Optimal Testing for Planted Satisfiability Problems

QUENTIN BERTHET*

Princeton University

Abstract. We study the problem of detecting planted solutions in a random satisfiability formula. Adopting the formalism of hypothesis testing in statistical analysis, we describe the minimax optimal rates. We also address algorithmic issues, and describe the performance of a new computationally efficient test. This result is compared to a related hypothesis on the hardness of detecting satisfiability of random formulas.

AMS 2000 subject classifications: Primary 62C20; secondary 68R01, 60C05.

Key words and phrases: Satisfiability problem, High-dimensional detection, Polynomial-time algorithms.

INTRODUCTION

We study in this paper the problem of detecting a planted solution in a random \( k\)-\textsc{SAT} formula of \( m \) clauses on \( n \) variables. This is formulated as an hypothesis testing problem: Given a formula \( \phi \), our goal is to decide whether it is a typical instance, drawn uniformly among all formulas, or if it has been drawn such that it is artificially satisfiable, by planting a solution.

In the field of statistics, there has been a recent resurgence of hypothesis testing, or detection problems, i.e. distinguishing null hypotheses with pure noise, against the presence of a structured signal in a high-dimensional setting. The seminal work of [Ing82, Ing98, DJ04], on the problem of detecting sparse, or weekly sparse signals has inspired a wide literature, such as [ITV10] for sparse linear regression, [ACCD11, BI13, ACV13, MW13] for small cliques, clusters, or communities in graphs and matrices, [ABBDL10] for general combinatorial structured signals in averages, and [ACBL12, BR12, BR13] for sparse principal components of covariance matrices. Several of these problems are combinatorial in nature, and the complexity of the class of possible signals (sparse vectors, cliques in a graph, small submatrices, or here the \( n\)-dimensional hypercube) has a direct influence on the statistical and algorithmic difficulties of the detection problem.

The minimax theory gives a formal definition of the statistical complexity of an hypothesis testing problem. The optimal rate of detection (in a minimax sense) is the sample size needed to identify with high probability the underlying distribution of given instances. It is said to be optimal when all methods fail to

*Partially supported by NSF grant CAREER-DMS-1053987.
significantly outperform a random guess given less samples than the rate, and when at least one test successfully distinguishes the two distributions, with high probability, given more samples than the rate (up to constants). This setting describes the interplay between the interesting parameters of a problem: sample size, ambient dimension, signal-to-noise ratio, sparsity, underlying dimension, etc.

This framework is particularly adapted to the study of random instances of \( k \)-SAT formulas: a random formula \( \phi \) can be interpreted as \( m \) independent, identically distributed clauses, each on \( k \) of the \( n \) variables. The uniform distribution is equivalent to pure noise, the absence of signal. Planting a solution is equivalent to changing the distribution of the clauses, dependent on an assignment \( x \in \{0, 1\}^n \). This planted satisfying assignment is the signal whose presence we seek to detect. The optimal rate of detection will describe how large \( m \) (the sample size) needs to be for detection to be possible, as a function of \( n \) (the ambient dimension), and \( k \), treated as a constant.

The properties of random instances of uniform \( k \)-SAT formulas have been widely studied in the probability and statistical physics literature. Particular attention has been paid to the notions of satisfiability thresholds (sharp changes of behavior when the clause-to-variable density ratio \( \Delta = m/n \) varies) [AP04, AM06, CO09, COP13], maximum satisfiability [ANP03] geometry of the space of solutions [ANP03, ART06, ACO08], and concentration of specific statistics [AM10, AM13]. The planted distribution has also been studied, often in order to create random instances that are known to be satisfiable, such as in [BHL+01, HJKN06, AGKS00, AJM04, ACO08, JMS05], and at high density in [AMZ06, CoKV07, FMV06]. More recently, the algorithmic complexity (in a specific computational model) of estimating the planted assignment has been studied in [FPV13].

One of the contributions of this work is the introduction of a statistical point of view to this problem. The use of tools from statistical analysis, such as the likelihood ratio and the statistical (or total variation) distance, highlights the importance of a specific statistic: the number of satisfying assignments. More specifically, we will study its behavior around its expected value. In order to obtain the optimal rates of detection, we prove new results concerning the concentration (or absence thereof) of this statistic. We also bring forth evidence of interesting phenomenons outside of the linear regime, usually the sole asymptotic regime considered in the study of random satisfiability formulas. We give optimal rates of detection for this statistical problem, and address algorithmic issues by studying a new polynomial-time test.

The two following subsections of the introduction are a short review of basic notions on \( k \)-SAT formulas and hypothesis testing problems in statistics, for a reader less familiar with one of these two literatures. Our hypothesis testing problem is formally described in Section 1. The optimal rates of detection are derived in Section 2, and the problem of testing in polynomial time is addressed in Section 3. The effect on the rates of detection of alternative choices of planting distributions is studied in Section 4.

**Preliminaries on \( k \)-SAT formulas**

Let \( n \) and \( m \) be positive integers. For all fixed positive integer \( k \), we denote by \( \mathcal{F}_{n,m}^k \) the set of boolean formulas on \( n \) variables that are the conjunction of
\( m \) disjunctions of \( k \) distinct literals. Formally, for all \( \phi \in \mathcal{F}_{n,m}^k \), we have for all \( x \in \{0,1\}^n \)

\[
\phi(x) = \bigwedge_{i=1}^m C_i(x),
\]

where for all \( i \in \{1,\ldots,m\} \), the clause \( C_i \) is the disjunction of \( k \) littorals on \( k \) distinct variables

\[
C_i(x) = x_{i_1} \lor \ldots \lor x_{i_k}, \quad x_{i_j} \in \{x_1, \bar{x}_1, \ldots, x_n, \bar{x}_n\}, \quad \text{and} \quad x_{i_j} \not\in \{x_{i_{j'}}, \bar{x}_{i_{j'}}\}.
\]

The \( k\text{-SAT} \) problem (short for satisfiability) is the name given to the task of determining whether a given formula \( \phi \) is satisfiable, i.e. if there exists \( x \in \{0,1\}^n \) such that \( \phi(x) \) evaluates to 'true'. We denote, for a given \( k\text{-SAT} \) formula \( \phi \), the set \( S(\phi) \) of satisfying assignments

\[
S(\phi) = \{ x \in \{0,1\}^n : \phi(x) = \text{true} \},
\]

and by \( Z(\phi) = |S(\phi)| \) the cardinality of this set, the number of satisfying assignments for \( \phi \). This definition is extended to single clauses and sub-formulas in general, with the notations \( S(C_i) \) and \( S(\phi_S) \) for the set of assignments satisfying respectively, the clause \( C_i \) or all the clauses in \( \phi_S \), when \( S \subset \{1,\ldots,n\} \). We often write \( Z \) when it is not ambiguous. We denote by \( \text{SAT} \) the set of formulas such that \( S(\phi) \neq \emptyset \), or equivalently \( Z(\phi) > 0 \).

For all \( x \in \{0,1\}^n \), \( x \in S(\phi) \) if and only if \( x \) satisfies all the clauses of \( \phi \). We can therefore write

\[
Z = \sum_{x \in \{0,1\}^n} \prod_{i=1}^m \mathbf{1}\{x \in S(C_i)\}.
\]

If a formula \( \phi \) is drawn uniformly at random, by choosing \( m \) independent clauses uniformly among the \( 2^k \binom{n}{k} \) possible choices, the expected value of \( Z \) can be easily derived. By applying successively linearity of the expectation, invariance of the distribution by any arbitrary flipping of the littorals, and independence of the clauses, it holds that

\[
\mathbf{E}[Z] = \sum_{x \in \{0,1\}^n} \mathbf{E} \left[ \prod_{i=1}^m \mathbf{1}\{x \in S(C_i)\} \right]
\]

\[
= 2^n \mathbf{E} \left[ \prod_{i=1}^m \mathbf{1}\{x_1 \in S(C_i)\} \right] = 2^n \prod_{i=1}^m \mathbf{E} \left[ \mathbf{1}\{x_1 \in S(C_i)\} \right]
\]

\[
= 2^n (1 - 2^{-k})^m = \left[2(1 - 2^{-k})^{\Delta}\right]^n,
\]

where \( \Delta = m/n \) is the clause-to-variable density ratio, and \( x_1 \) is the vector of all ones. From this closed-form formula, we can derive a simple upper bound for the probability of satisfiability of a uniformly drawn formula

\[
\mathbf{P}(\phi \in \text{SAT}) = \mathbf{P}(Z(\phi) > 0) \leq \mathbf{E}[Z(\phi)] = \left[2(1 - 2^{-k})^{\Delta}\right]^n.
\]

For \( \Delta \) large enough so that \( 2(1 - 2^{-k})^{\Delta} < 1 \), the probability is exponentially small in \( n \) (\( \Delta > 2^k \log(2) \) is sufficient). This density ratio is proved to be close
to optimal, up to a $O(k)$ in [AP04], and up to a constant independent of $k$ in [COP13]. To derive a lower bound for $\Delta$ in the satisfiability threshold for, the first intuition is to use a similar approach, through the Paley-Zygmund inequality

$$P(\phi \in \text{SAT}) = P(Z(\phi) > 0) \geq \frac{E[Z(\phi)^2]}{E[Z(\phi)]^2}.$$ 

It would be sufficient to prove that the final ratio is greater than a positive constant $c > 0$, uniformly in $n$. Indeed, it is noted in [FB99] that if $\phi$ is satisfiable with a uniformly positive probability (independent of $n$) for some $\Delta_k$, then it is satisfiable with probability exponentially close to 1 for $\Delta > \Delta_k$. Unfortunately, it holds that for some $\beta(k, \Delta) > 0$ (see [AP04])

$$E[Z^2] > [1 + \beta(k, \Delta)]^n E[Z^2].$$

To prove a lower bound for the satisfiability threshold, the authors use a similar “second moment” approach, with a modified random variable.

