Optimal Testing for Planted Satisfiability Problems

Quentin Berthet[*,†]

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 of detection. Our analysis relies on the study of the number of satisfying assignments, for which we prove new results. We also address algorithmic issues, and give a computationally efficient test with optimal statistical performance. This result is compared to an average-case hypothesis on the hardness of refuting 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-SAT formula of m clauses on n variables. This is formulated as a hypothesis testing problem: Given a formula ϕ, 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 guaranteed to be satisfiable, by planting a solution.

There is a resurgence in statistics of hypothesis testing 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 weakly sparse signals in high dimension has inspired a wide literature of detection problems. Examples include [ITV10] in the context of sparse linear regression, [ACCD11, BI13, ACV13, MW13] for small cliques or communities in graphs and matrices, [ABBDL10] for general combinatorial structured signals, and [ACBL12, BR12, BR13] for sparse principal components of covariance matrices. 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.

Minimax theory gives a formal definition of the statistical complexity of a hypothesis testing problem, in terms of the sample size needed to identify with high probability the underlying distribution of given instances. It describes the

[*] The author thanks Philippe Rigollet, Emmanuel Abbé, Dan Vilenchick and Amin Coja-Oghlan for very helpful discussions.

[†] Partially supported by NSF grant CAREER-DMS-1053987 when the author was at Princeton University, and by AFOSR grant FA9550-14-1-0098.

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 ϕ 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, CO13, DSS14], maximum satisfiability [ANP03] geometry of the space of solutions [ANP03, ART06, ACO08, KMRT+06, MRT09], 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]. Methods from statistical physics such as belief and survey propagation have been applied to this problem and rigorously studied [BMZ02, MPZ02, MZ02, CO10]. More recently, the algorithmic complexity (in a specific computational model) of estimating the planted assignment has been studied in [FPV13].

Here, the use of tools from statistical analysis, such as the likelihood ratio and the total variation distance, highlights the importance of a specific statistic: the number of satisfying assignments. More specifically, we study its deviations from its expected value. Optimal rates of detection are obtained by proving new results concerning the concentration (or absence thereof) of this statistic. We address algorithmic issues by showing that the optimal rates of detection can be obtained by a newly introduced polynomial-time test. We also show the effect of choosing a different planting distribution on the detection problem, particularly on the optimal rates of detection.

The following subsection introduces notations for k-SAT formulas. 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 detection rates of different choices for the planting distributions is studied in Section 4.

Notations for k-SAT formulas

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

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

where for all $i \in \{1, \dots, m\}$, the clause C_i is the disjunction of k literals on k distinct variables, i.e., the value of a variable or its negation

$$C_i(x) = \ell_{i,1} \vee \ldots \vee \ell_{i,k}\,,\ \ell_{i,j} \in \{x_1, \bar{x}_1, \dots, x_n, \bar{x}_n\}\,, \text{and } \ell_{i,j} \notin \{\ell_{i,j'}, \bar{\ell}_{i,j'}\}.$$

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

$$\mathcal{S}(\phi) = \big\{x \in \{0,1\}^n : \phi(x) = \text{'}\mathsf{true}\text{'}\big\}\,,$$

and by $Z(\phi) = |\mathcal{S}(\phi)|$ the number of satisfying assignments for ϕ. We often write Z when it is not ambiguous. For a subset S of $\{1, \dots, m\}$, we define the sub-formula

$$\phi_S = \bigwedge_{i \in S} C_i\,.$$

The definition of satisfying assignments extends to single clauses and sub-formulas in general, with the notations $\mathcal{S}(C_i)$ and $\mathcal{S}(\phi_S)$ for the set of assignments satisfying respectively, the clause C_i or the formula ϕ_S. We denote by SAT the set of satisfiable formulas: those with satisfying assignments.

1. PROBLEM DESCRIPTION

We are interested in distinguishing two distributions on $\mathcal{F}^k_{m,n}$, the *uniform*, and *planted* distributions. The uniform distribution, denoted by $\mathbf{P}_{\text{unif}}$, is generated by independently selecting each clause uniformly from the $2^k\binom{n}{k}$ possible choices. The planted distribution, denoted by $\mathbf{P}_{\text{planted}}$, 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 $\mathbf{P}_{x^*}$). Each clause is given as k literals, in a uniformly random order. We represent this as a hypothesis testing problem, on the observation $\phi \in \mathcal{F}^k_{m,n}$

$$\begin{aligned} H_0 &: \phi \sim \mathbf{P}_{\text{unif}} \\ H_1 &: \phi \sim \mathbf{P}_{\text{planted}} = \frac{1}{2^n} \sum_{x \in \{0,1\}^n} \mathbf{P}_x\,. \end{aligned}$$

It is also possible to consider the detection problem with composite alternative hypothesis over the $\mathbf{P}_x$. Our formulation is equivalent to choosing a uniform prior over the planted assignments, and to consider the distribution $\mathbf{P}_{\text{planted}}$, mixture of the $\mathbf{P}_x$. We will mention two regimes: the *linear regime*, when $m = \Delta n$, for some $\Delta > 0$, usually the only one considered in the probability theory literature; and the *square-root regime*, when $m = C\sqrt{n}$, for some $C > 0$, particularly relevant to the study of our statistical problem. We will often consider m, n large enough, but will mainly focus on non-asymptotic results.

