Probabilistic Analysis of the Unit Clause and Maximum Occurring Literal Selection Heuristics for the 3-Satisfiability Problem

by

Ming-Te Chao
Case Western Reserve University
Department of Computer Engineering and Science
Cleveland, OH 44106

and

John Franco
Indiana University
Department of Computer Science
Bloomington, IN 47405

TECHNICAL REPORT NO. 164

Probabilistic Analysis of the Unit Clause and Maximum Occurring Literal Selection Heuristics for the 3-Satisfiability Problem

by

Ming-Te Chao, Case Western Reserve University

and John Franco, Indiana University

December, 1984

This material is based on work supported by the U.S. Air Force under grant number AFOSR-84-0372.

ABSTRACT

An algorithm for the 3-Satisfiability problem is presented and a probabilistic analysis is performed. The analysis is based on an instance distribution which is parameterized to simulate a variety of sample characteristics. The algorithm assigns values to variables appearing in a given instance of 3-Satisfiability, one at a time, using the unit clause heuristic and a maximum occurring literal selection heuristic; at each step a variable is chosen randomly from a subset of variables which is usually large. The algorithm runs in polynomial time and it is shown that the algorithm finds a solution to a random instance of 3-Satisfiability with probability bounded from below by a constant greater than zero for a range of parameter values. The heuristics studied here can be used to select variables in a Backtrack algorithm for 3-Satisfiability. Experiments have shown that for about the same range of parameters as above the Backtrack algorithm using the heuristics finds a solution in polynomial average time.

1. Introduction

This paper is concerned with the probabilistic performance of two heuristics for the 3-Satisfiability problem (3-SAT). 3-SAT is the problem of determining whether all of a collection of 3-literal disjunctions (clauses) of Boolean variables are *true* for some truth assignment to the variables. This problem is NP-complete so there is no known polynomial time algorithm for solving it. 3-SAT is a special case of the Satisfiability problem (SAT) which is the problem of determining whether all of a collection of disjunctions of Boolean variables are *true* for some truth assignment to the variables.

The analysis is based on an equally likely instance distribution which has been used in other studies of algorithms for this problem. This model has two parameters: n, the number of disjunctions, and r, the number of variables from which disjunctions are composed. The model (which we refer to as $M(n,r,3)$) is described in greater detail in the next section. In [7] it was shown that, under $M(n,r,3)$, if $\lim_{n,r\to\infty}\frac{n}{r} > 5.2$ then random instances have no solution with probability approaching 1. In [2] it was reported that, according to experiments, random instances have no solution with probability approaching 1 if $\lim_{n,r\to\infty}\frac{n}{r} > 4$. In [1] it was shown that Backtracking solves 3-SAT in exponential average time for all limiting ratios of n to r which are constant. In [7] it was shown that a variant of the Davis-Putnam Procedure [3] which searches for all solutions to the given instance requires exponential time in probability under $M(n,r,3)$ for all limiting ratios of n to r which are constant. But, in [5] it was shown that the Pure-Literal heuristic can be used to solve random instances of 3-SAT in polynomial time with probability approaching 1 when $\lim_{n,r\to\infty}\frac{n}{r} < 1$. In this paper it is shown that the Unit-Clause heuristic and a maximum occurring literal selection heuristic can be used to solve random instances of 3-SAT in polynomial time with probability bounded from below by a constant when $\lim_{n,r\to\infty}\frac{n}{r} < 2.9$. A similar analysis shows that the Unit-Clause heuristic alone solves random instances in polynomial time with bounded probability when $\lim_{n,r\to\infty}\frac{n}{r} < 2.66$. These results are useful because they indicate the effectiveness of the two heuristics when used in a Backtrack algorithm for 3-SAT. Experiments suggest that Backtracking, using the two heuristics to determine which literal to consider at each step, will verify in polynomial average time that a solution exists for about the same range of limiting ratios of n to r.

There are a number of papers which investigate the probabilistic performance of SAT; these papers present results which are closely related to the results obtained for 3-SAT. These results are based on the constant-density model for SAT: construct each of n clauses independently by placing each of r variables independently in a clause with probability p and complementing those variables in each clause with probability 1/2. Average case results using the constant-density model or a variation are in [1], [8], [9], [10] and [11]. Probabilistic results using the constant-density model are in [6]. According to the results in [6], when the average number of literals in a clause is 3, random instances of SAT are nearly always proven to have no solutions in polynomial time.

2. 3-Satisfiability and The Probabilistic Model

The following terms are used to describe 3-SAT. Let $V = \{v_1, v_2 ... v_r\}$ be a set of r boolean variables. Associated with each variable v_i is a positive literal, denoted by v_i, and a negative literal, denoted by $\bar{v}_i$ and literal v_i has value *true* iff the variable v_i has value *true* and literal $\bar{v}_i$ has value *true* iff the variable v_i has value *false*. The literals v_i and $\bar{v}_i$ are said to be complementary. If l is a literal then $comp(l)$ is the literal which is complementary to l. A clause is a subset of the set of all literals associated with the variables of V such that no two literals in the subset are complementary. A truth assignment to V is an assignment of truth values to every variable in V. A clause c is satisfied by truth assignment t if at least one literal in c has value *true* under t. Let $A_i(V)$ denote the set of i-literal clauses that can be composed of literals associated with the variables of V. An instance I of 3-SAT is a collection of clauses chosen from $A_3(V)$ and the problem is to find a truth assignment to V which satisfies all clauses in I, if one exists, and to verify that no such truth assignment exists otherwise. A truth assignment which satisfies all clauses in I is said to be a solution to I.

The probabilistic model used for analysis is presented by describing the method used to construct random instances. A random instance of 3-SAT contains n clauses chosen uniformly, independently and with replacement from $A_3(V)$. The distribution associated with this construction is referred to as $M(n, r, 3)$.