**Preliminaries on hypothesis testing problems**

In an hypothesis testing (or detection) problem, the goal is to identify the underlying distribution of a dataset. Given an instance $X \in \mathcal{X}$ and two distributions $P_0$ and $P_1$ on $\mathcal{X}$, we aim to distinguish the two hypotheses

$$H_0 : X \sim P_0$$
$$H_1 : X \sim P_1.$$ 

The hypotheses can be more complex (e.g. composite alternatives, etc) but we will only consider here this simple setting. A test is a measurable function of the data $\Psi : \mathcal{X} \mapsto \{0, 1\}$ that estimates whether the instance was generated with distribution $P_0$ or $P_1$. We define the probability of error as the maximum of the probabilities of type I and type II error, formally

$$P_0(\Psi(\phi) = 1) \lor P_1(\Psi(\phi) = 0).$$

This quantity measures the success of any test $\Psi$. It is possible to obtain a lower bound for the risk of error of any test

$$P_0(\Psi(\phi) = 1) \lor P_1(\Psi(\phi) = 0) \geq \frac{1}{2} (P_0(\Psi(\phi) = 1) + P_1(\Psi(\phi) = 0) \geq \frac{1 - d_{TV}(P_0, P_1)}{2},$$

where the total variation (or statistical) distance $d_{TV}$ bounds the difference of probability of any event between two distributions on a probability space $\Omega$. Formally, we have for a finite space $\Omega$ (as it is the case here)

$$d_{TV}(P, Q) = \sup_{A \in 2^\Omega} |P(A) - Q(A)| = \frac{1}{2} \sum_{\omega \in \Omega} |P(\omega) - Q(\omega)|.$$ 

It is also possible to derive an upper bound for the probability of error of a specific test, the likelihood ratio test, defined by

$$\Psi(X) = 1 \{P_1(X) / P_0(X) > 1\}.$$
Indeed, for this test, the sum of type I and type II errors is equal to 

\[ 1 - d_{TV}(P_0, P_1) \]

and the following holds

\[
P_0(\Psi(\phi) = 1) \lor P_1(\Psi(\phi) = 0) \leq P_0(\Psi(\phi) = 1) + P_1(\Psi(\phi) = 0) \leq 1 - d_{TV}(P_0, P_1).
\]

The Neyman-Pearson lemma shows that this test is the most powerful of a given size (allowed type I error). When the data \( X \) is \( N \) i.i.d copies of a random variable, we can define the optimal rate of detection as a function of the parameters of the problem (the vector \( p \)), similarly to \( [Tsy09] \).

**Definition 1.** The rate \( N^*(p) \) is called optimal if for fixed \( \delta, \nu \in (0, 1/2) \), there exists constants \( c(\delta), c'(\nu) \) such that

- For \( N \geq c(\delta)N^*(p) \), there exists a test with probability of error smaller than \( \delta \).

- For \( N \leq c'(\nu)N^*(p) \), all tests have a probability of error greater than \( 1/2 - \nu \) (not significantly better than a random guess).

1. **Problem Description**

We are interested in distinguishing two distributions on \( F_{m,n} \), the uniform, and planted distributions. The uniform distribution, denoted by \( P_{\text{unif}} \), is generated by independently selecting each clause uniformly from the \( 2^k \binom{n}{k} \) possible choices. The planted distribution, denoted by \( P_{\text{planted}} \), is generated by randomly selecting an assignment \( x^* \) uniformly among the \( 2^n \) elements of \( \{0,1\}^n \), and then independently selecting all the clauses among the \( (2^k - 1) \binom{n}{k} \) clauses that are satisfied by \( x^* \) (denoted by \( P_{x^*} \)). We represent this as an hypothesis testing problem, on the observation \( \phi \in F_{m,n}^k \)

\[
H_0 : \phi \sim P_{\text{unif}} \\
H_1 : \phi \sim P_{\text{planted}} = \frac{1}{2^n} \sum_{x \in \{0,1\}^n} P_x.
\]

We will focus mainly on two regimes for the study of our testing problem.

- The linear regime, when \( m = \Delta n \), for some \( \Delta > 0 \).
- The square-root regime, when \( m = C\sqrt{n} \), for some \( C > 0 \).

For both regimes, we will often consider \( m, n \) large enough, but will mainly focus on non-asymptotic results.

We define a test as a measurable function \( \Psi : F_{m,n}^k \rightarrow \{0,1\} \), whose goal is to determine the underlying distribution of the observation \( \phi \). We define the probability of error as the maximum of the probabilities of type I and type II error, formally

\[
P_{\text{unif}}(\Psi(\phi) = 1) \lor P_{\text{planted}}(\Psi(\phi) = 0).
\]

This quantity is used here to measure the success of any test \( \Psi \). We will consider that a test is successful when its probability of error is smaller than some \( \delta \in (0,1) \), considered fixed for the whole problem, such as \( \delta = 0.05 \).
2. OPTIMAL TESTING

The goal of this section is to derive the optimal levels of detection for this hypothesis testing problem, i.e. describing how large $m$ should be for a test to be able to distinguish with high probability the two hypotheses. We will first describe the performances of several explicit tests, and prove that the performance of the likelihood ratio test is optimal, from an information theoretic point of view.

2.1 Satisfiability testing

As planting a satisfying assignment artificially makes the formula $\phi$ satisfiable, a natural test is to check whether $\phi \in \text{SAT}$. Indeed, on the one hand $P_{\text{planted}}(\phi \notin \text{SAT}) = 0$ by definition of the planted distribution, so this test has a probability of error of type II equal to zero. On the other hand, the behavior of $P_{\text{unif}}(\phi \in \text{SAT})$ as a function of $\Delta$ is a very well studied problem, and we use the following results from [COP13], where the transition is proved to happen around $2^k \log(2)$, in an interval of bounded size, independent of $k$.

**Proposition 2.** There exists $\tilde{\Delta}_k > 0$ smaller than $2^k \log(2)$, such that for $\Delta > \tilde{\Delta}_k$, when $m = \Delta n \to +\infty$,

$$P_{\text{unif}}(\phi \in \text{SAT}) \to 0.$$ 

There exists $\Delta_k > 0$ greater than $2^k \log(2) - O(1)$, such that for $\Delta < \Delta_k$, when $m = \Delta n \to +\infty$,

$$P_{\text{unif}}(\phi \in \text{SAT}) \to 1.$$ 

This proposition gives a precise idea of how successful the satisfiability test $\Psi_{\text{SAT}} = 1\{\phi \in \text{SAT}\}$ is. When $\Delta$ is greater than $\Delta_k$, the test will have a probability of error going to 0, and when $\Delta$ is smaller than $\Delta_k$, the error will converge to 1 (entirely because of the probability of a type I error). It is conjectured that this transition happens sharply at a constant.

When thinking of the formula $\phi$ as a sequence of i.i.d. clauses, $m$ can be interpreted as the sample size, and it is intuitive that the problem becomes easier when $\Delta$ increases. When $\Delta$ is small enough, there are not enough observations for the test $\Psi_{\text{SAT}}$ to perform well, and its probability of error converges to 1. It seems at first that in the linear regime, the same phenomenon will occur for any test, and that there must exist an absolute constant $\Delta'_k$ under which all tests will fail with high probability. We see in the following subsection that it is surprisingly not the case.

2.2 Likelihood-ratio testing

A test based on the likelihood ratio between the two candidate distributions can distinguish between them with high probability, in the square-root regime. When $m \geq C \sqrt{n}$, the probability of error of the likelihood ratio test is smaller than $\delta \in (0,1)$ for $m,n$ large enough, and $C$ greater than a constant $C_{k,\delta}$, we prove the following.