We define a test as a measurable function $\Psi : \mathcal{F}^k_{m,n} \to \{0,1\}$, whose goal is to determine the underlying distribution of the observation ϕ. We define the probability of error as the maximum of the probabilities of type I and type II error, formally

$$\mathbf{P}_{\text{unif}}(\Psi(\phi) = 1) \vee \mathbf{P}_{\text{planted}}(\Psi(\phi) = 0)\,.$$

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

We can make the simple observation that under the planted distribution, formulas are *guaranteed* to be satisfiable. This suggests to test satisfiability of the formula in order to solve the hypothesis testing problem. This test has a probability of error of type II equal to zero. Under the uniform distribution, the behavior of $\mathbf{P}_{\text{unif}}(\phi \in \mathsf{SAT})$ has been extensively studied, and a phase transition has been shown to exist in the linear regime of $m = \Delta n$, from satisfiability to unsatisfiability, around some Δ_k close to $2^k \log(2)$. We refer to [COP13, CO13] and references therein for more information, as well as [DSS14] for a proof of the sharpness of the phase transition, for k large enough. In this setting, when $\Delta > \Delta_k$, the satisfiability test $\Psi_{\mathsf{SAT}} = \mathbf{1}\{\cdot \in \mathsf{SAT}\}$ has a probability of error going to 0, and when $\Delta < \Delta_k$, the error will converge to 1 (entirely because of the probability of a type I error).

When thinking of the formula ϕ as a sequence of m i.i.d. clauses, m can be interpreted as the sample size, and the problem becomes easier when Δ increases. When Δ is too small, the probability of error of the test Ψ_{SAT} converges to 1. We see in the following section that this simple rate can be significantly improved.

2. OPTIMAL TESTING

In this section, we derive the optimal rate of detection for this problem, i.e., how large m should be for a test to be able to distinguish with high probability the two hypotheses. We prove that the *likelihood-ratio test* is successful in the square-root regime, and show that it is information-theoretic optimal.

2.1 Likelihood-ratio test

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 $m \geq C\sqrt{n}$ for a specific constant C, the probability of error of the likelihood-ratio test is smaller than $\delta \in (0, 1)$.

THEOREM 2.1. *For all $k \geq 2$, positive m, n, denote Ψ_{LR} the likelihood-ratio test defined by*

$$\Psi_{\mathsf{LR}}(\phi) = \mathbf{1}\{Z(\phi) > \mathbf{E}_{\text{unif}}[Z]\}\,. \tag{1}$$

For any $\delta \in (0, 1)$, there exists $\bar{C}_{k,\delta} > 0$ such that for $m \geq \bar{C}_{k,\delta}\sqrt{n}$, for m, n large enough, it holds

$$\mathbf{P}_{\text{unif}}(\Psi_{\mathsf{LR}}(\phi) = 1) \vee \mathbf{P}_{\text{planted}}(\Psi_{\mathsf{LR}}(\phi) = 0) \leq \delta\,.$$

PROOF. We first prove that the likelihood-ratio test has indeed form (1). For discrete distributions, the likelihood ratio is simply equal to the ratio of the two distributions. For all $\phi \in \mathcal{F}^k_{m,n}$, it holds

$$\frac{\mathbf{P}_{\text{planted}}(\phi)}{\mathbf{P}_{\text{unif}}(\phi)} = \frac{1}{2^n} \sum_{x \in \{0,1\}^n} \frac{\mathbf{P}_x(\phi)}{\mathbf{P}_{\text{unif}}(\phi)} \,.$$

To compute the probabilities in the above ratios, we can interpret the drawing of ϕ by placing m balls in $N = 2^k\binom{n}{k}$ bins independently - if it has distribution $\mathbf{P}_{\text{unif}}$ - or otherwise in the $N_k = (2^k - 1)\binom{n}{k}$ bins corresponding to clauses that are satisfied by x. Therefore, it holds for all ϕ

$$\frac{\mathbf{P}_x(\phi)}{\mathbf{P}_{\text{unif}}(\phi)} = \begin{cases} 0 & \text{if } x \notin \mathcal{S}(\phi) \\ \left(\frac{N}{N_k}\right)^m & \text{otherwise} \end{cases}$$

It can then be expressed in terms of $\mathbf{1}\{x \in \mathcal{S}(\phi)\}$, and $N/N_k = 1/(1-2^{-k})$

$$\begin{aligned} \frac{\mathbf{P}_{\text{planted}}}{\mathbf{P}_{\text{unif}}}(\phi) &= \frac{1}{2^n} \sum_{x \in \{0,1\}^n} \left(\frac{N}{N_k}\right)^m \mathbf{1}\{x \in \mathcal{S}(\phi)\} \\ &= \frac{1}{\mathbf{E}_{\text{unif}}[Z(\phi)]} \sum_{x \in \{0,1\}^n} \mathbf{1}\{x \in \mathcal{S}(\phi)\} = \frac{Z(\phi)}{\mathbf{E}_{\text{unif}}[Z(\phi)]} \,, \end{aligned}$$

by the known closed form of $\mathbf{E}_{\text{unif}}[Z(\phi)] = 2^n(1-2^{-k})^m$, which can be directly derived by linearity. The likelihood-ratio test is therefore indeed $\Psi_{\mathsf{LR}}(\phi) = \mathbf{1}\{Z(\phi) > \mathbf{E}_{\text{unif}}[Z(\phi)]\}$. It is now sufficient to prove $\mathbf{P}_{\text{unif}}(\Psi(\phi) = 1) + \mathbf{P}_{\text{planted}}(\Psi(\phi) = 0) \leq \delta$, as the maximum of two nonnegative numbers is smaller than their sum. By definition of the likelihood-ratio test,

