Towards the Rigorous Analysis of Evolutionary Algorithms on Random $k$-Satisfiability Formulas

Andrew M. Sutton

THE UNIVERSITY of ADELAIDE
Worst-case complexity can be overly pessimistic.

Instead of measuring worst-case runtime, measure *expected* runtime over a probability distribution on an ensemble of instances (Levin 1986).

Random $k$-satisfiability: use some random process to generate a Boolean formula over $n$ variables.

What is the expected runtime $T(n)$ of an algorithm over the set of all such formulas?
This talk

SAT in the EC community
Many empirical results ... but lack of rigorous runtime analysis

So, many opportunities to make progress!

Talk outline

- Analysis of RLS on planted model
- Uniform model
  - ideas about how low-density formulas might be easy for simple EAs (w.h.p.)
Some definitions.

A Boolean variable: \( x \in \{\text{true}, \text{false}\} \)

A set of \( n \) Boolean variables: \( \{x, y, z, w\} \)

A set of literals: \( \{x, \bar{x}, y, \bar{y}, z, \bar{z}, w, \bar{w}\} \)

A clause: \( (x \lor y \lor \bar{z}) \)

A Boolean formula: \( (x \lor y \lor \bar{z}) \land (\bar{y} \lor \bar{w} \lor z) \land \ldots \)

A formula \( F \) is satisfiable iff there exists an assignment (a mapping \( A : \{x, y, z, w\} \rightarrow \{\text{true}, \text{false}\} \)) such that \( F \) evaluates to true.
Random 3-SAT

Uniform model: \( \Psi^U_{n,m} \)

Choose \( m \) length-3 clauses uniformly at random (without replacement) from the set of nontrivial clauses on \( n \) variables.

Planted model: \( \Psi^P_{n,p} \)

First, choose an assignment \( x^* \) to \( n \) variables uniformly at random. Then, every length-3 clause that is satisfied by \( x^* \) is included with probability \( p \).
Random 3-SAT

Uniform model: example

\[(x_1 \lor x_2 \lor \bar{x}_4) \land (\bar{x}_2 \lor \bar{x}_3 \lor x_4)\]

Planted model: example

\[x^* = (1, 0, 0, 1)\]

\[(x_1 \lor x_2 \lor \bar{x}_4) \land (x_2 \lor \bar{x}_3 \lor \bar{x}_1)\]
Random 3-SAT

Uniform model: example

\((x_1 \lor x_2 \lor \overline{x_4}) \land (\overline{x_2} \lor \overline{x_3} \lor x_4)\)

Planted model: example

\(x^* = (1, 0, 0, 1)\)

\((x_1 \lor x_2 \lor \overline{x_4}) \land (x_2 \lor \overline{x_3} \lor \overline{x_1})\)
Random 3-SAT

Uniform model: example

\[(x_1 \lor x_2 \lor \bar{x}_4) \land (\bar{x}_2 \lor \bar{x}_3 \lor x_4)\]

Planted model: example

\[x^* = (1, 0, 0, 1)\]
\[(x_1 \lor x_2 \lor \bar{x}_4) \land (x_2 \lor \bar{x}_3 \lor \bar{x}_1)\]
Random 3-SAT

**Uniform model: example**

\[(x_1 \lor x_2 \lor \bar{x}_4) \land (\bar{x}_2 \lor \bar{x}_3 \lor x_4)\]

**Planted model: example**

\[x^* = (1, 0, 0, 1)\]

\[(x_1 \lor x_2 \lor \bar{x}_4) \land (x_2 \lor \bar{x}_3 \lor \bar{x}_1)\]
Random 3-SAT

Uniform model: example

\[(x_1 \lor x_2 \lor \bar{x}_4) \land (\bar{x}_2 \lor \bar{x}_3 \lor x_4)\]

Planted model: example

\[x^* = (1, 0, 0, 1)\]

\[(x_1 \lor x_2 \lor \bar{x}_4) \land (x_2 \lor \bar{x}_3 \lor \bar{x}_1)\]
Random 3-SAT

Uniform model: example

$$(x_1 \lor x_2 \lor \bar{x}_4) \land (\bar{x}_2 \lor \bar{x}_3 \lor x_4)$$

Planted model: example

$$x^* = (1, 0, 0, 1)$$

$$(x_1 \lor x_2 \lor \bar{x}_4) \land (x_2 \lor \bar{x}_3 \lor \bar{x}_1)$$
Proposition (Chernoff bounds).

