0,1\}^n$ such that for all $x \in A_j$, $f(x) = j$. Let t be an arbitrary iteration in the execution of the (1+1) EA and set $k := f(x^{(t)})$. As long as there is an unsolved subfunction f_i with respect to the assignment corresponding to $x^{(t)}$, the (1+1) EA can generate a strictly improving offspring by solving f_i and flipping no other bit outside of S_i. The resulting offspring must lie in some A_ℓ with $\ell > k$. The probability of this event is at least $(1-1/n)^{n-|S_i|}(1/n)^{|S_i|} \geq n^{-|S_i|}/e$, and the waiting time to increase the fitness level by at least one is bounded by $en^{|S_i|}$. Since there are at most $m = O(n)$ suboptimal fitness levels, the expected time until F is solved is bounded by $n^{O(\log n)} = 2^{o(n)}$. □

5. EXPERIMENTS

In this section we report numerical experiments that investigate the constants in the asymptotic bounds proved in this paper, and to explore the runtime character of the (1+1) EA at lower densities. In Figure 1(a) we investigate the runtime divided by $n \ln n$ as a function of $n = 10, 20, \ldots, 1000$ for the $\mathcal{P}_{n,m}$ model with $m/n = n$. For each value if n we generate 100 random 3-CNF formulas, and conduct 100 runs of the (1+1) EA, measuring the first iteration in which it finds a satisfying assignment. We then calculate the quartiles of the number of iterations to solve the formula at each value of n as a robust statistic for the runtime as a function of n. The plot is converging to a constant near e, providing empirical evidence that the runtime bound proved in this paper is tight, and suggests that the true runtime is concentrated around $en \ln n \pm O(n)$. We repeat this experiment for asymptotically lower densities and plot the results in Figure 1(b). In this case, we set $m/n = c \ln n$ for each random formula corresponding to the statement of Theorem 3. We determined $c = 4$ to be a sufficiently high constant, meaning that the (1+1) EA would not always converge to a solution on densities for $c < 4$. On the other hand, convergence was always observed for densities above $4 \ln n$. The behavior in 1(b) is very similar to the linear density case, and the true runtime appears to be concentrated around $en \ln n \pm O(n)$ for formulas of logarithmic densities.

5.1 Phase transition behavior

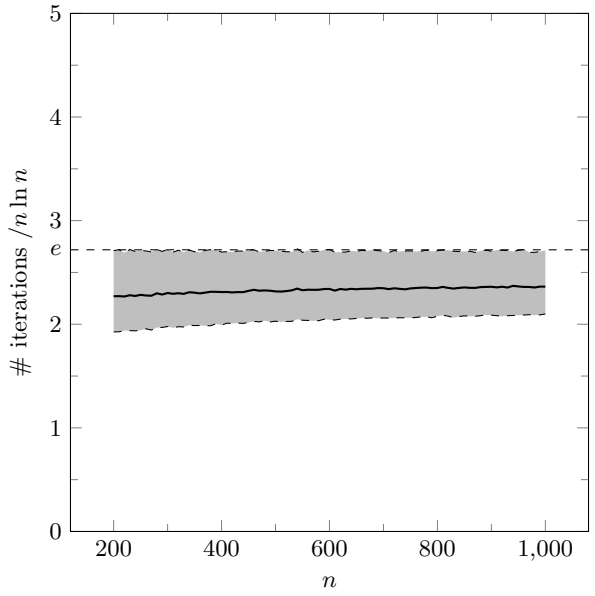
As discussed in Section 4, we conjecture that the (1+1) EA can also easily solve 3-CNF formulas at very low densities. To more precisely understand the behavior of the (1+1) EA on random planted 3-CNF formulas across the density spectrum, we performed numerical experiments and measured the time until a satisfying assignment was found at different densities for some distinct values of n.