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

Furthermore, by definition of the total variation distance

$$\begin{aligned} d_{TV}(\mathbf{P}_{\text{unif}}, \mathbf{P}_{\text{planted}}) &= \sum_{\substack{\phi \in \mathcal{F}^k_{m,n} \\ \mathbf{P}_{\text{unif}}(\phi) > \mathbf{P}_{\text{planted}}(\phi)}} \{\mathbf{P}_{\text{unif}} - \mathbf{P}_{\text{planted}}\}(\phi) \\ &= \sum_{\substack{\phi \in \mathcal{F}^k_{m,n} \\ Z(\phi)/\mathbf{E}[Z] < 1}} \left(1 - \frac{Z(\phi)}{\mathbf{E}[Z]}\right) \mathbf{P}_{\text{unif}}(\phi) \\ &= \mathbf{E}_{\text{unif}}\left[\left(1 - \frac{Z(\phi)}{\mathbf{E}[Z]}\right)_+\right] \,. \end{aligned}$$

The total variation distance between distributions of i.i.d. elements being nondecreasing in the sample size, we obtain by Lemma 2.2 that in the square-root regime, for C large enough and $m \geq C\sqrt{n}$,

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

This bound yields the desired result for some large enough constant $C_{k,\delta} > 0$. □

The proof of this theorem indicates that it is possible to distinguish the two distributions whenever Z is not concentrated around its expectation under the uniform distribution. 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 $Z/\mathbf{E}[Z]$ is much smaller than 1, with high probability.

LEMMA 2.2. *For all $k \geq 2$, C_0 an absolute constant, $m = C\sqrt{n}$, and C, n large enough, it holds with probability $1 - C_0/C^2$,for some constant $\gamma_k > 0$ that*

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

A stronger result, concerning the linear regime, can be derived similarly in order to answer a question regarding the behavior of Z with respect to its expectation. It is known [AM10] that for Δ small enough and $n \to +\infty$, $n^{-1}\log(Z)$ and $n^{-1}\mathbf{E}[\log(Z)]$ have the same limit, called the *quenched* average. In the following lemma, we prove that this limit is actually different from the constant $n^{-1}\log(\mathbf{E}[Z])$, called the *annealed* average, for all $\Delta > 0$.

LEMMA 2.3. *For all $k \geq 2$, $\Delta > 0$, and $m = \Delta n$ large enough, if $\phi \sim \mathbf{P}_{unif}$, 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 result is tangential to the problem at hand but of interest in and of itself. We show here that the quenched and annealed averages are different for all Δ and k, with a gap greater than $c_{k,\Delta}$, for which we give no explicit formula. This phenomenon is hinted at in [ACO08, CO09], and proven to hold for Δ large enough in [COP13], with an explicit lower bound for the gap. We provide a proof for Lemma 2.2 and 2.3 in Appendix A.

2.2 Information-theoretic lower bound

The proof of Theorem 2.1 also hints at a lower bounds for the statistical problem. The total variation 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.

The number of satisfying assignments is actually equal to its expectation whenever no variable appears in two different clauses. Indeed, when this is the case, the set of satisfying assignments can be described thus. There are m clauses on m distinct groups of k distinct variables. Each clause allows a specific group of k variables to take $2^k - 1$ values, and the $n - km$ remaining variables are free. 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, $Z = (2^k - 1)^m 2^{n-km} = 2^n(1 - 2^{-k})^m = \mathbf{E}[Z]$. This observation yields the following lower bound.

THEOREM 2.4. *For $\nu \in (0, 1/2)$, $m \leq 2\sqrt{\nu n}/k$, and m, n large enough, it holds that*

$$\inf_{\Psi} \left\{ \mathbf{P}_{\text{unif}}(\Psi(\phi) = 1) \vee \mathbf{P}_{\text{planted}}(\Psi(\phi) = 0) \right\} \geq \frac{1}{2} - \nu\,.$$

PROOF. We use the total variation bound, for any test Ψ

$$\begin{aligned} \mathbf{P}_{\text{unif}}(\Psi(\phi) = 1) \vee \mathbf{P}_{\text{planted}}(\Psi(\phi) = 0) &\geq \frac{1}{2}\big(\mathbf{P}_{\text{unif}}(\Psi(\phi) = 1) + \mathbf{P}_{\text{planted}}(\Psi(\phi) = 0)\big) \\ &\geq \frac{1 - d_{TV}(\mathbf{P}_{\text{unif}}, \mathbf{P}_{\text{planted}})}{2}\,. \end{aligned}$$

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

$$\begin{aligned} d_{TV}(\mathbf{P}_{\text{unif}}, \mathbf{P}_{\text{planted}}) &= \frac{1}{2} \sum_{\phi \in \mathcal{F}^k_{m,n}} |\mathbf{P}_{\text{unif}} - \mathbf{P}_{\text{planted}}|(\phi) \\ &= \frac{1}{2} \sum_{\phi \in F} |\mathbf{P}_{\text{unif}} - \mathbf{P}_{\text{planted}}|(\phi) + \frac{1}{2} \sum_{\phi \in F^c} |\mathbf{P}_{\text{unif}} - \mathbf{P}_{\text{planted}}|(\phi) \\ &= \frac{1}{2} \sum_{\phi \in F} \Big| \frac{Z(\phi)}{\mathbf{E}[Z]} - 1 \Big| \mathbf{P}_{\text{unif}}(\phi) + \frac{1}{2} \sum_{\phi \in F^c} |\mathbf{P}_{\text{unif}} - \mathbf{P}_{\text{planted}}|(\phi) \end{aligned}$$