**Theorem 3.** For all $k \geq 2$, positive $m,n$, denote $\Psi_{\text{LR}}$ the likelihood ratio test defined by

$$\Psi_{\text{LR}}(\phi) = 1\{Z(\phi) > E[Z]\}.$$
For all $\delta \in (0, 1)$, there exists $C_{k, \delta} > 0$ such that for $m \geq C_{k, \delta} \sqrt{n}$, for $m, n$ large enough, it holds

$$
P_{\text{unif}}(\Psi_{LR}(\phi) = 1) \lor P_{\text{planted}}(\Psi_{LR}(\phi) = 0) \leq \delta.
$$

**Proof.** We first prove that this is the right form for the likelihood ratio test. For all $\phi \in \mathcal{F}_{m,n}^k$, it holds

$$
\frac{P_{\text{planted}}(\phi)}{P_{\text{unif}}(\phi)} = \frac{1}{2^n} \sum_{x \in \{0,1\}^n} \frac{P_x(\phi)}{P_{\text{unif}}(\phi)} = \frac{1}{2^n} \sum_{x \in \{0,1\}^n} \frac{P_x(\phi)}{P_{\text{unif}}(\phi)}.
$$

To compute the probabilities in the above ratios, we can see that $\phi$ can be drawn by throwing $m$ balls in $N = 2^k \binom{n}{k}$ bins independently if it has distribution $P_{\text{unif}}$, or otherwise in the $N_x = \binom{2^k - 1}{k}$ bins corresponding to clauses that are satisfied by $x$. Denoting $b_1, \ldots, b_N$ the number of balls in each bin, this yields, for all $\phi$

$$
P_{\text{unif}}(\phi) = \frac{m!}{b_1! \ldots b_N!} \frac{1}{N^m}
$$

$$
P_x(\phi) = \begin{cases} 
\frac{m!}{b_1! \ldots b_N!} \frac{1}{N^m}, & \text{if } x \notin S(\phi) \\
0, & \text{otherwise}
\end{cases}
$$

Therefore, the likelihood ratio can be expressed in terms of $1\{x \in S(\phi)\}$

$$
\frac{P_{\text{planted}}(\phi)}{P_{\text{unif}}(\phi)} = \frac{1}{2^n} \sum_{x \in \{0,1\}^n} \left( \frac{N}{N_x} \right)^m 1\{x \in S(\phi)\} = \frac{1}{E[Z]} \sum_{x \in \{0,1\}^n} 1\{x \in S(\phi)\} = \frac{Z(\phi)}{E[Z]}.
$$

The likelihood ratio test is indeed $\Psi_{LR}(\phi) = 1\{Z(\phi) > E[Z]\}$. It is now sufficient to prove $P_{\text{unif}}(\Psi(\phi) = 1) + P_{\text{planted}}(\Psi(\phi) = 0) \leq \delta$, as the maximum of two non-negative numbers is smaller than their sum. By the statistical (or total variation) distance inequality,

$$
P_{\text{unif}}(\Psi_{LR}(\phi) = 1) + P_{\text{planted}}(\Psi_{LR}(\phi) = 0) = 1 - d_{TV}(P_{\text{unif}}, P_{\text{planted}}).
$$

Furthermore, by definition of the statistical distance

$$
d_{TV}(P_{\text{unif}}, P_{\text{planted}}) = \sum_{\phi \in \mathcal{F}_{m,n}^k} \{P_{\text{unif}} - P_{\text{planted}}\}(\phi)
$$

$$
= \sum_{\phi \in \mathcal{F}_{m,n}^k} \left( 1 - \frac{Z(\phi)}{E[Z]} \right) P_{\text{unif}}(\phi) = E \left[ \left( 1 - \frac{Z(\phi)}{E[Z]} \right)_{+} \right].
$$

In the square root regime, for $C$ large enough and $m \geq C \sqrt{n}$, we obtain by Lemma 4 (the statistical distance between distributions of i.i.d. elements is non-decreasing in the sample size)

$$
d_{TV}(P_{\text{unif}}, P_{\text{planted}}) \geq (1 - e^{-\gamma_k C^2/C_0})(1 - C_0/C^2).
$$

This bound yields the desired result for some large enough constant $C_{k, \delta} > 0$. $\square$
The proof of this theorem indicates that it is essentially possible to distinguish the two distributions whenever \( Z \) is not concentrated around its expectation for the uniform distribution. Indeed, the statistical distance, or total variation distance is equal to the expected distance (in absolute value) between the ratio \( \frac{Z}{\mathbf{E}[Z]} \) and 1. Our result is a consequence of the following lemma, that states that in the square-root regime, for a constant \( C \) large enough, the ratio \( \frac{Z}{\mathbf{E}[Z]} \) is much smaller than 1, with high probability.

**Lemma 4.** For all \( k \geq 2 \), \( m = C \sqrt{n} \), \( C_0 \) an absolute constant and \( C, n \) large enough, it holds with probability \( 1 - \frac{C_0}{C^2} \), for some constant \( \gamma_k > 0 \) that

\[
Z < e^{-\gamma_k C^2/C_0} \mathbf{E}[Z].
\]

The behavior of \( Z \) with respect to its expectation has been studied substantially in the linear regime. It is known that for \( \Delta \) small enough, \( n^{-1} \log(Z) \) converges to the limit of \( n^{-1} \mathbf{E}[\log(Z)] \), the quenched average (the concentration of this variable is studied in [AM10]). In the following lemma, we prove that this limit is different from the constant \( n^{-1} \log(\mathbf{E}[Z]) \), the annealed average, for all \( \Delta > 0 \). Lemma 4 states that in the square-root regime, the ratio between \( Z \) and \( \mathbf{E}[Z] \) is smaller than a constant, with high probability. In the linear regime, with probability converging to 1, this ratio is exponentially small.

**Lemma 5.** For all \( k \geq 2 \), \( \Delta > 0 \), and \( m = \Delta n \) large enough, if \( \phi \sim \mathbf{P}_{\text{unif}} \), it holds with probability \( 1 - o(1) \), for some constant \( c_{k, \Delta} > 0 \) that

\[
Z < e^{-c_{k, \Delta} n} \mathbf{E}[Z].
\]

This phenomenon is hinted at in [ACO08, CO09], and proven to hold for \( \Delta \) large enough in [COP13], with an explicit lower bound for the difference. We prove here that these averages are different for all \( \Delta \) and \( k \), with a gap greater than \( c_{k, \Delta} \), for which we give no explicit formula. We provide a proof for lemma 4 and 5 in Appendix A.

### 2.3 Information-theoretic lower bound

The proof of Theorem 3 also hints at the location of lower bounds for this statistical problem. The statistical distance \( d_{TV} \) between the uniform and planted distributions is close to 0 (and the statistical problem is impossible) when \( Z(\phi) \) is concentrated around its expectation. If \( m = 1 \), the satisfying assignments are exactly the ones that satisfy the only clause and \( Z(\phi) = 2^n(1 - 2^{-k}) = \mathbf{E}[Z] \). The statistical distance between the uniform and planted distribution is 0 and the two distributions are identical.

Similarly, the number of satisfying assignments will be equal to its expectation (and therefore, the two distributions will be indistinguishable) whenever no variable appears in two different clauses. Indeed, when this is the case, the set of satisfying assignments is well understood. There are \( m \) clauses on \( m \) distinct groups of \( k \) distinct variables. Each of these clauses allows a specific group of \( k \) variable to take exactly \( 2^k - 1 \) values, and the \( n - km \) remaining variables are free. When counting the satisfying assignments, there are therefore \( (2^k - 1)^m \) possible values for the constrained variables and \( 2^{n - km} \) possible values for the \( n - km \)
remaining. Overall, \( Z = (2^k - 1)^m 2^{n-km} = 2^n (1 - 2^{-k})^m = \mathbb{E}[Z] \). We use this observation to give the following lower bound.