3. The Algorithm SC_1

The algorithm we consider, called SC_1, takes as input a collection of clauses I and outputs "a solution exists" or "cannot determine whether a solution exists". SC_1 contains a single loop. At each iteration ofn the loop a literal is chosen and some clauses and literals are removed from I. Let $C_i^{I,\sigma}(j)$, for all $1 \leq i \leq 3$, denote the collection of clauses in I containing exactly i literals at the end of the j^{th} iteration where σ denotes the sequence of chosen literals. We shorten $C_i^{I,\sigma}(j)$ to $C_i(j)$. Then $C_i(0) = \phi$ for all $1 \leq i \leq 2$ and $|C_3(0)| = n$. If the $j+1^{st}$ chosen literal is l then the lines

Remove from I all clauses containing l
Remove from I all occurrences of $comp(l)$

Have the following effect

$$\begin{aligned}\forall\, 1 \leq i \leq 2\ C_i(j+1) = \{c : c \in C_i(j) \textit{ and } l \notin C_i(j) \textit{ and } comp(l) \notin C_i(j) \\ \textit{or } c \cup \{comp(l)\} \in C_{i+1}(j)\}\end{aligned}$$
$$C_3(j+1) = \{c : c \in C_3(j) \textit{ and } l \notin C_3(j) \textit{ and } comp(l) \notin C_3(j)\}.$$

In what follows $card(v, C_3(j))$ is the number of clauses in $C_3(j)$ which contain the literal v and $card(\bar{v}, C_3(j))$ is the number of clauses in $C_3(j)$ which contain the literal $\bar{v}$. Clauses in $C_1(j)$ are said to be unit clauses. Finally, $var(l)$ is the variable associated with literal l.

```
$SC_1(I)$:
        $j \leftarrow 0$
        Repeat
            If $|C_1(j)| = 0$ Then Begin
                Choose $v$ randomly from $V$
                $V \leftarrow V - \{v\}$
                If $card(\bar{v}, C_3(j)) > card(v, C_3(j))$ Then $l \leftarrow \bar{v}$ Else $l \leftarrow v$
                End
            Else Begin
                Choose $l$ randomly from $C_1(j)$
                $V \leftarrow V - \{var(l)\}$
                End
            Remove from $I$ all clauses containing $l$
            Remove from $I$ all occurrences of $comp(l)$
            $j \leftarrow j + 1$
        Until $I$ is empty or there exist two complementary unit clauses in $I$
        If $I$ is empty Then Output("a solution exists")
            Else Output("cannot determine whether a solution exists")
```

SC_1 runs in less than $O(r^2n)$ time since I must be empty after r iterations of the loop and the remove and *card* operations need look at no more than $r * n$ literals. An instance I of SAT has a solution if SC_1 run on I outputs "a solution exists": one solution to I may be found by assigning the value *true* to the variables whose positive literals were chosen and the value *false* to all other variables.

4. Analysis of SC_1

In this section it is shown that if instances are generated according to $M(n, r, 3)$ and $\lim_{n,r\to\infty} \frac{n}{r} < 2.9$ then for some $\epsilon > 0$, the probability that SC_1 outputs "a solution exists" is greater than ϵ.

The following theorem will be used to show how the collections of clauses in $C_i(j)$ are distributed.

Theorem 1:

Let V_{r-j} be the subset of variables associated with unchosen literals after j literals have been chosen. Suppose for all $1 \le i \le 3$ the clauses in $C_i(j)$ are independent and are equally likely to be any clause in $A_i(V_{r-j})$. Then for all $1 \le i \le 3$ the clauses in $C_i(j+1)$ are independent and equally likely to be any clause in $A_i(V_{r-j-1})$.

Proof:

Either the variable v is chosen randomly from V if $|C_1(j)| \neq 0$ or it is chosen randomly from $C_1(j)$. Consider the first case. Let c_1 and c_2 be two clauses in $C_i(j+1)$ and let $\hat{c}_1$ and $\hat{c}_2$ be the two clauses in $C_i(j)$ from which c_1 and c_2 were derived after the $j+1^{st}$ literal was chosen. Then

$$pr(c_1 = x) =$$

$$pr(\hat{c}_1 = x \textit{ or } \hat{c}_1 = x \cup \{v\} \textit{ and } \bar{v} \text{ was chosen } \textit{or } \hat{c}_1 = x \cup \{\bar{v}\} \text{ and } v \text{ was chosen})$$

$$= pr(\hat{c}_1 = y \textit{ or } \hat{c}_1 = y \cup \{v\} \textit{ and } \bar{v} \text{ was chosen } \textit{or } \hat{c}_1 = y \cup \{\bar{v}\} \text{ and } v \text{ was chosen})$$

$$= pr(c_1 = y).$$

Also, $pr(c_1 = x_1 \text{ and } c_2 = x_2) =$

$$\begin{aligned}
pr(&\hat{c}_1 = x_1 \text{ and } \hat{c}_2 = x_2 \text{ or } \hat{c}_1 = x_1 \text{ and } \hat{c}_2 = x_2 \cup \{v\} \text{ and } \bar{v} \text{ was chosen or}\\
&\hat{c}_1 = x_1 \text{ and } \hat{c}_2 = x_2 \cup \{\bar{v}\} \text{ and } v \text{ was chosen or}\\
&\hat{c}_1 = x_1 \cup \{v\} \text{ and } \hat{c}_2 = x_2 \text{ and } \bar{v} \text{ was chosen or}\\
&\hat{c}_1 = x_1 \cup \{\bar{v}\} \text{ and } \hat{c}_2 = x_2 \text{ and } v \text{ was chosen or}\\
&\hat{c}_1 = x_1 \cup \{v\} \text{ and } \hat{c}_2 = x_2 \cup \{v\} \text{ and } \bar{v} \text{ was chosen or}\\
&\hat{c}_1 = x_1 \cup \{\bar{v}\} \text{ and } \hat{c}_2 = x_2 \cup \{\bar{v}\} \text{ and } v \text{ was chosen})
\end{aligned}$$