As noticed above, for all $\phi \in F$, $Z(\phi) = \mathbf{E}[Z]$; the likelihood ratio is equal to 1. The first term of this equation is therefore equal to 0. This also implies that $\mathbf{P}_{\text{unif}}(\phi) = \mathbf{P}_{\text{planted}}(\phi)$ for all $\phi \in F$, and $\mathbf{P}_{\text{unif}}(F) = \mathbf{P}_{\text{planted}}(F)$. The second term is thus upper bounded by $\mathbf{P}_{\text{unif}}(F^c) = \mathbf{P}_{\text{planted}}(F^c)$. It is sufficient to prove that $\mathbf{P}_{\text{unif}}(F^c) \leq 2\nu$, a variant of the "birthday problem": We place a group of k balls in n distinct bins uniformly at random, m times independently. The probability that none of these m groups intersect is equal to $\mathbf{P}_{\text{unif}}(F)$. 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 $k^2 i/n$ (the expected number of such collisions). As $k^2(m-1)/n < 1/2$ (for fixed ν and n large enough) the following holds

$$\mathbf{P}_{\text{unif}}(F) \geq \prod_{i=1}^{m-1} \Big(1 - \frac{k^2 i}{n}\Big) > \prod_{i=1}^{m-1} e^{-2k^2 i/n} = e^{-k^2(m-1)(m-2)/n} > 1 - k^2 m^2/n\,.$$

This gives the desired result. □

From the last two theorems, we can conclude that the *optimal rate of detection* is $m^* = \sqrt{n}$. When $m = C\sqrt{n}$, detection is possible with probability of error smaller than δ, for C greater than some constant $\bar{C}_{k,\delta}$, by using the likelihood-ratio test. It is impossible to distinguish the two hypotheses with error probability smaller than $1/2 - \nu$ for $C < \underline{C}_{k,\nu} := 2\sqrt{\nu}/k$. No effort has been made to optimize (or even quantify) the constant $\bar{C}_{k,\delta}$, as a function of k and δ.

3. POLYNOMIAL-TIME TESTING

For $k \geq 2$, computing the outcome of the likelihood-ratio test involves solving a #P-complete problem [Val79], and for $k \geq 3$, even computing the outcome of the satisfiability test Ψ_{SAT} (which is already suboptimal) is equivalent to solving a NP-hard problem. 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 NP-complete [Coo71, Lev73], at the root of the web of problems known to be in the same class [Kar72]. None of the tests described above can be computed in a computationally efficient manner. It is therefore legitimate to examine the performance of tests that can be computed in polynomial time.

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 (for a large enough Δ) is presented in [CoKV07], for the case $k = 3$ (their results extend to any fixed k). A similar problem is studied as well in [FPV13]. This method can be used as a tool for detection: in the unsatisfiable regime (when Δ is large enough), the existence of a satisfying assignment is a sufficient reason to reject the null. The main issue of this approach is that the regime of detection is not optimal: m needs to be of order n (linear regime), when only $\sqrt{n}$ (square-root regime) is required for the likelihood-ratio test.

3.1 Variable coupling test

The proof that the likelihood-ratio test has a low probability of error in the optimal regime is based on the fact that there is a large number of variables that appear more than once, and on the fact that under the null distribution, a couple of literals based on the same variable have equal probability to have the same sign or opposite signs. We can use this fact to design a test that runs in polynomial time and achieves the optimal rate of detection.

We recall that in each clause, the literals are given in a uniformly random order. Let T be the number of variables (among the n possible) that appear more than once as the first literal of a clause of ϕ (according to the random ordering in the data) and P (resp. D) the number of those for which the first two occurrences (according to the natural order of the clauses) of the same variable have the same sign (resp. different signs), so that $P + D = T$. The following holds

THEOREM 3.1. *For all $k \geq 2$, $m, n > 0$ and $\delta \in (0, 1)$, denote Ψ_{COU} the test defined by*

$$\Psi_{\mathsf{COU}}(\phi) = \mathbf{1}\{P/T > 1/2 + 1/[2(2^k - 1)]^2\} \,,$$

and

$$\tilde{C}_{k,\delta} := [2(2^k - 1)]^2 \sqrt{2 \log(2/\delta)} \vee \sqrt{1024/\delta} \,.$$

For $m \geq \tilde{C}_{k,\delta} \sqrt{n}$, it holds

$$\mathbf{P}_{\text{unif}}(\Psi_{\mathsf{COU}}(\phi) = 1) \vee \mathbf{P}_{\text{planted}}(\Psi_{\mathsf{COU}}(\phi) = 0) \leq \delta \,.$$

PROOF. For each variable that appears at least twice as the first literal of a clause, consider the probability that the two first occurrences (according to the natural order of the clauses) of a variable as the first literal of a clause (according to the random ordering in the data) have the same value. It is equal to $1/2$ under the uniform distribution, and conditionally on the value of T, $P \sim \mathcal{B}(T, 1/2)$. Under the planted distribution, each literal has independently probability $(1 + 1/(2^k - 1))/2$ to have the same value as the corresponding variable in x_i^*, and probability $(1 - 1/(2^k - 1))/2$ to have a different value. Overall, the probability that these two literals have the same sign under the planted distribution is

$$\frac{1}{4}\big(1 + \frac{1}{2^k - 1}\big)^2 + \frac{1}{4}\big(1 - \frac{1}{2^k - 1}\big)^2 = \frac{1}{2} + \frac{1}{2(2^k - 1)^2} \,.$$