**Theorem 6.** For \( m \leq 2 \sqrt{\nu n} / k \), it holds that

\[
\inf \{ P_{\text{unif}}(\Psi(\phi) = 1) \lor P_{\text{planted}}(\Psi(\phi) = 0) \} \geq \frac{1}{2} - \nu.
\]

**Proof.** We use the statistical distance bound, for any test \( \Psi \)

\[
P_{\text{unif}}(\Psi(\phi) = 1) \lor P_{\text{planted}}(\Psi(\phi) = 0) \geq \frac{1}{2} (P_{\text{unif}}(\Psi(\phi) = 1) + P_{\text{planted}}(\Psi(\phi) = 0))
\]

\[
\geq 1 - d_{TV}(P_{\text{unif}}, P_{\text{planted}}).
\]

We denote by \( F \) the set of formulas where no variable appears in two different clauses.

\[
d_{TV}(P_{\text{unif}}, P_{\text{planted}}) = \frac{1}{2} \sum_{\phi \in F_{m,n}} |P_{\text{unif}} - P_{\text{planted}}|(|\phi|)
\]

\[
= \frac{1}{2} \sum_{\phi \in F} |P_{\text{unif}} - P_{\text{planted}}|(|\phi|) + \frac{1}{2} \sum_{\phi \in F^c} |P_{\text{unif}} - P_{\text{planted}}|(|\phi|)
\]

\[
= \frac{1}{2} \sum_{\phi \in F} \left| \frac{Z(\phi)}{\mathbb{E}[Z]} - 1 \right| P_{\text{unif}}(\phi) + \frac{1}{2} \sum_{\phi \in F^c} |P_{\text{unif}} - P_{\text{planted}}|(|\phi|)
\]

As noticed above, for all \( \phi \in F \), \( Z(\phi) = \mathbb{E}[Z] \), and the likelihood ratio is equal to 1, as \( P_{\text{unif}}(\phi) = P_{\text{planted}}(\phi) \). Therefore the first term of this equation is equal to 0. Furthermore, \( P_{\text{unif}}(F) = P_{\text{planted}}(F) \), and the second term is upper bounded by \( P_{\text{unif}}(F^c) = P_{\text{planted}}(F^c) \). To obtain the desired lower bound, it is sufficient to prove that \( P_{\text{unif}}(F^c) \leq 2\nu \). This is a variant of the “birthday problem”: We place a group of \( k \) balls in distinct bins uniformly at random, \( m \) times independently. The probability that none of these \( m \) groups intersect is equal to \( P_{\text{unif}}(F) \). When \( i \) groups have already been drawn, occupying \( ki \) bins, the probability that one of the next \( k \) balls falls in an occupied bin is smaller than \( k^2i/n \) (the expected number of such collisions, under the hypergeometric distribution). As \( k^2(m-1)/n < 1/2 \), the following holds

\[
P_{\text{unif}}(F) \geq \prod_{i=1}^{m-1} \left( 1 - \frac{k^2i}{n} \right) > \prod_{i=1}^{m-1} e^{-2k^2i/n} = e^{-k^2(m-1)(m-2)/n} > 1 - k^2m^2/n.
\]

This gives the desired result. \( \square \)

From the last two results, we can conclude that the optimal rate of detection, as defined in Definition 1, is \( m^* = \sqrt{n} \). When \( m = C \sqrt{n} \), we know that detection is possible with probability of error smaller than \( \delta \), for \( C \) greater than some constant \( C_{k,\delta} \). This is achieved by using the likelihood ratio test based on the number of satisfying assignments \( Z(\phi) \). It is impossible to distinguish the two hypotheses with error probability smaller than \( 1/2 - \nu \) for \( C < C_{k,\nu} := 2\sqrt{\nu}/k \): no test performs significantly better than a coin flip. No effort has been made to optimize (or even quantify) the constant \( C_{k,\delta} \), as a function of \( k \) and \( \delta \).
3. POLYNOMIAL-TIME TESTING

For \( k \geq 2 \), computing the outcome of the likelihood-ratio test involves solving an \#P-complete problem, and for \( k \geq 3 \), even computing the outcome of the satisfiability test \( \Psi_{\text{SAT}} \) (which is already suboptimal) is equivalent to solving an \( \text{NP} \)-hard problem. It is therefore legitimate to examine what can be achieved in polynomial time. The testing methods described in the previous section are not computationally efficient: determining if a formula is satisfiable is the quintessential hard problem, the first known to be \( \text{NP} \)-complete [Coo71, Lev73], at the root of the web of problems known to be in the same class [Kar72]. The \( k\text{-SAT} \) problem is \( \text{NP} \)-complete for \( k \geq 3 \), so none of the tests described above can be computed in a computationally efficient manner.

3.1 Testing with candidate assignments

Finding a satisfying assignment in formulas that are known to be satisfiable has been the focus of substantial efforts [BMZ02, Fla02, KV06, CoKV07]. A polynomial-time algorithm that does so in the linear regime is presented in [CoKV07], for the case \( k = 3 \) (their results extend to any fixed \( k \)). More specifically, they prove that for a large enough non-explicit \( \Delta \), if \( \phi \) is distributed uniformly on the set of satisfiable formulas (we denote this distribution by \( P_{\text{SAT}} \)), their algorithm outputs in a polynomial time \( P(n) \) a candidate assignment \( x_{\text{Alg}}(\phi) \), which is satisfying with high probability.

This can be used as a tool for detection. In their analysis, they prove that their approach is still valid when the distribution \( P_{\text{SAT}} \) is replaced by the planted distribution \( P_{\text{planted}} \). Therefore, one could design the following test: On any given formula \( \phi \), let this algorithm run for a time \( P(n) \), and reject the null if and only if the assignment \( x_{\text{Alg}}(\phi) \) returned by the algorithm is satisfying (which is checked in linear time). If \( \phi \sim P_{\text{planted}} \), the test will be correct with high probability. If \( \phi \sim P_{\text{unif}} \), it is unsatisfiable with high probability, as soon as \( \Delta > \Delta_k \) by Proposition 2, and the returned assignment \( x_{\text{Alg}}(\phi) \) cannot be satisfying.

There are two issues with using this algorithm in a detection setting: the polynomial time \( P(n) \) is unknown, and the large enough constant \( \Delta \) is non-explicit.

3.2 Majority vector test

To address those issues, we design a test that can run in time \( O(mn^2m\log(m)) \), and has a probability of error lower than any \( \delta \in (0, 1) \), for \( \Delta \) greater than an explicit \( \Delta_{k,\delta} \). The majority test \( \Psi_{\text{MAJ}} \) focuses on a tool used in the algorithm of [CoKV07]: the majority vector \( M \in \mathbb{R}^n \), defined in the following way. For \( i \in \{1, \ldots, n\} \), \( M_i = N_+(x_i) - N_-(x_i) \), where \( N_+(x_i) \) (resp. \( N_-(x_i) \)) is the number of times the literal \( x_i \) (resp. \( \bar{x}_i \)) appears in the formula \( \phi \). Another formulation is the following, more suited to our point of view. For \( j \in \{1, \ldots, m\} \) we define for each clause \( C_j \) the vector \( S_j \in \mathbb{R}^n \) whose \( i \)-th coefficient is equal to 0 if \( x_i \) is absent of \( C_j \), and to 1 (resp. \(-1\)) if it appears positively (resp. negatively) in this clause. This is consistent with our setting where variables appear once at most in each clause. For a clause on 5 variables, the vector associated to the clause \( C(x) = \bar{x}_1 \lor \bar{x}_2 \lor x_3 \) is therefore \( S = (-1, -1, 1, 0, 0) \). With this notation, \( M \) is equal to the sum of these \( m \) vectors

\[
M = S_1 + \ldots + S_m.
\]
This formulation allows us to study $M$ as the sum of $m$ i.i.d. random vectors. Under the null, when $\phi \sim P_{\text{unif}}$, each $S_i$ is uniformly distributed on the set of vectors with coefficients in $\{-1, 0, 1\}$ with $\ell_0$ (or $\ell_1$) norm equal to $k$. On the other hand, when $\phi \sim P_{x^*}$ for some $x^* \in \{0, 1\}$, for each choice of non-zero coefficients of $S$, there is a “forbidden” sign pattern: the literals in the clause cannot be all different from those in $x^*$, so that the clause is satisfied by this assignment. This property allows us to distinguish the uniform and planted distributions, by examining the $\ell_2$ norm of the vector $M$.

**Theorem 7.** For all $k \geq 2$, positive $m, n$ and $\delta \in (0, 1)$, denote $\Psi$ the test defined by

$$\Psi_{\text{MAJ}}(\phi) = 1\{|M|_2 > \sqrt{km}(1 + \sqrt{2\log(1/\delta)})\}$$

and

$$\tilde{\Delta}_{k,\delta} := \frac{2^k - 1}{k^2}(1 + 2\sqrt{2\log(1/\delta)})^2.$$ 

For and $m \geq \tilde{\Delta}_{k,\delta} n$, it holds

$$P_{\text{unif}}(\Psi_{\text{MAJ}}(\phi) = 1) \lor P_{\text{planted}}(\Psi_{\text{MAJ}}(\phi) = 0) \leq \delta.$$

**Proof.** Let $C_k := \{S \in \mathbb{R}^n : S_i \in \{-1, 0, 1\}, |S|_0 = k\}$, and $f : C_k^n \to \mathbb{R}$ be the function with input $(S_1, \ldots, S_m)$ (each in $C_k$), such that

$$f(S_1, \ldots, S_m) = |S_1 + \ldots + S_m|_2.$$ 

The following holds, for any $(S_1, \ldots, S_m) \in C_k^m$ and $i \in \{1, \ldots, m\}$, by the triangular inequality

$$\max_{S_i' \in C_k} |f(S_1, \ldots, S_i, \ldots, S_m) - f(S_1, \ldots, S_i', \ldots, S_m)| \leq \sqrt{2k}.$$ 

By the McDiarmid or (bounded differences) inequality, it holds for all distributions that

$$P(f - \mathbb{E}f \geq \varepsilon) \leq \exp\left(-\varepsilon^2 / 2km\right).$$