$= pr(\hat{c}_1 = y_1$ and $\hat{c}_2 = y_2$ or $\hat{c}_1 = y_1$ and $\hat{c}_2 = y_2 \cup \{v\}$ and $\bar{v}$ was chosen or
$\hat{c}_1 = y_1$ and $\hat{c}_2 = y_2 \cup \{\bar{v}\}$ and v was chosen or
$\hat{c}_1 = y_1 \cup \{v\}$ and $\hat{c}_2 = y_2$ and $\bar{v}$ was chosen or
$\hat{c}_1 = y_1 \cup \{\bar{v}\}$ and $\hat{c}_2 = y_2$ and v was chosen or
$\hat{c}_1 = y_1 \cup \{v\}$ and $\hat{c}_2 = y_2 \cup \{v\}$ and $\bar{v}$ was chosen or
$\hat{c}_1 = x_1 \cup \{\bar{v}\}$ and $\hat{c}_2 = y_2 \cup \{\bar{v}\}$ and v was chosen)

$= pr(c_1 = y_1$ and $c_2 = y_2)$.

Consider the second case. The $j + 1^{st}$ chosen variable is equally likely to be any of $r - j$ variables and is selected independently of clauses in $C_i(j)$ for all $2 \le i \le 3$. Hence for all $2 \le i \le 3$ we may use the proof of the first case. For $i = 1$ the result follows from the independence and equal liklihood of the unit clauses.

Corollary 1:

For all $0 \le j \le r$ and $1 \le i \le 3$ all clauses in $C_i(j)$ are independent and equally likely to be any clause in $A_i(V_{r-j})$.

Proof:

By induction on j. The basis step holds because of the assumed distribution on instances given to SC_1. The induction step holds because of theorem 1.

Because of corollary 1 a system of differential equations for finding the expected number of clauses in $C_i(j)$ for all $2 \le i \le 3$ may be obtained. Let $n_i(j)$ denote the number of clauses in $C_i(j)$, let $w_i(j)$ denote the number of $i-$literal clauses added to $C_i(j)$ as a result of choosing the j^{th} variable and let $z_i(j)$ denote the number of clauses eliminated from $C_i(j)$ as a result of choosing the j^{th} variable. These three terms depend on I and σ but this dependence is omitted from the terms for the sake of simplicity. The $w_i(j)$ term may be thought of as representing the "rate of flow" of clauses into $C_i(j)$ when the j^{th} variable is chosen and the $z_i(j)$ term may be thought of as representing the "rate of flow" of clauses out of $C_i(j)$ when the j^{th} variable is chosen. If the average rate of flow into $C_1(j)$ is always less than 1 the number of clauses in $C_1(j)$ will not, in probability, grow very large since at least one clause is removed from $C_1(j)$ whenever $C_1(j) \neq \phi$. In this case the probability that a complementary pair of clauses exists in $C_1(j)$ for some j is small. However, if the average rate of flow into $C_1(j)$ rises above 1 for a constant fraction of the values of $\frac{j}{r}$ then the number of clauses in $C_1(j)$ gets large for a fraction of the values of $\frac{j}{r}$ since the flow out of $C_1(j)$ [illegible] nptotically no more than one unless $|C_1(j)|$ is large. In this case the probability that there is a complementary pair of clauses in $C_1(j)$ for some j is near 1. Since, as will be seen from the analysis below, if the expected flow into $C_1(j)$ goes above $1+\epsilon$ for any $\epsilon > 0$ then it stays above 1 for a constant fraction of values of $\frac{j}{r}$, the point at which $E\{w_1(j)\}$ is around 1 is a critical one regarding the probabilistic performance of SC_1.

We now develop the differential equations for finding $E\{w_1(j)\}$, solve them and find the condition on $\frac{n}{r}$ which causes $E\{w_1(j)\} < 1$. Later, it will be shown that this implies SC_1 finds a satisfying truth assignment when one exists with probability greater than some positive constant.

Clearly, for $1 \leq i \leq 3$

$$n_i(j+1) = n_i(j) + w_i(j+1) - z_i(j+1).$$

Taking expectations gives

$$E\{n_i(j+1)\} = E\{n_i(j)\} + E\{w_i(j+1)\} - E\{z_i(j+1)\}$$

which can be written

$$E\{n_i(j+1)\} - E\{n_i(j)\} = E\{w_i(j+1)\} - E\{z_i(j+1)\}. \tag{1}$$

For large r we can approximate (1) by

$$\frac{dE\{n_i(j)\}}{dj} = E\{w_i(j+1)\} - E\{z_i(j+1)\}. \tag{2}$$

But, for all $1 \leq i \leq 3$

$$\begin{aligned} E\{z_i(j+1)\} &= E\{E\{z_i(j+1)/n_i(j)\}\} \\ &= E\{\frac{i * n_i(j)}{r-j}\} = \frac{i * E\{n_i(j)\}}{r-j} \end{aligned} \tag{3a}$$

because of corollary 1. Also,

$$\begin{aligned} E\{w_1(j+1)\} &= E\{E\{w_1(j+1)/n_2(j)\}\} \\ &= E\{\frac{2 * n_2(j)}{2(r-j)}\} = \frac{E\{n_2(j)\}}{r-j} \end{aligned} \tag{3b}$$

and

$$E\{w_3(j+1)\} = 0.$$

Finally, $E\{w_2(j+1)\} =$

$$3\frac{E\{n_3(j)\}}{2(r-j)} - H_2(j+1) * pr(j+1^{st} \text{ chosen literal does not come from } C_1(j)) \quad (4)$$

where $H_2(j+1)$ is the average number of extra clauses removed from $C_3(j)$ given the $j+1^{st}$ chosen literal does not come from $C_1(j)$. Therefore (2), for $i = 3$ can be written

$$\frac{dE\{n_3(j)\}}{dj} = -\frac{3 * E\{n_3(j)\}}{r-j}. \quad (5)$$

The solution to this differential equation under the assumption that $E\{n_3(0)\} = n$ is

Theorem 2:

$$E\{n_3(j)\} = (1 - \frac{j}{r})^3 n.$$

Proof:

Straightforward solution to (5).