On the $\mathcal{P}_{n,m}$ model, for three distinct values of n, i.e., $n \in \{100, 300, 1000\}$, we generate formulas using 100 equidistant values of m such that the constraint density ranges from 1 to 10. For each distinct density value, we generate 100 formulas from the random $\mathcal{P}_{n,m}$ model and run the (1+1) EA 100 times on each formula. Runs that do not complete in at most 10^7 iterations are halted and removed from consideration. Of the runs that do not fail, the median runtime as a function of constraint density for these trials is plotted in Figure 2(a). We also plot the percentage of runs that failed as a function of constraint density in Figure 2(b).

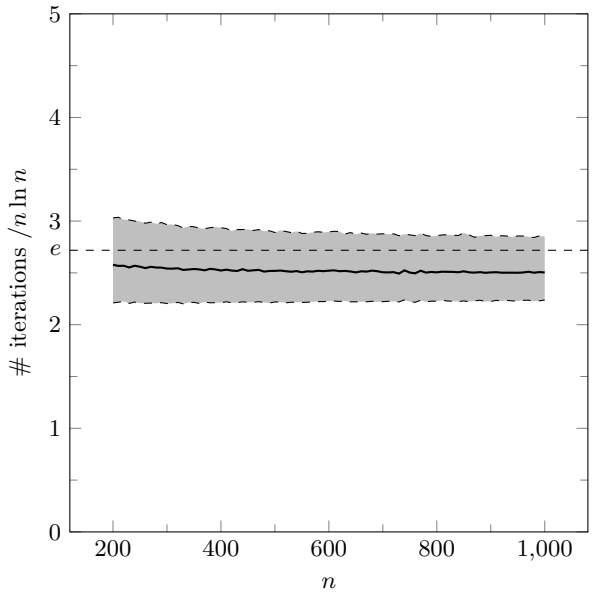
In these results, we also observe the classical easy-hard-easy pattern similar to the one that occurs for complete DPLL solvers on the uniform random model [16, 7]. This corresponds to the so-called *phase transition* phenomenon in random satisfiability where formulas near the sat/unsat transition have high decision complexity [14].

Remarkably, our experiments suggest that there is also a critical density in the *planted* model $\mathcal{P}_{n,m}$ for the (1+1) EA at which formulas are on average more difficult to optimize. We also observe that the hardness peak for the (1+1) EA occurs close to density values of $m/n \approx 4.26$, which is the critical density for DPLL solvers on the uniform model $\mathcal{U}_{n,m}$. This corresponds to the conjectured satisfiability threshold r_3 for random unfiltered, unplanted formulas.

Below the hardness peak, the (1+1) EA finds a satisfying assignment quickly, and we conjecture that there exists a constant $c < 4.26$ such that the (1+1) EA runs in polynomial time with high probability on random satisfiable formulas with density at most c. As density increases beyond the critical point, the empirical running time in Figure 2(a) appears to converge again toward $en \ln n$ for each n value. Theorem 3 establishes an asymptotic bound on the density at which most formulas become easy again. An interesting open problem is the location of the critical density below which formulas become difficult on average for the (1+1) EA.

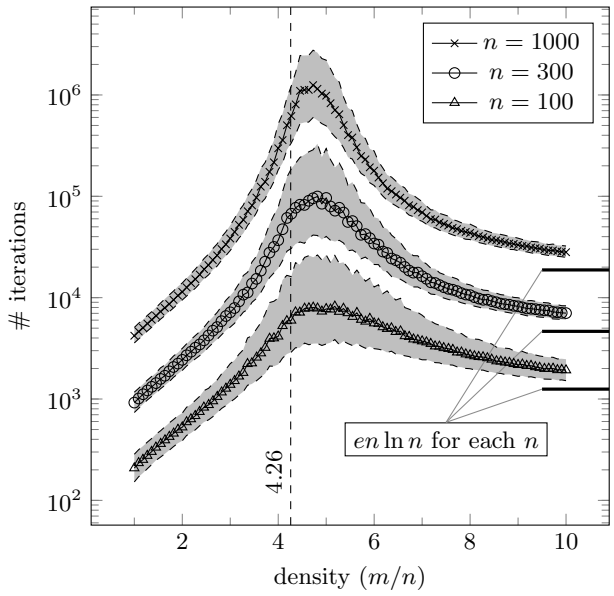


(a) $m = n^2$ (constraint density is $\Theta(n)$).