Therefore, with probability $1 - \delta$, when $\phi \sim P_{\text{unif}}$

$$|v|_2 \leq \mathbb{E}_{\text{unif}}[|v|_2] + \sqrt{2km \log(1/\delta)} \leq \sqrt{\mathbb{E}_{\text{unif}}[|v|_2^2]} + \sqrt{2km \log(1/\delta)},$$

by Jensen’s inequality. Furthermore, under this distribution, the $S_i$’s are independent, centered random vectors, and the following holds

$$\mathbb{E}_{\text{unif}}[|M|_2^2] = \mathbb{E}_{\text{unif}}[\sum_{i=1}^m |S_i|^2_2] = \sum_{i=1}^m \mathbb{E}_{\text{unif}}|S_i|^2_2.$$

Overall, under the null hypothesis, with probability $1 - \delta$

$$(1) \quad |M|_2 \leq \sqrt{km} \left(1 + \sqrt{2\log(1/\delta)}\right).$$
For $\phi \sim \mathbf{P}_{x^*}$ for some $x^* \in \{0, 1\}^n$, let $z^* = 2x^* - 1 \in \{-1, 1\}^n$. The distribution of each coefficient of $S$ is the following

$$S^{(j)} = \begin{cases} 
  z^*_j & \text{with probability } \frac{1}{2} \left( 1 + \frac{1}{2^{k-1}} \right) \frac{k}{n}, \\
  0 & \text{with probability } 1 - \frac{k}{n}, \\
  -z^*_j & \text{with probability } \frac{1}{2} \left( 1 - \frac{1}{2^{k-1}} \right) \frac{k}{n}.
\end{cases}$$

Indeed, there is a probability $1 - k/n$ that $x_j$ is not in the clause, and if it is, out of the $2^k - 1$ allowed sign patterns for $S$, it appears as $x^*_j$ in $2^k - 1$ of them and as $\bar{x}^*_j$ in the $2^k - 1 - 1$ remaining.

Therefore, $\mathbf{E}_{x^*}[M] = m \mathbf{E}_{x^*}[S] = \Delta \frac{k}{2^{k-1}} z^*$. The expectation of $|M|_2$ is constant for all $x^* \in \{0, 1\}^n$ (the distribution of $|M|_2$ is invariant by change of planted assignment) and $|\mathbf{E}_{x^*}[M]|_2 = \Delta \frac{k}{2^{k-1}} \sqrt{n} = \frac{\Delta k}{2^{k-1}} \sqrt{km}$. With probability $1 - \delta$, when $\phi \sim \mathbf{P}_{\text{planted}}$, McDiarmid’s inequality yields

$$|M|_2 \geq \mathbf{E}_{\text{planted}}[|M|_2] - \sqrt{2km \log(1/\delta)}$$

$$\geq \frac{1}{2^n} \sum_{x^* \in \{0, 1\}^n} \mathbf{E}_{x^*}[|M|_2] - \sqrt{2km \log(1/\delta)}$$

$$\geq \mathbf{E}_{x^*}[|M|_2] - \sqrt{2km \log(1/\delta)}$$

$$\geq |\mathbf{E}_{x^*}[M]|_2 - \sqrt{2km \log(1/\delta)},$$

by Jensen’s inequality. Under the alternative hypothesis, with probability $1 - \delta$

$$|M|_2 \geq \sqrt{km} \left( \frac{\Delta k}{2^k-1} - \sqrt{2 \log(1/\delta)} \right).$$

Together, equations (1) and (2) give the desired result.

Therefore, the majority test $\Psi_{\text{MAJ}}$ enables to distinguish the uniform and planted distributions with probability $1 - \delta$, in the linear regime, for $\Delta$ greater than an explicit $\Delta_{k, \delta}$. It runs in polynomial time $O(mn \land m \log(m))$, treating $k$ as a constant.

### 3.3 Hardness hypothesis on random instances

This result can be contrasted with an hypothesis by Feige, formulated in [Fei02]. To prove hardness of approximation results (in the worst-case), the author proposes an hypothesis on the hardness of determining the satisfiability of 3-SAT formulas on average:

**Hypothesis 8 ([Fei02]).** “Even when $\Delta$ is an arbitrarily large constant independent of $n$, there is no polynomial time algorithm that refutes most 3CNF formulas with $n$ variables and $m = \Delta n$ clauses, and never wrongly refutes a satisfiable formula.”

Formally, it is conjectured in this hypothesis that for all $\Delta > 0$, in the linear regime, there is no test $\Psi$ that runs in polynomial time such that $\mathbf{P}_{\text{unif}}(\Psi = 1) \leq 1/2$, and $\mathbf{P}_1(\Psi = 0) = 0$, for any distribution $\mathbf{P}_1$ supported on SAT. In particular, in our testing problem, this hypothesis states that no test that runs
in polynomial time has a type I error smaller than $1/2$ and a type II error equal to 0. Theorem 7 is consistent with this hypothesis. We show that it is possible to distinguish those two hypotheses with probability of error smaller than any $\delta \in (0, 1)$, for $\Delta$ greater than an explicit constant $\Delta_{k, \delta}$. Interestingly, this result shows that in the linear regime, while it is conjectured and widely believed that it is impossible to distinguish those distributions with a completely one-sided error, it is possible to design a test with small type I and type II errors simultaneously.

There has been a recent interest in the notions of optimal rates for polynomial-time algorithms. More specifically, there is a growing literature on limitations, beyond those imposed by information theory, to statistical performance for computationally efficient procedures. Such phenomenon have been hinted at [DGR98, Ser00, CJ13, SSST12], and studied in specific computational models, such as in [FGR+13, FPV13]. More recently, these barriers have been proven to hold for various supervised tasks such as in [DLS13], based on a primitive on random 3-SAT instances, and unsupervised problems in statistics in [BR13] and the subsequent [MW13, Che13], based on a hardness hypothesis for the planted clique problem.

The above discussion shows the complexity of using Feige’s hypothesis as a primitive to prove computational lower bounds for statistical problems: it does not imply that it is impossible to detect planted distributions in a computationally efficient manner in the linear regime, and is extremely sensitive to the allowed probability of type I and type II errors. There is however a significant gap (from the square-root to the linear regime) between the optimal rate of detection and those possible for polynomial-time methods, akin to the problems cited above. Therefore, the question of computational lower bounds for detection of planted satisfiability remains open, and the following can be conjectured.

**Conjecture 9.** For any $\delta \in (0, 1/2)$, any $a > b > 0$, and all randomized tests $\xi$ that run in polynomial-time, there exists a positive constant $\Gamma$ such that for all $n^{1-a} < \Gamma m < n^{1-b}$ it holds that

$$P_{\text{unif}}(\xi(\phi) = 1) \land P_{\text{planted}}(\xi(\phi) = 1) \geq \delta.$$ 

### 4. ALTERNATIVE CHOICES FOR PLANTING DISTRIBUTIONS

The tests described in Theorems 3 and 7 exploit a fundamental difference between the two distributions. Planting a satisfying assignment $x^* \in \{0, 1\}^n$ breaks the symmetry of the uniform distribution. The likelihood ratio $Z/E[Z]$ is affected by the imbalances in interactions between variables - whether they appear in different clauses with the same sign or not. Similarly, the majority test is based on the bias in the majority vector $M$, under the planted distribution.

This asymmetry is a characteristic of our choice of planting distribution. In this section, we observe that the rates of estimation are different for other natural choices of distribution on SAT, the set of satisfiable formulas. Such an example is $P_{\text{SAT}}$, the uniform distribution on SAT. The statistical problem becomes.

$$H_0 : \phi \sim P_{\text{unif}}$$

$$\bar{H}_1 : \phi \sim P_{\text{SAT}}.$$ 

Using the notations of Proposition 2, the following asymptotic rates of detection hold.
Theorem 10. For $m \geq \Delta_k n$, when $m, n \to \infty$

$$\inf_{\Psi} \{\mathbb{P}_{\text{unif}}(\Psi(\phi) = 1) \vee \mathbb{P}_{\text{SAT}}(\Psi(\phi) = 0)\} \to 0.$$ 

For $m \leq \Delta_k n$, when $m, n \to \infty$

$$\inf_{\Psi} \{\mathbb{P}_{\text{unif}}(\Psi(\phi) = 1) \vee \mathbb{P}_{\text{SAT}}(\Psi(\phi) = 0)\} \to \frac{1}{2}.$$ 

Proof. The first result (upper bound) is direct from the first part of Proposition 2. In this regime, $\mathbb{P}_{\text{SAT}}(\phi \not\in \text{SAT}) = 0$, and $\mathbb{P}_{\text{unif}}(\phi \in \text{SAT}) \to 0$, which yields the desired result.