Therefore, conditionally on the value of T, P has distribution $\mathcal{B}(T, 1/2 + 1/[2(2^k - 1)^2])$. By Hoeffding's inequality, the following holds for all $\varepsilon > 0$

$$\mathbf{P}_{\text{unif}}\big(P/T > 1/2 + \varepsilon \,|\, T\big) \leq \exp(-2\varepsilon^2 T)$$
$$\mathbf{P}_{\text{planted}}\big(P/T < 1/2 + 1/[2(2^k - 1)^2] - \varepsilon \,|\, T\big) \leq \exp(-2\varepsilon^2 T)$$

By Lemma A.1, and by definition of $\tilde{C}_{k,\delta}$, $T \geq \tilde{C}^2_{k,\delta}/4$ with probability at least $1-\delta/2$. Let $\varepsilon = 1/[2(2^k-1)]^2$, and condition on the event $T \geq \tilde{C}^2_{k,\delta}/4$. The previous yields, for $C_{k,\delta} \geq \sqrt{2\log(2/\delta)}/\varepsilon$

$$\mathbf{P}_{\text{unif}}\big(P/T > 1/2 + 1/[2(2^k-1)]^2 \,|\, T\big) \leq \delta/2$$
$$\mathbf{P}_{\text{planted}}\big(P/T < 1/2 + 1/[2(2^k-1)]^2 \,|\, T\big) \leq \delta/2\,.$$

Which gives the desired result by a simple union bound. □

3.2 Hardness hypothesis on random instances

The result of Theorem 3.1 can be contrasted with a hypothesis by Feige, formulated in [Fei02], to prove hardness of approximation results in the worst case. We recall the proposed assumption on the hardness of determining the satisfiability of 3-SAT formulas on average:

"Even when Δ 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, in a statistical language, it is conjectured in this hypothesis that for all $\Delta > 0$, in the linear regime, there is no test Ψ 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. At first sight, this is in apparent contradiction with theorem 3.1. Interestingly, this result shows that up to the optimal square-root regime it is possible to design a test with small type I and type II errors simultaneously, even though it is conjectured and widely believed that it is impossible to distinguish those distributions with a completely one-sided error.

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 the statistical performance of computationally efficient procedures. Such phenomena 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, WBS14], based on a hardness hypothesis for the planted clique problem. The above discussion shows the difficulty 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.

4. ALTERNATIVE CHOICES FOR PLANTING DISTRIBUTIONS

The tests described in Theorems 2.1 and 3.1 exploit a fundamental difference between the two considered distributions. Planting a satisfying assignment $x^* \in \{0,1\}^n$ breaks the symmetry of the uniform distribution. The likelihood ratio

$Z/\mathbf{E}[Z]$ is affected by the imbalances in interactions between variables. Similarly, the variable coupling test is based on the bias in the signs of chosen literals, under the planted distribution.

This asymmetry is a characteristic of our choice of the planting distribution. In this section, we observe that the rates of detection are different for other natural choices of distribution on SAT, the set of satisfiable formulas. Such an example is $\mathbf{P}_{\mathsf{SAT}}$, the uniform distribution on SAT. In this new statistical problem, the alternative hypothesis becomes $\tilde{H}_1 : \phi \sim \mathbf{P}_{\mathsf{SAT}}$.

It is a fundamentally different statistical problem: its optimal rate of detection is the linear regime $m^* = n$, achieved by the satisfiability test Ψ_{SAT}. Indeed, as shown in a simple remark in Section 1, this test is successful in the satisfiable part of the linear regime. Furthermore, as $\mathbf{P}_{\mathsf{SAT}}$ is the uniform distribution on SAT, or $\mathbf{P}_{\text{unif}}(\,\cdot\,|\phi \in \mathsf{SAT})$, the total variation distance $d_{TV}(\mathbf{P}_{\text{unif}}, \mathbf{P}_{\mathsf{SAT}})$ is equal to $\mathbf{P}_{\text{unif}}(\phi \notin \mathsf{SAT})$. As explained before, this probability vanishes to 0 for Δ small enough, which yields the matching lower bound. From a statistical point of view, this modified hypothesis testing problem is a significantly harder task than the detection of planted satisfiability.

Among all distributions on satisfiable formulas, the closest in total variation 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, KMZ12] 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.

APPENDIX A: PROOFS OF TECHNICAL RESULTS

Lemma 2.2 and 2.3 are a consequence of the following result on the number of variables that appear at least twice in the formula. For simplicity of the proof, we only consider the first literal of each clause, which is sufficient to our objective.

LEMMA A.1. *Let ϕ be a random formula of $\mathcal{F}^k_{m,n}$ with distribution $\mathbf{P}_{unif}$. Let T be the number of variables (among the possible n) that appear more than once as the first literal of a clause of ϕ.*

- *Let $\Delta > 0$, and $m = \Delta n$. There exists positive constants ε_Δ and r_Δ such that*
$$\mathbf{P}(T < \varepsilon_\Delta n) \leq \frac{r_\Delta}{n}\,.$$
- *Let $C > 0$, and $m = C\sqrt{n}$. It holds that*
$$\mathbf{P}(T < C^2/4) \leq \frac{576}{C^2}\,.$$

PROOF. We prove this deviation bounds in the two regimes.