Let $X_1, X_2, \ldots, X_n$ be independent Poisson trials such that for $1 \leq i \leq n$, $\Pr(X_i = 1) = p_i$, where $0 < p_i < 1$. Let $X = \sum_{i=1}^{n} X_i$, $\mu = E(X) = \sum_{i=1}^{n} p_i$. Then the following inequalities hold for any $0 < \delta \leq 1$.

\[
\Pr(X \geq (1 + \delta)\mu) \leq e^{-\mu\delta^2/3}
\]
\[
\Pr(X \leq (1 - \delta)\mu) \leq e^{-\mu\delta^2/2}
\]
Randomized local search

Fitness function

Given a formula $F$ on $n$ variables and $m$ clauses,

$$f : \{0, 1\}^n \rightarrow \{0, 1, \ldots, m\}$$

counts the number of clauses satisfied by an assignment.

Algorithm 1: RLS

Choose $x \in \{0, 1\}^n$ uniformly at random;

while stopping criteria not met do
  $y \leftarrow x$;
  Choose $i \in \{1, \ldots, n\}$ uniformly at random;
  $y_i \leftarrow 1 - x_i$;
  if $f(y) > f(x)$ then $x \leftarrow y$

*Note the strict inequality in the selection step.*
Suppose $x^*$ is the planted solution to a formula $F \sim \Psi_{n,p}^P$.

We define a potential function $\varphi : \{0, 1\}^n \rightarrow \{0, 1, \ldots, n\}$:

$$\varphi(x) = d(x, x^*).$$

**Definition.**

We call an assignment $x$ bad if, for some $\epsilon > 0$,

$$\varphi(x) > \left(\frac{1}{2} + \epsilon\right)n.$$

An assignment which is not bad is called good.
Lemma 1

Suppose $x$ is chosen uniformly at random from $\{0, 1\}^n$. Then

$$\Pr(x \text{ is bad}) \leq e^{-\epsilon^2 n}.$$ 

Proof.

The probability that $x_i = x_i^*$ is exactly $1/2$ for all $i$. Thus let $X_i := [x_i = x_i^*]$ and the lemma follows from the Chernoff bound with $\mu = n/2$ and $\delta = 2\epsilon$. \qed
Consider an orientation of a hypercube graph $G = (V, E)$ where $V = \{0, 1\}^n$ and

$$(x, y) \in E \iff \phi(y) < \phi(x).$$
The underlying search space

**Definition.**
We label the directed edge \((x, y)\) in \(G\) **deceptive** if \(f(x) \geq f(y)\).

**Claim.**
1. Each planted solution induces an orientation of \(G\).
2. Each formula \(F\) induces a labeling of \(G\).

What is the fraction of edges that are labeled deceptive?
The underlying search space

\[ G = (V, E) \], with vertices labeled by fitness, deceptive edges in red:
Let $G' = (V', E')$ be the subgraph of $G$ induced by the set of all good nodes:

$$V' = \{x \in \{0, 1\}^n : x \text{ is good}\}$$

$$(x', y') \in E' \iff x', y' \in V' \text{ and } \varphi(y) < \varphi(x).$$

**Lemma.**

Let $(x, y) \in E'$. The probability that $(x, y)$ is labeled deceptive during the construction of $F$ is at most $2e^{-cpn^2}$ for a constant $c$. 
Most “good” edges are nondeceptive

Proof.
Let $S :=$ set of all clauses on $n$ variables that are:
- satisfied by $x^*$
- unsatisfied by $x$
- satisfied by $y$
Let $U :=$ set of all clauses on $n$ variables that are:
- satisfied by $x^*$
- satisfied by $x$
- unsatisfied by $y$

The edge $(x, y)$ is labeled deceptive iff $F$ has at least as many clauses in $U$ as in $S$. 
Most “good” edges are nondeceptive

Let $R_U (R_S)$ be the number of clauses from $U (S)$ in the formula. Then the probability that $(x, y)$ is labeled deceptive is $\Pr(R_U \geq R_S)$.

$R_U (R_S)$ is a binomial random variable with $|U| (|S|)$ trials and probability $p$.

W.L.O.G., suppose $x^* = (1, 1, \ldots, 1)$.

Since $d(x, y) = 1$, $x$ and $y$ differ in the (say) $i$-th bit.