The second result (lower bound) is also a consequence of Proposition 2. We use the statistical distance (or total variation) bound, for any test $\Psi$

$$\mathbb{P}_{\text{unif}}(\Psi(\phi) = 1) \vee \mathbb{P}_{\text{SAT}}(\Psi(\phi) = 0) \geq \frac{1}{2}(\mathbb{P}_{\text{unif}}(\Psi(\phi) = 1) + \mathbb{P}_{\text{SAT}}(\Psi(\phi) = 0)) \geq \frac{1 - d_{\text{TV}}(\mathbb{P}_{\text{unif}}, \mathbb{P}_{\text{SAT}})}{2}.$$ 

As $\mathbb{P}_{\text{SAT}}$ is the uniform distribution on SAT, or $\mathbb{P}_{\text{unif}}(\cdot | \phi \in \text{SAT})$, the total variation distance $d_{\text{TV}}(\mathbb{P}_{\text{unif}}, \mathbb{P}_{\text{SAT}})$ is equal to $\mathbb{P}_{\text{unif}}(\phi \not\in \text{SAT})$. The second part of Proposition 2 therefore yields the desired result.

This altered statistical problem is fundamentally different. Its optimal rate of detection is the linear regime $m^* = n$, achieved by the satisfiability test $\Psi_{\text{SAT}}$. If the satisfiability threshold conjecture holds and $\Delta_k = \Delta_k$, the constant is also optimal. This altered hypothesis testing problem is a significantly harder task than the detection of planted satisfiability. For all choices of alternative distribution $\mathbb{P}_1$ supported on SAT, the satisfiability test allows to discriminate the two hypotheses, with probability of error converging to 0, for $m \geq \Delta_k n$.

Among all distributions on satisfiable formulas, the closest in statistical distance to the uniform distribution (and therefore the choice of alternative that yields the hardest statistical problem) is the uniform distribution on SAT. Other distributions used to generate formulas that are hard to solve, with hidden solutions (usually, with no immediate asymmetry) as in [AJM04, BHL+01, JMS05] are candidates to create detection problems with optimal rate of detection in the linear regime. Such an example is the uniform distribution on formulas that are not-all-equal, or NAE satisfiable.

As noticed in Section 3 the polynomial-time algorithm proposed by [CoKV07] can be used to design a test, successful with a probability of error converging to 0, for $\Delta$ large enough, in the linear regime. For this distribution, the rates of detection possible by any testing function, or by one that runs in polynomial time, are similar.

Acknowledgements

The author thanks Philippe Rigollet, Emmanuel Abbé, Dan Vilenchick and Amin Coja-Oghlan for very helpful discussions.
APPENDIX A: PROOF OF TECHNICAL RESULTS

Proof of Lemma 4 and 5. For all \( x \in \{0,1\}^n \), \( x \in S(\phi) \) if and only if \( x \) satisfies all the clauses of \( \phi \). We can therefore write

\[
Z = \sum_{x \in \{0,1\}^n} \prod_{i=1}^m 1\{x \in S(C_i)\}.
\]

We recall that this yields, for \( \phi \) drawn uniformly \( \mathbb{E}[Z] = 2^n(1 - 2^{-k})^m \).

In the proof of Theorem 6, we use that \( Z \) is equal to its expectation when the \( km \) variables in the formula are distinct. In the linear regime, or in the square-root regime for a large enough constant, it is not the case, with high probability. The interactions between the clauses that share the same variable will create an imbalance between couples of clauses where the same variables appears with the same sign, and those where it appears with a different one.

We compute the conditional expectation of \( Z \), given the first variable of each clause, and wether the first two occurrences of every variable (when there are two or more) are the same literal or not. Formally, we denote \( G = (G_1, \ldots, G_n) \) the partition of \( \{1, \ldots, m\} \) in \( n \) sets (allowing some of them to be empty), where

\[
G_i = \{ j \in \{1, \ldots, m\} : C_j(x) \in \{x_i \land \ldots, \bar{x}_i \land \ldots\} \},
\]

and \( \sigma = (\sigma_1, \ldots, \sigma_n) \), where \( \sigma_i = 0 \) if there are less than two elements in \( G_i \), \( \sigma_i = 1 \) if the first two elements of \( G_i \) have the same first literal (either both \( x_i \) or both \( \bar{x}_i \)), and \( \sigma_i = -1 \) otherwise. By linearity of expectation, it holds

\[
\mathbb{E}[Z \mid (G, \sigma)] = \sum_{x \in \{0,1\}^n} \mathbb{E}
\left[ 1\{x \in S(\phi)\} \mid (G, \sigma) \right].
\]

We now observe that this conditional expectation is constant, for all \( x \in \{0,1\}^n \). Indeed, let \( x_0 \) be the assignment of all zeroes, and \( t_x \) be the literal-flipping transformation such that \( t_x(x_0) = x \), and \( T_x \) the corresponding literal-flipping transformation on formulas. For all \( x \), it holds

\[
\phi(x) = \phi(t_x(x_0)) = (T_x\phi)(x_0).
\]

For all \( x \), \( T_x\phi \) also has distribution \( \mathbf{P}_{\text{unif}} \), and \( (G, \sigma) \) is invariant by this transformation. Therefore, it holds

\[
\mathbb{E}[Z \mid (G, \sigma)] = \sum_{x \in \{0,1\}^n} \mathbb{E}
\left[ 1\{x \in S(\phi)\} \mid (G, \sigma) \right] = 2^n \mathbb{E}
\left[ 1\{x_0 \in S(\phi)\} \mid (G, \sigma) \right].
\]

The assignment \( x_0 \) will satisfy the formula \( \phi \) if and only if it satisfies all the sub-formulas \( \phi_{G_1}, \ldots, \phi_{G_n} \) (the empty formula is always satisfied). Given \( (G, \sigma) \),
the events \( \{x_0 \in S(\phi_{G_i})\} \) are independent: the sub-formulas are satisfied by \( x_0 \) if and only if every clause contains at least one negated literal, which occurs independently, conditioned on \((G, \sigma)\). We can therefore compute the conditional expectation

\[
E \left[ 1 \{x_0 \in S(\phi)\} \mid (G, \sigma) \right] = E \left[ \prod_{i=1}^{n} 1 \{x_0 \in S(\phi_{G_i})\} \mid (G, \sigma) \right] = \prod_{i=1}^{n} E \left[ 1 \{x_0 \in S(\phi_{G_i})\} \mid (G, \sigma) \right] = \prod_{i=1}^{n} E \left[ 1 \{x_0 \in S(\phi_{G_i})\} \mid (G_i, \sigma_i) \right]
\]

The product terms can be expressed as a function of \( g_i = |G_i| \). If \( \sigma_i = 0 \), in the case of \( g_i < 2 \), treating separately the cases \( g_i = 0 \) or \( 1 \), we have

\[
E \left[ 1 \{x_0 \in S(\phi_{G_i})\} \mid (G_i, \sigma_i = 0) \right] = \left( 1 - \frac{1}{2^k} \right)^{g_i}.
\]

If there are at least two elements in \( G_i \), we have

\[
E \left[ 1 \{x_0 \in S(\phi_{G_i})\} \mid (G_i, \sigma_i = 1) \right] = \frac{1}{2} \left[ 1 + \left( 1 - \frac{1}{2^{k-1}} \right)^2 \right] \left( 1 - \frac{1}{2^k} \right)^{g_i-2}
\]

\[
E \left[ 1 \{x_0 \in S(\phi_{G_i})\} \mid (G_i, \sigma_i = -1) \right] = \left( 1 - \frac{1}{2^{k-1}} \right) \left( 1 - \frac{1}{2^k} \right)^{g_i-2}.
\]

Overall, this yields

\[
E \left[ 1 \{x_0 \in S(\phi_{G_i})\} \mid (G_i, \sigma_i) \right] = \left[ 1 + \frac{\sigma_i}{2^k(1 - 2^{-k})^2} \right] \left( 1 - \frac{1}{2^k} \right)^{g_i}.
\]

If we denote \( p \) (resp. \( d \)) the number of groups for which \( \sigma_i = 1 \) (resp. \( -1 \)), we obtain

\[
E[Z \mid (G, \sigma)] = 2^p \left( 1 - \frac{1}{2^k} \right)^n \left[ 1 + \frac{1}{2^k(1 - 2^{-k})^2} \right]^p \left[ 1 - \frac{1}{2^k(1 - 2^{-k})^2} \right]^d.
\]

It is possible to design a set of \((G, \sigma)\), event of probability close to 1, for which this expectation has the desired value. To do so, we study the behavior of \( p \) and \( d \), the number of variables that appear at least twice among the first variables of the clauses, for which respectively \( \sigma_i = 1 \) or \(-1\).