In order to solve (2) for $i = 2$ we must first find $H_2(j+1)$ and the probability that the $j+1^{st}$ chosen literal does not come from $C_1(j)$. It suffices to find a lower bound for $H_2(j+1)$ and the probability mentioned since we require only an upper bound on $E\{w_1(j)\}$.

Theorem 3:

$$H_2(j+1) \geq \frac{8}{9\sqrt{2\pi}} E\{\sum_{y=1}^{n_3(j)} \sqrt{y} \binom{n_3(j)}{y} (\frac{3}{r-j})^y (1 - \frac{3}{r-j})^{n_3(j)-y}\}$$

Proof:

The probability that a particular literal appears in x clauses given the variable associated with that literal appears in y clauses is $\binom{y}{x}(\frac{1}{2})^x(\frac{1}{2})^{y-x}$. Hence the expected number of clauses containing the least frequently occurring literal associated with the chosen variable given y is

$$2\sum_{x=0}^{\lfloor \frac{y}{2} \rfloor} x \binom{y}{x} (\frac{1}{2})^y = \frac{y}{2} - \frac{\frac{y+1}{2}\binom{y}{\lfloor \frac{y}{2} \rfloor}}{2^y} \qquad \text{if } y \text{ is odd}$$

and

$$2\sum_{x=0}^{\frac{y}{2}-1} x\binom{y}{x}(\frac{1}{2})^y + \frac{y}{2}\binom{y}{\frac{y}{2}}(\frac{1}{2})^y = \frac{y}{2} - \frac{\frac{y}{2}\binom{y}{\frac{y}{2}}}{2^y} \qquad \text{if } y \text{ is even}$$

Let

$$G(y) = \begin{cases} \frac{\binom{y}{\frac{y}{2}}}{2^y} & y \text{ even} \\ \frac{(1+\frac{1}{y})\binom{y}{\lfloor\frac{y}{2}\rfloor}}{2^y} & y \text{ odd} \end{cases}$$

Then

$$E\{H_2(j+1)/n_3(j)\} = \frac{1}{2}\sum_{y=0}^{n_3(j)} yG(y)\binom{n_3(j)}{y}(\frac{3}{r-j})^y(1-\frac{3}{r-j})^{n_3(j)-y}$$

$$> \frac{8}{9\sqrt{2\pi}}\sum_{y=0}^{n_3(j)} \sqrt{y}\binom{n_3(j)}{y}(\frac{3}{r-j})^y(1-\frac{3}{r-j})^{n_3(j)-y}$$

since, by stirling's formula, $G(y) > \frac{16}{9\sqrt{2\pi y}}$. Taking the expectation gives the desired result.

If $E\{n_3(j)\}$ and $r-j$ are large and $\frac{n}{r}$ is a constant, since $\lim_{n,r\to\infty}\frac{E\{n_3(j)\}}{r-j}$ is bounded by a constant and since $n_3(j)$ is binomially distributed then the lower bound for $H_2(j+1)$ may be approximated by the expression

$$\frac{8}{9\sqrt{2\pi}}\beta\sqrt{\frac{3*E\{n_3(j)\}}{r-j}}$$

where β depends on the value of the expression under the large square root sign. A few values of β are as follows:

$\frac{3*E\{n_3(j)\}}{r-j}$	β
1	.7731
2	.891
4	.96
8	.983
16	.992

But, $E\{n_3(j)\} = (1-\frac{j}{r})^3 n$ so, for $1 \leq j \leq \delta r$ where δ is any constant between zero and one, the lower bound for $H_2(j+1)$ is approximately

$$\frac{8\sqrt{3}}{9\sqrt{2\pi}}\beta(1-\frac{j}{r})\sqrt{\frac{n}{r}}. \tag{6}$$

Only a lower bound for the probability that the $j+1^{st}$ chosen literal does not come from $C_1(j)$ still needs to be found.

Theorem 4:

$$\Pr(j+1^{st} \text{ chosen literal does not come from } C_1(j)) > 1 - E\{w_1(j)\} \tag{7}$$

for all j from 1 to the point at which $E\{w_1(j)\}$ is maximum.

Proof:

The probability that the $j+1^{st}$ chosen literal comes from $C_1(j)$ is the probability that $C_1(j) \neq \phi$. The probability that $C_1(j) \neq \phi$ is less than the rate at which clauses enter $C_1(j)$ which is $E\{w_1(j)\}$. This is because at least one clause is removed from $C_1(j)$ if $C_1(j) \neq \phi$ and $E\{w_1(j)\}$ is rising with j. The probability required is therefore greater than $1 - E\{w_1(j)\}$.