Linear regime

We first place ourselves in the linear regime $m = \Delta n$. The first literals of the clauses of the random formula can be interpreted as being drawn by independently placing m balls uniformly in n bins, and T_i is 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

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

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

$$T = T_1 + \ldots + T_n\,.$$

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

$$\mathbf{Var}[T] = n\mathbf{Var}[T_1] + n(n-1)\big[\mathbf{E}[T_1T_2] - \mathbf{E}[T_1]\mathbf{E}[T_2]\big]\,.$$

We control the last term

$$\begin{aligned}
\mathbf{E}[T_1T_2] &= \mathbf{P}[T_1=1, T_2=1] = \mathbf{P}[T_1=1|T_2=1]\mathbf{P}[T_2=1]\\
&= \mathbf{P}[T_1=1|T_2=1]\mathbf{E}[T_2]\\
&= \Big[1-\Big[\Big(1-\frac{1}{n}\Big)^{m-2} + (m-2)\Big(1-\frac{1}{n}\Big)^{m-3}\frac{1}{n}\Big]\Big]\mathbf{E}[T_2]
\end{aligned}$$

Therefore, we obtain the bound

$$\mathbf{E}[T_1T_2] - \mathbf{E}[T_1]\mathbf{E}[T_2] \le \Big[1-\Big(1-\frac{1}{n}\Big)^2 + \Delta\Big(1-\Big(1-\frac{1}{n}\Big)^2\Big)\Big]\mathbf{E}[T_2] \le \frac{3+3\Delta}{n}\,.$$

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

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

Square-root regime

This proof is a simple modification of the proof of the linear regime with the same notations, for $m = C\sqrt{n}$. We derive the expectation and variance of T

$$\begin{aligned}
\mathbf{E}[T_i] &= 1-\Big[\Big(1-\frac{1}{n}\Big)^m + m\Big(1-\frac{1}{n}\Big)^{m-1}\frac{1}{n}\Big]\\
&= 1-\Big[\Big(1-\frac{1}{n}\Big)^{C\sqrt{n}} + \frac{C}{\sqrt{n}}\Big(1-\frac{1}{n}\Big)^{C\sqrt{n}-1}\Big]\\
&= 1-\Big[1-\frac{C}{\sqrt{n}}+\frac{C^2}{2n}+o\Big(\frac{1}{n}\Big)+\frac{C}{\sqrt{n}}-\frac{C^2}{n}+o\Big(\frac{1}{n}\Big)\Big] = \frac{C^2}{2n}+o\Big(\frac{1}{n}\Big)\,.
\end{aligned}$$

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

$$\mathbf{Var}[T] = n\mathbf{Var}[T_1] + n(n-1)\big[\mathbf{E}[T_1T_2] - \mathbf{E}[T_1]\mathbf{E}[T_2]\big]\,.$$

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

$$\mathbf{E}[T_1T_2] - \mathbf{E}[T_1]\mathbf{E}[T_2] \le \Big[1-\Big(1-\frac{1}{n}\Big)^2 + \frac{C}{\sqrt{n}}\Big(1-\Big(1-\frac{1}{n}\Big)^2\Big)\Big]\mathbf{E}[T_2] \le \frac{3}{n}\times C^2/n\,.$$

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

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

□

Proof of Lemma 2.2 and 2.3. For all $x \in \{0,1\}^n$, $x \in \mathcal{S}(\phi)$ if and only if x satisfies all the clauses of ϕ. We can therefore write

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

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

In the proof of Theorem 2.4, 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 whether the first two occurrences of every variable (when there are two or more) are the same literal or not. Formally, we denote $G = (G_1, \dots, G_n)$ the partition of $\{1, \dots, m\}$ in n sets (allowing some of them to be empty), where

$$G_i = \big\{ j \in \{1, \dots, m\} : C_j(x) \in \{x_i \wedge \dots, \bar{x}_i \wedge \dots\} \big\}\,,$$

and $\sigma = (\sigma_1, \dots, \sigma_n)$, 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

$$\mathbf{E}[Z \,|\, (G,\sigma)] = \sum_{x \in \{0,1\}^n} \mathbf{E}\Big[\mathbf{1}\{x \in \mathcal{S}(\phi)\} \,|\, (G,\sigma)\Big]\,.$$

We now observe that this conditional expectation is constant, for all $x \in \{0,1\}^n$. Indeed, let e_0 be the assignment of all zeroes, and t_x be the literal-flipping transformation such that $t_x(e_0) = x$, and T_x the corresponding literal-flipping transformation on formulas. For all x, it holds

$$\phi(x) = \phi(t_x(e_0)) = (T_x\phi)(e_0)\,.$$

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

$$\begin{aligned}
\mathbf{E}[Z \,|\, (G,\sigma)] &= \sum_{x \in \{0,1\}^n} \mathbf{E}\Big[\mathbf{1}\{x \in \mathcal{S}(\phi)\} \,|\, (G,\sigma)\Big] \\
&= \sum_{x \in \{0,1\}^n} \mathbf{E}\Big[\mathbf{1}\{e_0 \in \mathcal{S}(T_x\phi)\} \,|\, (G,\sigma)\Big] \\
&= 2^n \mathbf{E}\Big[\mathbf{1}\{e_0 \in \mathcal{S}(\phi)\} \,|\, (G,\sigma)\Big]\,.
\end{aligned}$$

The assignment e_0 will satisfy the formula ϕ if and only if it satisfies all the sub-formulas $\phi_{G_1}, \dots, \phi_{G_n}$ (the empty formula is always satisfied). Given (G, σ), the events $\{e_0 \in \mathcal{S}(\phi_{G_i})\}$ are independent: the sub-formulas are satisfied by e_0 if and only if every clause contains at least one negated literal, which occurs independently, conditioned on (G, σ). We can therefore compute the conditional expectation

$$\begin{aligned} \mathbf{E}\Big[\mathbf{1}\{e_0 \in \mathcal{S}(\phi)\} \,|\, (G,\sigma)\Big] &= \mathbf{E}\Big[\prod_{i=1}^n \mathbf{1}\{e_0 \in \mathcal{S}(\phi_{G_i})\} \,|\, (G,\sigma)\Big] \\ &= \prod_{i=1}^n \mathbf{E}\Big[\mathbf{1}\{e_0 \in \mathcal{S}(\phi_{G_i})\} \,|\, (G,\sigma)\Big] \\ &= \prod_{i=1}^n \mathbf{E}\Big[\mathbf{1}\{e_0 \in \mathcal{S}(\phi_{G_i})\} \,|\, (G_i,\sigma_i)\Big] \end{aligned}$$