Indeed, for a large \( p + d \), with \( p \) and \( d \) close to \((p + d)/2\), this expectation is significantly smaller than \( E[Z] \). Indeed, for all \( t \in (0, 1) \), the function \( f_t : \alpha \mapsto (1 + t)^{1+\alpha} (1 - t)^{-1-\alpha} \) is continuous and \( f_t(0) = 1 - t^2 \), so there exists \( \alpha_t \in (0, 1) \) such that \( f_t(\alpha) < 1 - t^2/2 \) for all \(|\alpha| < \alpha_t\). Therefore, there exists \( \alpha_k \in (0, 1) \) such that

\[
\left[ 1 + \frac{1}{2^k(1 - 2^{-k})^2} \right]^{1+\alpha} \left[ 1 - \frac{1}{2^k(1 - 2^{-k})^2} \right]^{1-\alpha} < 1 - \frac{1}{2^{4k+1} (1 - 2^{-k})^4} := e^{-\gamma_k},
\]

for all \(|\alpha| < \alpha_k\), for some \( \gamma_k > 0 \).
For every variable, we denote $T_i = |\sigma_i| \in \{0, 1\}$.

**Linear regime, Lemma 5**

We control $p$ and $d$ in the regime $m = \Delta n$. We can see the random formula as being drawn by independently placing $m$ balls uniformly in $n$ bins, and $T_i$ being the indicator of the event “there are at least two balls in bin $i$”. This is the complement of having either one or no ball in bin $i$, which yields

$$E[T_i] = 1 - \left[\left(1 - \frac{1}{n}\right)^m + m\left(1 - \frac{1}{n}\right)^{m-1}\frac{1}{n}\right] = 1 - \left[\left(1 - \frac{\Delta}{m}\right)^m + \Delta\left(1 - \frac{\Delta}{m}\right)^{m-1}\right],$$

which has limit $1 - (1 + \Delta)e^{-\Delta} = 2\varepsilon_\Delta > 0$. Therefore, for $m$ large enough, $E[T_i] > \varepsilon_\Delta$. By, definition of $p, d$ and $T_i$, we have

$$p + d = T = T_1 + \ldots + T_n.$$ 

Therefore, it holds $E[T] = E[T_1 + \ldots + T_n] > n\varepsilon_\Delta$. These variables are not independent and the variance is less simple

$$\text{Var}[T] = n\text{Var}[T_1] + n(n - 1)\left[E[T_1T_2] - E[T_1]E[T_2]\right].$$

We control the last term

$$E[T_1T_2] = P[T_1 = 1, T_2 = 1] = E[T_1 = 1|T_2 = 1]E[T_2 = 1] = \left[1 - \left(1 - \frac{1}{n}\right)^m + (m - 2)\left(1 - \frac{1}{n}\right)^{m-2}\right]E[T_2]$$

Therefore, we obtain the bound

$$E[T_1T_2] - E[T_1]E[T_2] \leq \left[1 - \left(1 - \frac{1}{n}\right)^2 + \Delta\left(1 - \left(1 - \frac{1}{n}\right)^2\right)\right]E[T_2] \leq \frac{3 + 3\Delta}{n}.$$

Overall, this yields $\text{Var}[T] \leq (4 + 3\Delta)n$. We now apply Chebyshev’s inequality, with $r_\Delta = (3 + 3\Delta)/(E[T_1] - \varepsilon_\Delta)^2$

$$P[T < \varepsilon_\Delta n] \leq \frac{\text{Var}[T]}{(E[T_1] - \varepsilon_\Delta)^2n^2} \leq \frac{r_\Delta}{n}.$$ 

Of these $T$ variables, between $T/2(1 + \alpha_k)$ and $T/2(1 - \alpha_k)$ will have their first two occurrences with the same literal, with probability greater than $1 - e^{-\alpha_k^2\varepsilon_\Delta n/2}$, by Hoeffding’s inequality. We call $B$ the event $T \geq n\varepsilon_\Delta$ and $p \in (T/2(1 - \alpha_k), T/2(1 + \alpha_k))$. By the above, $P(B) = 1 - o(1)$. For $(G, \sigma)$ in the event $B$, it holds

$$E[Z | (G, \sigma)] = 2^d\left(1 - \frac{1}{2^k}\right)^m \left[1 + \frac{1}{2^k(1 - 2^{-k})^2}\right]^p \left[1 - \frac{1}{2^k(1 - 2^{-k})^2}\right]^d$$

$$< 2^d\left(1 - \frac{1}{2^k}\right)^m (e^{-\gamma_k})^{T/2} < e^{-\gamma_k\varepsilon_\Delta n/2}E[Z] := e^{-2c_k\Delta n}E[Z].$$
Therefore $E[Z | B] < e^{-2ck, \Delta n} E[Z]$. We can now conclude by conditioning on $B$ and using Markov’s inequality

$$P(Z > e^{-ck, \Delta n} E[Z]) = P(Z > e^{-ck, \Delta n} E[Z] | B) P(B) + P(Z > e^{-ck, \Delta n} E[Z] | B^c) P(B^c) \leq P(Z > e^{-ck, \Delta n} E[Z] | B) + P(B^c) \leq \frac{E[Z | B]}{e^{-ck, \Delta n} E[Z]} + P(B^c) \leq e^{-ck, \Delta n} + P(B^c).$$

Which yields the desired result.

**Square-root regime, Lemma 4**

This proof is a simple modification of the proof of the linear regime in lemma 5 with the same notations, for $m = C\sqrt{n}$. We write

$$E[Z | (G, \sigma)] = 2^n \left(1 - \frac{1}{2^k}\right)^m \left[1 + \frac{1}{2^{2k}(1-2^{-k})^2}\right] \left[1 - \frac{1}{2^{2k}(1-2^{-k})^2}\right]^{d}$$

We can derive the expectation and variance of $T$

$$E[T_i] = 1 - \left[ \left(1 - \frac{1}{n}\right)^m + m \left(1 - \frac{1}{n}\right)^{m-1} \frac{1}{n} \right]$$

$$= 1 - \left[ \left(1 - \frac{1}{n}\right)^{C\sqrt{n}} + \frac{C}{\sqrt{n}} \left(1 - \frac{1}{n}\right)^{C\sqrt{n}-1} \right]$$

$$= 1 - \left[ 1 - \frac{C}{\sqrt{n}} + \frac{C^2}{2n} + o\left(\frac{1}{n}\right) \right] = \frac{C^2}{2n} + o\left(\frac{1}{n}\right).$$

Therefore, for $n$ large enough $E[T_i] \in (C^2/3n, C^2/n)$ and $E[T_i] \in (C^2/3, C^2)$. For the variance, as in the linear regime it holds

$$\text{Var}[T] = n \text{Var}[T_i] + n(n-1) [E[T_1 T_2] - E[T_1] E[T_2]].$$

We obtain in a similar way the following bound, for $n$ large enough

$$E[T_1 T_2] - E[T_1] E[T_2] \leq \left[ 1 - \left(1 - \frac{1}{n}\right)^2 + \frac{C}{\sqrt{n}} \left(1 - \left(1 - \frac{1}{n}\right)^2 \right) \right] E[T_2] \leq \frac{3}{n} \times C^2/n.$$

Therefore, $\text{Var}[T] \leq 4C^2$, and we have, using Chebyshev’s inequality

$$P[T \geq C^2/4] \leq \frac{\text{Var}[T]}{(C^2/3 - C^2/4)^2} \leq \frac{576}{C^2}.$$

As in the linear regime, of these $T$ variables, between $T/2(1 + \alpha_k)$ and $T/2(1 - \alpha_k)$ will have their first two occurrences with the same literal, with probability greater than $1 - e^{-\alpha_k^2 C^2/8}$, by Hoeffding’s inequality. We call $B$ the event $T \geq C^2/4$ and $p \in (T/2(1 - \alpha_k), T/2(1 + \alpha_k))$. By the above, $P(B) = 1 - O(1/C^2)$. For $(G, \sigma)$ in the event $B$, it holds

$$E[Z | (G, \sigma)] = 2^n \left(1 - \frac{1}{2^k}\right)^m \left[1 + \frac{1}{2^{2k}(1-2^{-k})^2}\right] \left[1 - \frac{1}{2^{2k}(1-2^{-k})^2}\right]^{d}$$

$$< 2^n \left(1 - \frac{1}{2^k}\right)^m (e^{\gamma_k})^{T/2} < e^{-\gamma_k C^2/8} E[Z].$$
Therefore $E[Z | B] < e^{-\gamma k C^2/8} E[Z]$. We can now conclude by conditioning on $B$ and using Markov’s inequality

$$P(Z > e^{-\gamma k C^2/16} E[Z]) = P(Z > e^{-\gamma k C^2/16} E[Z] | B)P(B) + P(Z > e^{-\gamma k C^2/16} E[Z] | B^c)P(B^c) \leq \frac{E[Z | B]}{e^{-\gamma k C^2/16} E[Z]} + P(B^c) \leq e^{-\gamma k C^2/8} + P(B^c).$$

This yields the second result, for $C$ large enough, and some absolute constant $C_0$. 

\[\square\]

REFERENCES

[ABBDL10] Louigi Addario-Berry, Nicolas Broutin, Luc Devroye, and Gábor Lugosi, On combinatorial testing problems, Ann. Statist. 38 (2010), no. 5, 3063–3092. MR2722464 (2011k:62035)

[ACBL12] Ery Arias-Castro, Sébastien Bubeck, and Gábor Lugosi, Detection of correlations, Ann. Statist. 40 (2012), no. 1, 412–435. MR3014312

[ACCD11] Ery Arias-Castro, Emmanuel J. Candès, and Arnaud Durand, Detection of an anomalous cluster in a network, Ann. Statist. 39 (2011), no. 1, 278–304. MR2797847 (2012a:62130)

[ACO08] Dimitris Achlioptas and Amin Coja-Oghlan, Algorithmic barriers from phase transitions, FOCS ’08 Proceedings of the 49th Annual IEEE Symposium on Foundations of Computer Science (2008), 793–802.