Substituting (6),(7), $(1-\frac{j}{r})^3 n$ for $E\{n_3(j)\}$ and $\frac{E\{n_2(j)\}}{r-j}$ for $E\{w_1(j)\}$ (from (3b)) into (4) and substituting the result and (3a) into (2) with i set to 2 gives

$$\frac{dE\{n_2(j)\}}{dj} = \frac{3}{2}(1-\frac{j}{r})^2\frac{n}{r} - \frac{2 * E\{n_2(j)\}}{r(1-\frac{j}{r})} + \frac{8\sqrt{3}}{9\sqrt{2\pi}}\beta\sqrt{\frac{n}{r}}\,\frac{E\{n_2(j)\}}{r}$$

$$-\frac{8\sqrt{3}}{9\sqrt{2\pi}}\beta\sqrt{\frac{n}{r}}\,(1-\frac{j}{r}). \tag{8}$$

For the moment suppose β is constant. Then the solution to (8) with boundary condition $E\{n_2(0)\} = 0$ and with α substituted for $\frac{8\sqrt{3}}{9\sqrt{2\pi}}$ is

$$E\{n_2(j)\} = (1-\frac{j}{r})^2 e^{\frac{j}{r}\alpha\beta\sqrt{\frac{n}{r}}}\left[\frac{3}{2}\frac{n}{\alpha\beta\sqrt{\frac{n}{r}}}(1-e^{-\frac{j}{r}\alpha\beta\sqrt{\frac{n}{r}}})\right.$$

$$+r\alpha\beta\sqrt{\frac{n}{r}}\ln(1-\frac{j}{r}) - r(\alpha\beta\sqrt{\frac{n}{r}})^2(\ln(1-\frac{j}{r})+\frac{j}{r})$$

$$+r\frac{(\alpha\beta\sqrt{\frac{n}{r}})^3}{2}(\ln(1-\frac{j}{r})+\frac{j}{r}+\frac{j^2}{2r^2})-\ldots\Bigg].$$

Thus, from (3b)

$$E\{w_1(j)\}=(1-\frac{j}{r})e^{\frac{j}{r}\alpha\beta\sqrt{\frac{n}{r}}}\left[\frac{3}{2}\frac{1}{\alpha\beta\sqrt{\frac{n}{r}}}\frac{n}{r}(1-e^{-\frac{j}{r}\alpha\beta\sqrt{\frac{n}{r}}})\right.$$

$$+\alpha\beta\sqrt{\frac{n}{r}}\ln(1-\frac{j}{r})-(\alpha\beta\sqrt{\frac{n}{r}})^2(\ln(1-\frac{j}{r})+\frac{j}{r})$$

$$\left.+\frac{1}{2}(\alpha\beta\sqrt{\frac{n}{r}})^3(\ln(1-\frac{j}{r})+\frac{j}{r}+\frac{j^2}{2r^2})-\ldots\right]. \qquad (9)$$

The expression on the right in (9) has a maximum in the vicinity of and greater than $j=\frac{r}{2}$. We call the point at which the maximum occurs j_o. If $\frac{n}{r}=2.9$ then $3*E\{n_3(j_o)\}/(r-j_o)\approx 1.9$ so $\beta\approx .89$ at $j=j_o$. Since $3*E\{n_3(j)\}/(r-j)<\frac{3n}{r}<9$, $\beta>.89$ for all $0\leq j<j_o$ so (8) with β set to .89 gives an upper bound on $E\{n_2(j)\}$ and therefore $E\{w_1(j)\}$ up to j_o. It can be seen from (9) that $E\{w_1(j)\}$ for $1\leq j\leq j_o$ is less than 1 when $\beta=.89$ and $\frac{n}{r}=2.9$

The solution to (8) with $\beta=0$ is an upper bound on $E\{n_2(j)\}$ in the range $j_o\leq j<r$. When divided by $(r-j)$ and an appropriate boundary condition is added this solution is an upper bound for $E\{w_1(j)\}$ in the range $j_o\leq j<r$ and has value equal to the value of $E\{w_1(j_o)\}$ at $j=j_o$. Since this bound is maximal at $j=\frac{r}{2}$ the maximum value of this bound in the range $j_o\leq j<r$ is equal to the maximum value of the first bound in the range $0\leq j\leq j_o$. Hence

Theorem 5:

Given that inputs to SC_1 are distributed according to $M(n,r,3)$,

$$E\{w_1(j)\}<1 \text{ for all } 0\leq j<r \text{ when } \lim_{n,r\rightarrow\infty}\frac{n}{r}<2.9$$

We now prove the main result

Theorem 6:

SC_1 verifies that a solution exists for satisfiable instances generated according to $M(n,r,3)$ with probability greater than ϵ for some $\epsilon > 0$ when $\lim_{n,r\to\infty} \frac{n}{r} < 2.9$.

Proof:

From theorem 5 $E\{w_1(j)\} < 1$ for all $0 \le j < r$ when $\lim_{n,r\to\infty} \frac{n}{r} < 2.9$. From corollary 1 the clauses entering $C_1(j+1)$ from $C_2(j)$ are statistically independent. Suppose all clauses entering $C_1(j+1)$ are regarded as entering $C_1(j+1)$ in some order which is decided arbitrarily. Then the probability that the q^{th} clause entering $C_1(j+1)$ is complementary to no clause in $C_1(j+1)$ is

$$\left(1 - \frac{1}{2(r-j)}\right)^{n_1(j)+q-1}.$$

Therefore, the probability that none of the clauses entering $C_1(j+1)$ is complementary to any clause in $C_1(j+1)$ is

$$\left(1 - \frac{1}{2(r-j)}\right)^{n_1(j)*w_1(j)+w_1(j)*(w_1(j)-1)/2}$$

so the probability that no complementary pair is encountered during a run of SC_1 is

$$\sum\prod_{j=0}^{r-1}\left(1 - \frac{1}{2(r-j)}\right)^{n_1(j)*w_1(j)+w_1(j)*(w_1(j)-1)/2} pr(\ldots n_1(j), w_1(j)\ldots)$$

$$> \sum\prod_{j=0}^{r-1}\left(1 - \frac{1}{2r}\right)^{\frac{2r}{r-j}(n_1(j)*w_1(j)+w_1(j)*(w_1(j)-1)/2)} pr(\ldots n_1(j), w_1(j)\ldots)$$

$$= \sum\left(1 - \frac{1}{2r}\right)^{2r\sum_{j=0}^{r-1}\frac{2n_1(j)*w_1(j)+w_1(j)*(w_1(j)-1)}{2(r-j)}} pr(\ldots n_1(j), w_1(j)\ldots). \quad (10)$$