Since $x$ is good, $\varphi(x) \leq (1/2 + \epsilon)n$.

Claim.

$$|S| = \binom{n - 1}{2}, \quad |U| = \binom{n - 1}{2} - \binom{n - \varphi(x)}{2}.$$ 

E.g., $x = 0001$, $y = 1001$, $\implies (x_1 \lor x_2 \lor \bar{x}_4) \in S$

E.g., $x = 0101$, $y = 1101$, $\implies (\bar{x}_1 \lor \bar{x}_2 \lor x_3) \in U$
Most “good” edges are nondeceptive

So the probability that \((x, y)\) is labeled deceptive is

\[
\Pr(R_U \geq R_S) \leq \Pr(R_U \geq a) + \Pr(R_S \leq a) \leq 2e^{-cpn^2}.
\]

The inequality comes from the Chernoff bound and \(c\) is a constant depending only on \(\epsilon\).

Lemma.
The fraction of deceptive edges in \(G'\) is at most \(ne^{-n(cp-\ln 2)}\).

Proof.
By the previous lemma, the fraction of deceptive edges in \(G'\) is at most \(|E'| \times 2e^{-cpn^2}\) and \(|E'| \leq |E| = n2^{n-1} \).
Claim.

Starting from a random initial solution \( x^{(0)} \), RLS gets to a local optimum \( \hat{x} \) in expected polynomial time.

We can think of RLS moving through \( G \). RLS moves along an edge \( (x, y) \) of \( G \) when it replaces the current solution \( x \) with a neighboring solution \( y \) of better fitness.
Theorem 1.
Probability RLS finds $x^*$ in expected polynomial time starting from a good initial solution $x^{(0)}$ is at least $1 - ne^{-n(cp-n-\ln 2)}$.

Proof.
If $x^{(0)} \in V'$ and RLS never moves along a deceptive edge, then the following are true.

1. RLS never leaves $G'$
2. $\hat{x} = x^*$

The probability that an arbitrary edge in $G'$ is deceptive is at most $ne^{-n(cp-n-\ln 2)}$. □

Theorem 2.
Suppose $p = \Omega(1/n)$. Then the probability that RLS succeeds on any random planted formula $F$ is $1 - o(1)$.

Proof.
$x^{(0)} \in V'$ w.h.p., and finds $x^*$ w.h.p. □
Now consider a formula $F \sim \Psi_{n,m}^U$.

**Definition.**

The **clause density** of $F$ is $m/n$.

Low-density formulas are *underconstrained*, high-density formulas are *overconstrained*.

**Satisfiability threshold conjecture**

For all $k \geq 3$ there exists a real number $r_c(k)$ such that

\[
\lim_{n \to \infty} \Pr\{F \text{ is satisfiable}\} = \begin{cases} 
1 & m/n < r_c(k); \\
0 & m/n > r_c(k). 
\end{cases}
\]
The uniform model

\[ \Pr \{ F \text{ is satisfiable} \} \]

\[ r_c \]

\[ \frac{m}{n} \]
The pure literal heuristic

**Definition.**

A literal $\ell$ is called *pure* in a set of clauses if its negation does not occur in that set.

**Example:**

$$(x_1 \lor \bar{x}_2 \lor \bar{x}_4) \land (\bar{x}_2 \lor x_4 \lor x_3)$$

Pure literals: $x_1$, $\bar{x}_2$, $x_3$.

**Algorithm 2:** The pure literal heuristic (PLH).

```
while $\mathcal{C}$ contains pure literals do
    Select a literal $\ell$ which is pure in $\mathcal{C}$;
    $\ell \leftarrow \text{true}$;
    $\mathcal{C} \leftarrow \mathcal{C} \setminus \{C \in \mathcal{C} : \ell \in C\}$;
```

The pure literal heuristic

**Theorem** (due to Broder et al., 1993).

Let $F \sim \Psi_{n,m}^U$ be a uniform random 3-SAT formula where $m/n < 1.63$. PLH succeeds on $F$ with probability $1 - o(1)$.
Ideas for the (1+1) EA

If PLH succeeds on a 3-SAT formula $F$, then PLH must succeed on every subset of clauses from $F$.

The (1+1) EA can simulate the first step of PLH since, as long as there are pure literals in $F$, a fitter solution can be obtained by setting them correctly.

Open question.

Can the (1+1) EA efficiently simulate PLH?