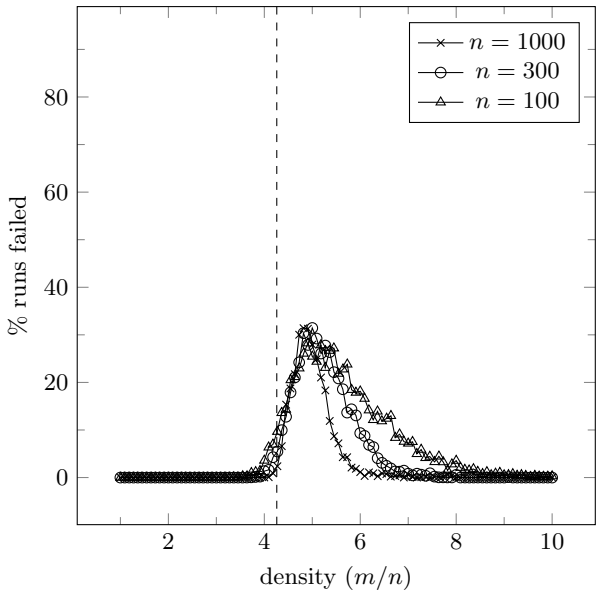


(b) $m = 4n \ln n$ (constraint density is $\Theta(\log n)$).

Figure 1: Median runtime of the (1+1) EA divided by $n \ln n$ as a function of n for the $\mathcal{P}_{n,m}$. The shaded region denotes the interquartile range. The statistics are taken from 100 runs each on 100 random formulas generated for each value of n.



(a) Runtime statistics. The marked lines denote the median runtime. The shaded regions denote the interquartile range.



(b) Percentage of runs (out of 10000) requiring $\geq 10^7$ iterations at each density value.

Figure 2: Results for the (1+1) EA on the $\mathcal{P}_{n,m}$ model controlling m for constraint density. The statistics are taken from 100 runs each on 100 random formulas generated for each value of m/n.

6. CONCLUSIONS

We have presented an improved runtime analysis of the (1+1) EA for randomly constructed 3-CNF formulas. Investigating the fitness distance correlation for high density formulas, we have shown an improved bound of $O(n \log n)$ on the (1+1) EA. In extension to the investigations in [23] , the $O(n \log n)$ bound holds for formulas of logarithmic density with probability $1 - o(1)$. Our complementary experimental investigations imply the leading constants in our asymptotic bounds are low, and extend the investigations to other density ratios.

Acknowledgements

The research leading to these results has received funding from the Australian Research Council (ARC) under grant agreement DP140103400 and from the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement no 618091 (SAGE).

7. REFERENCES

[1] Dimitris Achlioptas. Random satisfiability. In Armin Biere, Marijn Heule, Hans van Maaren, and Toby Walsh, editors, *Handbook of Satisfiability*, volume 185 of *Frontiers in Artificial Intelligence and Applications*, pages 245–270. IOS Press, 2009.

[2] Dimitris Achlioptas, Amin Coja-Oghlan, and Federico Ricci-Tersenghi. On the solution-space geometry of random constraint satisfaction problems. *Random Structures and Algorithms*, 38(3):251–268, 2011.

[3] Mikhail Alekhnovich and Eli Ben-Sasson. Linear upper bounds for random walk on small density random 3-CNFs. *SIAM Journal on Computing*, 36(5):1248–1263, 2007.

[4] Lee Altenberg. Fitness distance correlation analysis: An instructive counterexample. In Thomas Bäck, editor, *Proceedings of the Seventh International Conference on Genetic Algorithms*, pages 57–64. Morgan Kaufmann, 1997.

[5] Anne Auger and Benjamin Doerr. *Theory of Randomized Search Heuristics: Foundations and Recent Developments*. World Scientific Publishing Co., Inc., 2011.

[6] Eli Ben-Sasson, Yonatan Bilu, and Danny Gutfreund. Finding a randomly planted assignment in a random 3-CNF. Manuscript, 2002.