If the sum in the exponent of (10) is less than $\frac{\kappa n}{r}$ (where κ is a constant) with probability bounded from below by 2/3 then (10) is bounded from below by $\frac{2}{3}(1-\frac{1}{2r})^{2n\kappa}$ which approaches a constant as r approaches infinity if the limiting ratio of n to r is constant. To show that the sum in the exponent of (10) is less than $\frac{\kappa n}{r}$ with probability greater than 2/3 we show that the expectation of the sum is bounded from above by $\frac{\kappa n}{3r}$ and apply markov's inequality.

To show that the expectation of the sum in the exponent of (10) is less than $\frac{\kappa n}{3r}$ we need only show that the expectation of each term in the sum is less than $\frac{\kappa n}{3r^2}$. Denote by $p_1(j)$ the j^{th} term in the sum. Then

$$E\{p_1(j)\}$$

$$\leq \frac{1}{2(r-j)}\left(E\{w_1^2(j)\} + \sum_{s=0}^{n}\sum_{t=0}^{n} 2 * s * t * pr(n_1(j) = t, w_1(j) = s)\right). \qquad (11)$$

The second term within parentheses is bounded by $\gamma_1(1-\frac{j}{r})\frac{n}{r}$ for $j < r - r^{8/9}$ and by $\gamma_2(1-\frac{j}{r})\frac{n}{r}$ for $j \geq r - r^{8/9}$ where γ_1 and γ_2 are constants greater than zero. Consider the first case, $1 \leq j < r - r^{8/9}$. Suppose SC_1 is modified so that all literals not chosen from $C_1(j)$ are chosen randomly from the set of all unchosen literals and suppose that $\hat{n}_2(j)$ and $\hat{w}_1(j)$ have the same meaning as $n_2(j)$ and $w_1(j)$ except applied to the modified SC_1. Define $n_l = E\{\hat{n}_2(j)\} - n^{3/4}$ and $n_u = E\{\hat{n}_2(j)\} + n^{3/4}$. It is easy to see that $\hat{n}_2(j)$ is binomially distributed with mean $E\{\hat{n}_2(j)\}$ proportional to $\frac{j}{r}(1-\frac{j}{r})^2 n$ so the probability that $n_l < \hat{n}_2(j) < n_u$ is greater than $1 - 2e^{-n^{3/2}/E\{\hat{n}_2(j)\}}$ from [4] and this is greater than $1 - e^{-\sqrt{n}}$ since $E\{\hat{n}_2(j)\} < n$. The double sum of (11) can be bounded from above by using $\hat{w}_1(j)$ for $w_1(j)$. We do so and split the result into three parts:

$$\sum_{s=0}^{n}\sum_{t=0}^{n}\sum_{u=0}^{n_l} 2 * s * t * pr(n_1(j) = t, \hat{n}_2(j) = u, \hat{w}_1(j) = s)$$

$$+\sum_{s=0}^{n}\sum_{t=0}^{n}\sum_{u=n_l}^{n_u} 2 * s * t * pr(n_1(j) = t, \hat{n}_2(j) = u, \hat{w}_1(j) = s)$$

$$+\sum_{s=0}^{n}\sum_{t=0}^{n}\sum_{u=n_u}^{n} 2 * s * t * pr(n_1(j) = t, \hat{n}_2(j) = u, \hat{w}_1(j) = s)$$

$$< \frac{4n^2}{e\sqrt{n}} + 2 * E\{\hat{w}_1(j)\} * E\{n_1(j)\} \qquad (12)$$

in the limit since $\frac{n}{r} < 2.9$ and $|n_u - n_l| \to 0$. But $E\{\hat{w}_1(j)\}$ may be shown to be proportional to $(1 - \frac{j}{r})\frac{n}{r}$ by solving (8) with $\beta = 0$ and dividing by $r - j$. Also, $E\{n_1(j)\}$ is bounded by a constant for all $1 \leq j \leq r$ since $E\{w_1(j)\} < 1$ and at least one clause is removed from $C_1(j)$ if $C_1(j) \neq \phi$. So (12) is less than $\gamma_1(1 - \frac{j}{r})\frac{n}{r}$ where γ_1 is a constant greater than zero. Now consider the case $r - r^{8/9} \leq j < r$. In this range $E\{\hat{w}_1(j)\}$ is proportional to $(1 - \frac{j}{r})\frac{n}{r}$ and is decreasing with increasing j. Clearly, in this range

$$\sum_{s=0}^{n}\sum_{t=0}^{n} 2 * s * t * pr(n_1(j) = t, w_1(j) = s)$$

$$< 2 * E\{\hat{w}_1(j)\} * E\{n_1(r - r^{8/9})\} < \gamma_2(1 - \frac{j}{r})\frac{n}{r}$$

as $r \to \infty$.