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

$$\mathbf{E}\Big[\mathbf{1}\{e_0 \in \mathcal{S}(\phi_{G_i})\} \,|\, (G_i, \sigma_i = 0)\Big] = \Big(1 - \frac{1}{2^k}\Big)^{g_i} .$$

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

$$\begin{aligned} \mathbf{E}\Big[\mathbf{1}\{e_0 \in \mathcal{S}(\phi_{G_i})\} \,|\, (G_i, \sigma_i = 1)\Big] &= \frac{1}{2}\Big[1 + \Big(1 - \frac{1}{2^{k-1}}\Big)^2\Big]\Big(1 - \frac{1}{2^k}\Big)^{g_i - 2} \\ \mathbf{E}\Big[\mathbf{1}\{e_0 \in \mathcal{S}(\phi_{G_i})\} \,|\, (G_i, \sigma_i = -1)\Big] &= \Big(1 - \frac{1}{2^{k-1}}\Big)\Big(1 - \frac{1}{2^k}\Big)^{g_i - 2} . \end{aligned}$$

Overall, this yields

$$\mathbf{E}\Big[\mathbf{1}\{e_0 \in \mathcal{S}(\phi_{G_i})\} \,|\, (G_i, \sigma_i)\Big] = \Big[1 + \frac{\sigma_i}{2^{2k}(1 - 2^{-k})^2}\Big]\Big(1 - \frac{1}{2^k}\Big)^{g_i} .$$

Recall that we denote P (resp. D) the number of groups for which $\sigma_i = 1$ (resp. -1). It holds that

$$\mathbf{E}[Z \,|\, (G,\sigma)] = 2^n \Big(1 - \frac{1}{2^k}\Big)^m \Big[1 + \frac{1}{2^{2k}(1 - 2^{-k})^2}\Big]^P \Big[1 - \frac{1}{2^{2k}(1 - 2^{-k})^2}\Big]^D .$$

It is possible to design a set of (G, σ), 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 $T = P + D$, with P and D close to $(P+D)/2$, this expectation is significantly smaller than $\mathbf{E}[Z]$. 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

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

for all $|\alpha| < \alpha_k$, for some $\gamma_k > 0$.

For every variable, we denote $T_i = |\sigma_i| \in \{0,1\}$, and $T = T_1 + \ldots + T_n$. We now prove independently the two lemmas.

Linear regime, Lemma 2.3

We control P and D in the regime $m = \Delta n$. By lemma A.1, it holds that

$$\mathbf{P}[T < \varepsilon_\Delta n] \leq \frac{r_\Delta}{n}\,.$$

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, $\mathbf{P}(B) = 1 - o(1)$. For (G,σ) in the event B, it holds

$$\begin{aligned}
\mathbf{E}[Z\,|\,(G,\sigma)] &= 2^n\Big(1-\frac{1}{2^k}\Big)^m \Big[1 + \frac{1}{2^{2k}(1-2^{-k})^2}\Big]^P \Big[1 - \frac{1}{2^{2k}(1-2^{-k})^2}\Big]^D \\
&< 2^n\Big(1-\frac{1}{2^k}\Big)^m (e^{-\gamma_k})^{T/2} < e^{-\gamma_k \varepsilon_\Delta n/2}\mathbf{E}[Z] := e^{-2c_{k,\Delta} n}\,\mathbf{E}[Z]\,.
\end{aligned}$$

Therefore $\mathbf{E}[Z\,|\,B] < e^{-2c_{k,\Delta}n}\,\mathbf{E}[Z]$. We can now conclude by conditioning on B and using Markov's inequality

$$\begin{aligned}
\mathbf{P}(Z > e^{-c_{k,\Delta}n}\,\mathbf{E}[Z]) &= \mathbf{P}(Z > e^{-c_{k,\Delta}n}\,\mathbf{E}[Z]\,|\,B)\mathbf{P}(B) + \\
&\quad\; \mathbf{P}(Z > e^{-c_{k,\Delta}n}\,\mathbf{E}[Z]\,|\,B^c)\mathbf{P}(B^c) \\
&\leq \mathbf{P}(Z > e^{-c_{k,\Delta}n}\,\mathbf{E}[Z]\,|\,B) + \mathbf{P}(B^c) \\
&\leq \frac{\mathbf{E}[Z\,|\,B]}{e^{-c_{k,\Delta}n}\,\mathbf{E}[Z]} + \mathbf{P}(B^c) \\
&\leq e^{-c_{k,\Delta}n} + \mathbf{P}(B^c)\,.
\end{aligned}$$

Which yields the desired result.