[7] James M Crawford and Larry D Auton. Experimental results on the crossover point in satisfiability problems. In *Proceedings of the Eleventh National Conference on Artificial Intelligence (AAAI-93)*, pages 21–27, 1993.

[8] Benjamin Doerr and Leslie Ann Goldberg. Adaptive drift analysis. *Algorithmica*, 65(1):224–250, 2013.

[9] Benjamin Doerr, Daniel Johannsen, and Carola Winzen. Multiplicative drift analysis. *Algorithmica*, 64(4):673–697, 2012.

[10] Alan Frieze and Stephen Suen. Analysis of two simple heuristics on a random instance of k-SAT. *Journal of Algorithms*, 20(2):312–355, 1996.

[11] Thomas Jansen. On classifications of fitness functions. In Leila Kallel, Bart Naudts, and Alex Rogers, editors, *Theoretical Aspects of Evolutionary Computing*, Natural Computing Series, pages 371–385. Springer Berlin Heidelberg, 2001.

[12] Thomas Jansen. *Analyzing Evolutionary Algorithms - The Computer Science Perspective*. Natural Computing Series. Springer, 2013.

[13] Terry Jones and Stephanie Forrest. Fitness distance correlation as a measure of problem difficulty for genetic algorithms. In Larry J. Eshelman, editor, *Proceedings of the Sixth International Conference on Genetic Algorithms*, pages 184–192. Morgan Kaufmann, 1995.

[14] Scott Kirkpatrick and Bart Selman. Critical behavior in the satisfiability of random Boolean expressions. *Science*, 264(5163):1297–1301, 1994.

[15] Michael Krivelevich and Dan Vilenchik. Solving random satisfiable 3CNF formulas in expected polynomial time. In *Proceedings of the Seventeenth Symposium on Discrete Algorithms (SODA'06)*, pages 454–463, 2006.

[16] David Mitchell, Bart Selman, and Hector Levesque. Hard and easy distributions of SAT problems. In *Proceedings of the Tenth National Conference on Artificial Intelligence (AAAI-92)*, pages 459–465, 1992.

[17] Michael Mitzenmacher. Tight thresholds for the pure literal rule. Technical Report 1997-011, Digital SRC, 1997.

[18] Michael Molloy. Cores in random hypergraphs and Boolean formulas. *Random Structures and Algorithms*, 27(1):124–135, 2005.

[19] Frank Neumann and Carsten Witt. *Bioinspired Computation in Combinatorial Optimization: Algorithms and Their Computational Complexity*. Springer, 2010.

[20] R. J. Quick, Victor J. Rayward-Smith, and G. D. Smith. Fitness distance correlation and ridge functions. In A. E. Eiben, Thomas Bäck, Marc Schoenauer, and Hans-Paul Schwefel, editors, *Proceedings of the Fifth International Conference on Parallel Problem Solving from Nature (PPSN V)*, volume 1498 of *Lecture Notes in Computer Science*, pages 77–86. Springer, 1998.

[21] Jeanette Schmidt-Pruzan and Eli Shamir. Component structure in the evolution of random hypergraphs. *Combinatorica*, 5(1):81–94, 1985.

[22] Tobias Storch. Finding large cliques in sparse semi-random graphcs by simple randomized search heuristics. *Theoretical Computer Science*, 386:114–131, 2007.

[23] Andrew M. Sutton and Frank Neumann. Runtime analysis of evolutionary algorithms on randomly constructed high-density satisfiable 3-CNF formulas. In Thomas Bartz-Beielstein, Jürgen Branke, Bogdan Filipic, and Jim Smith, editors, *Proceedings of the Thirteenth International Conference on Parallel Problem Solving from Nature (PPSN XIII)*, volume 8672 of *Lecture Notes in Computer Science*, pages 942–951. Springer, 2014.

[24] Carsten Witt. Fitness levels with tail bounds for the analysis of randomized search heuristics. *Information Processing Letters*, 114(1–2):38 – 41, 2014.