We now need to find a bound on $E\{w_1^2(j)\}$. Let $\hat{w}_1(j)$ be as before. Clearly, $E\{w_1^2(j)\} \leq E\{\hat{w}_1^2(j)\}$. But $\hat{w}_1(j)$ is distributed binomially hence $E\{\hat{w}_1^2(j)\} = \sigma^2(\hat{w}_1(j)) + (E\{\hat{w}_1(j)\})^2 < E\{\hat{w}_1(j)\} + (E\{\hat{w}_1(j)\})^2$ and

$$E\{w_1^2(j)\} < \gamma_3 * (1 - \frac{j}{r})\frac{n}{r}.$$

Let $\gamma = \max\{\gamma_1, \gamma_2\}$. Substituting $\gamma(1 - \frac{j}{r})\frac{n}{r}$ for the double sum in (11) and then $\gamma_3 * (1 - \frac{j}{r})\frac{n}{r}$ for $E\{w_1^2(j)\}$ in the resulting inequality gives

$$E\{p_1(j)\} \leq \left(\frac{\gamma_3 + \gamma}{2}\right)\frac{n}{r^2} = \frac{\kappa}{3} * \frac{n}{r^2}.$$

From this the expectation of the sum in the exponent of (9) is less than $\frac{\kappa n}{3r}$. By markov's inequality the probability that the sum is greater than $\frac{\kappa n}{r}$ is less than 1/3. Therefore, the probability that the sum is less than $\frac{\kappa n}{r}$ is greater than 2/3. Thus (10) is greater than $\frac{2}{3}(1 - \frac{1}{2r})^{2n\kappa}$ which approaches $\frac{2}{3}e^{-\frac{n\kappa}{r}}$ as r approaches infinity. Let $\epsilon = \frac{2}{3}(1 - \frac{1}{2r})^{2n\kappa}$.

The Unit-Clause heuristic and the maximum occurring literal heuristic have been incorporated into a Backtrack algorithm for 3-SAT and experiments run. The algorithm is

```
$BA(I)$ :
    If there exist two complementary unit clauses in $I$ Then return $UNSAT$
    Else If $I = \phi$ Then return $SAT$
    Else If there is a unit clause in $I$ Then Begin
        While there is a unit clause $\{l\}$ in $I$ Do Begin
            Remove from $I$ all clauses containing $l$
            Remove from $I$ all occurrences of $comp(l)$
            End
        Return $BA(I)$
        End
    Else Begin
        Choose a variable $v$ which is present in $I$
        If $card(\bar{v}, C_3) > card(v, C_3)$ Then $l \leftarrow \bar{v}$ Else $l \leftarrow v$
        $I_1 = \{c : c \in I$ and $l, comp(l) \notin c$ or $c \cup \{l\} \in I\}$
        $I_2 = \{c : c \in I$ and $l, comp(l) \notin c$ or $c \cup \{comp(l)\} \in I$
        If $BA(I_1) = SAT$ Then return $SAT$
        Else If $BA(I_2) = SAT$ Then return $SAT$ Else return $UNSAT$
```

Algorithm BA was run on random instances of 3-SAT generated according to $M(n, r, 3)$ with $\frac{n}{r}$ set to 2.4, 2.6, 2.8, 3.0, 3.2, 3.4 and 3.6 for r ranging from 10 to 200 in steps of 10. At each data point the average number of calls to BA per instance was computed for 100 instances. The results are presented in figure 1. Note that for $\frac{n}{r} \leq 2.6$ the performance curves are practically straight lines, for $\frac{n}{r} = 2.8$ there are occasional peaks and for $\frac{n}{r} \geq 3.0$ the performance curves rise dramatically. Upon looking at the performance of individual instances for the case $\frac{n}{r} = 2.8$ it was noted that the peaks were due to a few runs that required many calls to BA.

5. A Modification to SC_1

In this section we discuss why, in SC_1, if $C_1(j) = \phi$ then the $j+1^{st}$ chosen literal is chosen only on the number of occurrences of that literal and its complement in $C_3(j)$ and not in $C_2(j)$. Suppose that the $j+1^{st}$ literal is chosen on the number of times it occurs in $C_3(j)$ and $C_2(j)$ if $C_1(j) = \phi$. Assume the most optimistic case: the literal appears in more clauses of both $C_3(j)$ and $C_2(j)$ than its complement. Let $E\{w_1^*(j)\}$ denote the new average "flow" of clauses into $C_1(j)$. Then

$$E\{w_1^*(j)\} = E\{w_1(j)\} - H_1(j)(1 - E\{w_1^*(j)\})$$

where $H_1(j)$ is the extra number of clauses removed from the "flow" into $C_1(j)$ when the chosen literal is not a unit clause and $1 - E\{w_1^*(j)\}$ is the probability (to within $O(\frac{1}{r})$) that the chosen literal is not a unit clause. So

$$E\{w_1^*(j)\} = \frac{E\{w_1(j)\} - H_1(j)}{1 - H_1(j)}.$$

Thus $E\{w_1^*(j)\} < 1$ is equivalent to $E\{w_1(j)\} < 1$ and no benefit is gained by considering the number of occurrences of the chosen literal in $C_2(j)$.

6. Conclusions

We have presented an algorithm for 3-SAT based on the Unit-Clause and maximum occurring literal heuristics and have shown that this algorithm finds a solution to a random instance of 3-SAT in polynomial time with probability bounded from below by a constant under $M(n, r, 3)$ when $\lim_{n,r\to\infty} \frac{n}{r} < 2.9$. Experiments indicate that a Backtrack algorithm containing these two heuristics performs extremely well probabilistically over the same range of values of the limiting ratio of $\frac{n}{r}$. The method used to get these results has the advantages of providing intuition and being general enough to be used on other algorithms for 3-SAT and other NP-complete problems. The method can be used to show that the Unit-Clause heuristic alone finds a solution to a random instance of 3-SAT in polynomial time with probability bounded from below by a constant under $M(n, r, 3)$ when $\lim_{n,r\to\infty} \frac{n}{r} < 2.66$: the analysis is the same as presented here except that $\beta = 0$.