[ACV13] Ery Arias-Castro and Nicolas Verzelen, Community detection in random networks, Arxiv Preprint (2013).

[AGKS00] Dimitris Achlioptas, Carla Gomes, Henry Kautz, and Bart Selman, Generating satisfiable problem instances, IN AAAI/IAAI (2000), 256–261.

[AJM04] Dimitris Achlioptas, Haixia Jia, and Cristopher Moore, Hiding satisfying assignments: two are better than one, IN PROCEEDINGS OF AAAI’04 24 (2004), 131–136.

[AM06] Dimitris Achlioptas and Cristopher Moore, Random k-sat: Two moments suffice to cross a sharp threshold, SIAM Journal on Computing 36 (2006), no. 3, 740–762.

[AM10] Emmanuel Abbe and Andrea Montanari, On the concentration of the number of solutions of random satisfiability formulas, Random Structures & Algorithms (2010).

[AM13] , Conditional random fields, planted constraint satisfaction, and entropy concentration, Arxiv Preprint (2013).

[AMZ06] Fabrizio Altarelli, Rémi Monasson, and Francesco Zamponi, Can rare sat formulas be easily recognized? on the efficiency of message passing algorithms for k-sat at large clause-to-variable ratios, CoRR abs/cs/0609101 (2006).
Dimitris Achlioptas, Assaf Naor, and Yuval Peres, *On the fraction of satisfiable clauses in typical formulas*, EXTENDED ABSTRACT IN FOCS’03 (2003), 362–370.

Dimitris Achlioptas and Yuval Peres, *The threshold for random k-sat is $2^k \ln 2 - o(k)$*, J. Amer. Math. Soc. 17 (2004), 947–973.

Dimitris Achlioptas and Federico Ricci-Tersenghi, *On the solution-space geometry of random constraint satisfaction problems*, STOC ’06: Proceedings of the thirty-eighth annual ACM symposium on Theory of computing (2006), 130–139.

Wolfgang Barthel, Alexander K. Hartmann, Michele Leone, Federico Ricci-Tersenghi, Martin Weigt, and Riccardo Zecchina, *Hiding solutions in random satisfiability problems: A statistical mechanics approach*, CoRR cond-mat/0111153 (2001).

Cristina Butucea and Yuri I. Ingster, *Detection of a sparse submatrix of a high-dimensional noisy matrix*, Bernoulli (to appear) (2013).

A. Braunstein, M. Mézard, and R. Zecchina, *Survey propagation: an algorithm for satisfiability*.

Quentin Berthet and Philippe Rigollet, *Optimal detection of sparse principal components in high dimension*, Ann. Statist. 41 (2012), no. 4, 1780–1815.

Cristina Butucea and Yuri I. Ingster, *Detection of a sparse submatrix of a high-dimensional noisy matrix*, Bernoulli (to appear) (2013).

Amin Coja-Oghlan, *Random constraint satisfaction problems*, Electronic Proceedings in Theoretical Computer Science 9 (2009), 32–37.

Amin Coja-oghlan, Michael Krivelevich, and Dan Vilenchik, *Why almost all k-cnf formulas are easy*, PROCEEDINGS OF THE 13TH INTERNATIONAL CONFERENCE ON ANALYSIS OF ALGORITHMS (2007).

S. A. Cook, *The complexity of theorem proving procedures*, Proceedings of the Third Annual ACM Symposium (New York), ACM, 1971, pp. 151–158.

Amin Coja-Oghlan and Konstantinos Panagiotou, *Going after the k-sat threshold*, STOC ’13 Proceedings of the 45th annual ACM symposium on Symposium on theory of computing (2013), 705–714.

Scott E. Decatur, Oded Goldreich, and Dana Ron, *Computational sample complexity*, SIAM JOURNAL ON COMPUTING 29 (1998).

David Donoho and Jiashun Jin, *Higher criticism for detecting sparse heterogeneous mixtures*, Ann. Statist. 32 (2004), no. 3, 962–994. MR2065195 (2005e:62066)

Amit Daniely, Nati Linial, and Shai Shalev Shwartz, *More data speeds up training time in learning halfspaces over sparse vectors*, Arxiv Preprint (2013).

Elhad Friedgut and Jean Bourgain, *Sharp thresholds of graph prop-
erties, and the k-sat problem, J. AMER. MATH. SOC 12 (1999), 1017–1054.

[Fei02] Uriel Feige, Relations between average case complexity and approximation complexity, Proceedings of the Thirty-Fourth Annual ACM Symposium on Theory of Computing (New York), ACM, 2002, pp. 534–543 (electronic). MR2121179

[FGR+13] Vitaly Feldman, Elena Grigorescu, Lev Reyzin, Santosh Vempala, and Ying Xiao, Statistical algorithms and a lower bound for planted clique, Proceedings of the Forty-Fifth Annual ACM Symposium on Theory of Computing, STOC 2013, 2013.

[Fla02] Abraham Flaxman, A spectral technique for random satisfiable 3cnf formulas, SODA '03 Proceedings of the fourteenth annual ACM-SIAM symposium on Discrete algorithms (2002), 357–363.

[FMV06] Uriel Feige, Elchanan Mossel, and Dan Vilenchik, Complete convergence of message passing algorithms for some satisfiability problems, IN RANDOM (2006), 339–350.

[FPV13] Vitaly Feldman, Will Perkins, and Santosh Vempala, On the complexity of random satisfiability problems with planted solutions, Arxiv Preprint (2013).

[HJKN06] Harri Haanpää, Matti Järvisalo, Petteri Kaski, and Ilkka Niemelä, Hard satisfiable clause sets for benchmarking equivalence reasoning techniques, Journal on Satisfiability, Boolean Modeling and Computation 2 (2006), no. 1, 27–46.

[Ing82] Yu. I. Ingster, The asymptotic efficiency of tests for a simple hypothesis against a composite alternative, Teor. Veroyatnost. i Primenen. 27 (1982), no. 3, 587–592. MR673934 (84m:62040)

[Ing98] Yuri I. Ingster, Minimax detection of a signal for $l^n$-balls, Math. Methods Statist. 7 (1998), no. 4, 401–428 (1999). MR1680087 (2000f:62012)

[ITV10] Yuri I. Ingster, Alexandre B. Tsybakov, and Nicolas Verzelen, Detection boundary in sparse regression, Electron. J. Stat. 4 (2010), 1476–1526.

[JMS05] Haixia Jia, Cristopher Moore, and Doug Strain, Generating hard satisfiable formulas by hiding solutions deceptively, IN AAAI (2005), 384–389.

[Kar72] Richard M. Karp, Reducibility among combinatorial problems, Complexity of computer computations (Proc. Sympos., IBM Thomas J. Watson Res. Center, Yorktown Heights, N.Y., 1972), Plenum, New York, 1972, pp. 85–103. MR0378476 (51 #14644)

[KV06] Michael Krivelevich and Dan Vilenchik, Solving random satisfiable 3cnf formulas in expected polynomial time, IN PROC. 17TH ACM-SIAM SYMP. ON DISCRETE ALGORITHMS (2006), 454–463.

[Lev73] Leonid Levin, Universal search problems, Problemy Peredachi Informatsii 9 (1973), no. 3, 115–116.

[MW13] Zongming Ma and Yihong Wu, Computational barriers in minimax submatrix detection, Arxiv Preprint (2013).

[Ser00] Rocco A. Servedio, Computational sample complexity and attribute-efficient learning, Journal of Computer and System Sciences 60
Shai Shalev-Shwartz, Ohad Shamir, and Eran Tomer, *Using more data to speed-up training time*, Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics April 21-23, 2012 La Palma, Canary Islands., JMLR W&CP, vol. 22, 2012, pp. 1019–1027.

Alexandre B. Tsybakov, *Introduction to nonparametric estimation*, Springer Series in Statistics, Springer, New York, 2009, Revised and extended from the 2004 French original, Translated by Vladimir Zaiats. MR2724359 (2011g:62006)