Square-root regime, Lemma 2.2

As in the linear regime, we control P and D when $m = C\sqrt{n}$. Lemma A.1 yields

$$\mathbf{P}[T \geq C^2/4] \leq \frac{576}{C^2}\,.$$

Again, 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, $\mathbf{P}(B) = 1 - O(1/C^2)$. For (G,σ) in the event B, it holds

$$\begin{aligned}
\mathbf{E}[Z\,|\,(G,\sigma)] &= 2^n\Big(1-\frac{1}{2^k}\Big)^m \Big[1 + \frac{1}{2^{2k}(1-2^{-k})^2}\Big]^P \Big[1 - \frac{1}{2^{2k}(1-2^{-k})^2}\Big]^D \\
&< 2^n\Big(1-\frac{1}{2^k}\Big)^m (e^{-\gamma_k})^{T/2} < e^{-\gamma_k C^2/8}\mathbf{E}[Z]\,.
\end{aligned}$$

Therefore $\mathbf{E}[Z \mid B] < e^{-\gamma_k C^2/8}\,\mathbf{E}[Z]$. We can now conclude by conditioning on B and using Markov's inequality

$$\begin{aligned}\mathbf{P}(Z > e^{-\gamma_k C^2/16}\,\mathbf{E}[Z]) &= \mathbf{P}(Z > e^{-\gamma_k C^2/16}\,\mathbf{E}[Z] \mid B)\mathbf{P}(B) + \\ &\quad \mathbf{P}(Z > e^{-c_{k,\Delta} n}\,\mathbf{E}[Z] \mid B^c)\mathbf{P}(B^c) \\ &\le \mathbf{P}(Z > e^{-\gamma_k C^2/16}\,\mathbf{E}[Z] \mid B) + \mathbf{P}(B^c) \\ &\le \frac{\mathbf{E}[Z \mid B]}{e^{-\gamma_k C^2/16}\,\mathbf{E}[Z]} + \mathbf{P}(B^c) \\ &\le e^{-\gamma_k C^2/8} + \mathbf{P}(B^c)\,.\end{aligned}$$

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

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.

[ACO08] Dimitris Achlioptas and Amin Coja-Oghlan, *Algorithmic barriers from phase transitions*, 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*, 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).

[ANP03] Dimitris Achlioptas, Assaf Naor, and Yuval Peres, *On the fraction of satisfiable clauses in typical formulas*, EXTENDED ABSTRACT IN FOCS'03 (2003), 362–370.

[AP04] 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.

[ART06] 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.

[BHL$^+$01] 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).

[BI13] Cristina Butucea and Yuri I. Ingster, *Detection of a sparse submatrix of a high-dimensional noisy matrix*, Bernoulli (to appear) (2013).

[BMZ02] A. Braunstein, M. Mézard, and R. Zecchina, *Survey propagation: an algorithm for satisfiability*.

[BR12] Quentin Berthet and Philippe Rigollet, *Optimal detection of sparse principal components in high dimension*, Ann. Statist. **41** (2012), no. 4, 1780–1815.

[BR13] ______, *Complexity theoretic lower bounds for sparse principal component detection*, J. Mach. Learn. Res. (COLT) **30** (2013), 1046–1066.

[Che13] Yudong Chen, *Incoherence-optimal matrix completion*.

[CJ13] Venkat Chandrasekaran and Michael I. Jordan, *Computational and statistical tradeoffs via convex relaxation*, Proceedings of the National Academy of Sciences (2013).

[CO09] Amin Coja-Oghlan, *Random constraint satisfaction problems*, Electronic Proceedings in Theoretical Computer Science **9** (2009), 32–37.

[CO10] ______, *On belief propagation guided decimation for random k-sat*.

[CO13] ______, *The asymptotic k-sat threshold*.

[CoKV07] 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).

[Coo71] S. A. Cook, *The complexity of theorem proving procedures*, Proceedings of the Third Annual ACM Symposium (New York), ACM, 1971, pp. 151–158.

[COP13] 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.

[DGR98] Scott E. Decatur, Oded Goldreich, and Dana Ron, *Computational sample complexity*, SIAM JOURNAL ON COMPUTING **29** (1998).

[DJ04] David Donoho and Jiashun Jin, *Higher criticism for detecting sparse heterogeneous mixtures*, Ann. Statist. **32** (2004), no. 3, 962–994. MR2065195 (2005e:62066)

[DLS13] Amit Daniely, Nati Linial, and Shai Shalev Shwartz, *More data speeds up training time in learning halfspaces over sparse vectors*, Arxiv Preprint (2013).

[DSS14] Jian Ding, Allan Sly, and Nike Sun, *Proof of the satisfiability conjecture for large k.*

[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 Fourty-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)

[KMRT$^+$06] Florent Krzakala, Andrea Montanari, Federico Ricci-Tersenghi, Guilhem Semerjian, and Lenka Zdeborova, *Gibbs states and the set of solutions of random constraint satisfaction problems.*

[KMZ12] Florent Krzakala, Marc Mézard, and Lenka Zdeborová, *Reweighted belief propagation and quiet planting for random k-sat.*

[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.

[MPZ02] M. Mézard, G. Parisi, and R. Zecchina, *Analytic and algorithmic solution of random satisfiability problems.*

[MRT09] Andrea Montanari, Ricardo Restrepo, and Prasad Tetali, *Reconstruction and clustering in random constraint satisfaction problems.*

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

[MZ02] Marc Mezard and Riccardo Zecchina, *The random k-satisfiability problem: from an analytic solution to an efficient algorithm.*

[Ser00] Rocco A. Servedio, *Computational sample complexity and attribute-efficient learning*, Journal of Computer and System Sciences **60** (2000), no. 1, 161–178.

[SSST12] 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.

[Val79] Leslie G. Valiant, *The complexity of enumeration and reliability problems*, SIAM J. Comput. **8** (1979), no. 3, 410–421.

[WBS14] Tengyao Wang, Quentin Berthet, and Richard J. Samworth, *Statistical and computational trade-offs in estimation of sparse principal components*, Arxiv Preprint (2014).

Department of Computing
and Mathematical Sciences
California Institute of Technology
Pasadena, CA 91125, USA
(qberthet@caltech.edu)