References

[1] Brown, C.A. and Purdom, P.W., "An average time analysis of backtracking," *SIAM J. Comput.* **10** (1981), pp 583-593.

[2] Chao, M.T., "Probabilistic analysis and performance measurement of algorithms for the Satisfiability problem," Ph.D. Dissertation, Case Western Reserve University (1984).

[3] Davis, M. and Putnam, H., "A computing procedure for quantification theory," *J.ACM* **7** (1960), pp 201-215.

[4] Erdos, P. and Spencer, J., *Probabilistic Methods in Combinatorics*, Academic Press, 1974.

[5] Franco, J., "Probabilistic analysis of the pure literal heuristic for the satisfiability problem," *Annals of Operations Research* **1** (1984), pp 273-289.

[6] Franco, J., "Sensitivity of probabilistic results on algorithms for NP-complete problems to input distributions," Case Western Reserve University (1983).

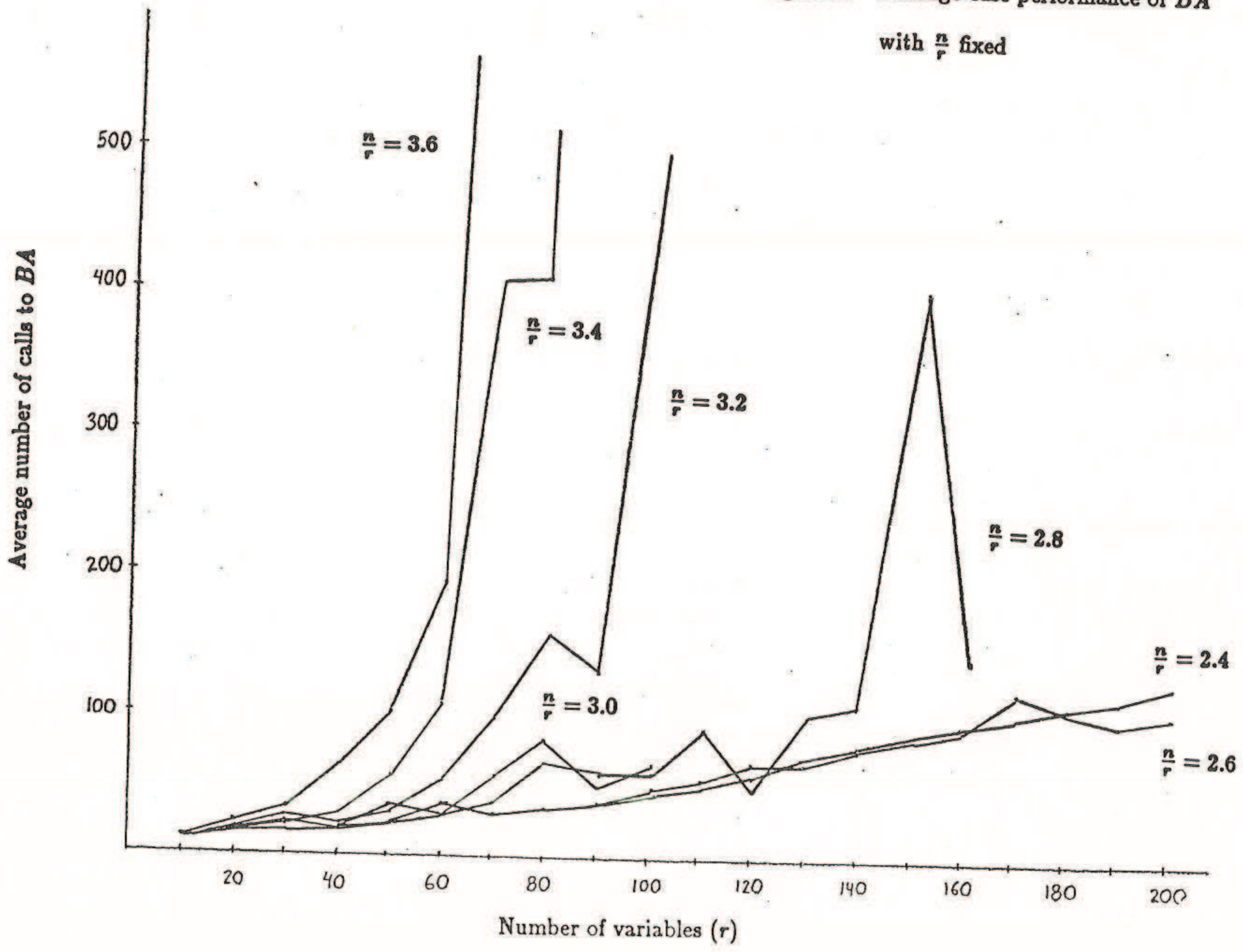
[7] Franco, J. and Paull, M., "Probabilistic analysis of the Davis Putnam Procedure for solving the satisfiability problem," *Discrete Applied Mathematics* **5** (1983), pp 77-87.

[8] Goldberg, A., "Average case complexity of the satisfiability problem," proc. 4^{th} Workshop on Automated Deduction (1979), pp 1-6.

[9] Goldberg, A., Purdom,P.W. and Brown, C.A., "Average time analysis of simplified Davis-Putnam procedures," *Information Processing Letters* **15** (1982), pp 72-75.

[10] Purdom, P.W., "Search rearrangement backtracking and polynomial average time," *Artificial Intelligence* **21** (1983), pp 117-133.

[11] Purdom, P.W. and Brown, C.A., "The pure literal rule and polynomial average time," to appear in *SIAM J. Comput.*.

Figure 1. - Average case performance of *BA* with $\frac{n}{r}$ fixed