SATISFIABILITY DECAY ALONG CONJUNCTIONS OF PSEUDO-RANDOM CLAUSES

ELI SHAMIR

Institute of Mathematics and School of Computer Science
The Hebrew University of Jerusalem

Abstract. k-SAT is a fundamental constraint satisfaction problem. It involves $S(m)$, the satisfaction set of the conjunction of m clauses, each clause a disjunction of k literals. The problem has many theoretical, algorithmic and practical aspects.

When the clauses are chosen at random it is anticipated (but not fully proven) that, as the density parameter m/n (n the number of variables) grows, the transition of $S(m)$ to being empty, is abrupt: It has a "sharp threshold", with probability $1 - o(1)$.

In this article we replace the random ensemble analysis by a pseudo-random one: Derive the decay rule for individual sequences of clauses, subject to combinatorial conditions, which in turn hold with probability $1 - o(1)$.

This is carried out under the big relaxation that k is not constant but $k = \gamma \log n$, or even $r \log\log n$. Then the decay of S is slow, "near-perfect" (like a radioactive decay), which entails sharp thresholds for the transition-time of S below any given level, down to $S = \emptyset$.

1. Introduction

Propositional Logic has always been the ground floor of reasoning. In recent decades truth-value expressions and Boolean functions play a direct prominent role in specification and verification of designs, in computational modelling, combinatorial algorithms and complexity [5, 6].

Questions from theory and applications in these areas often reduce to satisfiability of logical formulas expressing constraints on the values of the Boolean variables. The study of the structure of the satisfying set S, decision algorithms for $S = \emptyset$, search algorithms for elements of S and related issues became a lively topic of interdisciplinary research.

A popular and transparent way to present constraints, abbreviated CNF-SAT or simply SAT, is by a conjunctive normal form

$$\mathbf{C}(m) = C(1) \wedge \cdot \wedge C(m); \quad C(i) = y_1 \vee \cdots \vee y_k \tag{1.1}$$

where each y_j is a literal, i.e. a variable x_i or its negation $\overline{x}_i$, $1 \le i \le n$. To satisfy the function f presented in (1.1), truth-value assignment (valuation) must satisfy all the m constraints, so all $C(i)$ should be true. To refute f, it suffices to refute one clause.

Definition 1.2. The refutation cell of $C(i)$ in (1.1) is

$$q = q(i)\colon y_1 = 0, \ldots, y_k = 0; \tag{1.3}$$

thus q is a k-codimensional subspace of $Q = \{0,1\}^n$, the space of all valuations over the n variables.

The satisfaction set of (1.1) is denoted $S(m) \subseteq Q$, and $S(0) = Q$. Upon joining another clause $C(m+1)$, $S(m)$ depletes to

$$S(m+1) = S(m) - S(m) \cap q(m+1) \tag{1.5}$$

This relation is very general. It holds whenever the constraints are present as a conjunction of "clauses", usually each clause depends on few variables, the refutation cell of a clause is derived from the type of the clause [see Section 7]. In this article we study the depletion, or decay, process of S for $k-SAT$ formulas like (1.1), where the width (number of literals) in a clause is uniformly k. The length ("time" in Section 5) m will be a power of n, and $n \to \infty$.

Remark 1.6. The dual cover process. Notice that $q(i)$ in (1.3) is the satisfying set of the clause dual to $C(i)$, namely $\overline{y}_1 \wedge \cdots \wedge \overline{y}_k$. This gives a dual point of view, of a **cover process** (in Q) formed by the union of the q cells, i.e. the valuation satisfying the disjunctive formula (DNF) dual to (1.1). Keeping both views is useful: in Section 5 we use the "coupon collector" random cover paradigm as the "end game" of the decay process.

In perspective, or briefly outlined in Section 2, most k-SAT studies treat the random ensemble, where k is constant and $\mathbf{C}(m)$ is an element of a suitable sample space. A sharp threshold ("phase-transition") behavior is anticipated:

$$\begin{aligned} \text{Prob}\,\{S(f,m) \neq \emptyset\} &= 1 - o(1) \quad \text{if } m/n < c_1, \\ &= o(1) \quad \text{if } m/n > c_2, \\ \text{and} \quad c_i &= c \pm o(1) \quad \text{as } n \to \infty. \end{aligned} \tag{1.7}$$

The case $k = 3$ is the hardest, extensively studied and experimented with, for good reasons [5, 11, 14].

Relaxation, k grows slowly with n. In this case the cells q are not "huge", but of relative size $o(1)$ in Q. The depletion process (1.5) is slow.

For simplicity, we take $k = \gamma \log n$, (the basis of log is 2). But (see Section 4) our analysis works for slower rates, down to $k = r \log\log n, r > 1$. In this situation, we are able to replace the random ensemble study by an induction proof (on m), of a tight decay rule for "typical" individual sequence $\mathbf{C}(m)$, subject to natural combinatorial conditions which are "pseudo-random" (section 3). The decay rule is termed "near perfect decay," abbreviated NPD. It's like a radioactive decay with near-constant rate.

Intermediate level thresholds. In Section 5 we show that NPD entails sharp thresholds, not only for crossing to $S(m) = \emptyset$, but also for size (S) crossing

below some fraction of Q. In Section 6 we show how this leads to an efficiently-testable unique representation in the pseudo-random classes $\{\mathbf{C}(m)\}$, which in turn is very handy for a design of Learning algorithms, reconstructing a hidden formula in the class from a small sample of its values.

2. Brief survey, Mostly for Random k-SAT

Molloy's 2001 article [14] and Clote–Kranakis' book [5] give extensive surveys. For recent results, see [1, 2].

The phase transition of S in k-SAT, from non-empty to empty, is complex and intriguing, serves as a meeting ground for interdisciplinary research, each culture employs its typical arsenal. The relation (1.7) as formulated, is rigorously proved only for $k = 2$, where the threshold value is $c_1 = c_2 = 1$. The proof relies on a combinatorial criterion for $S = \emptyset$ [4, 14], where the threshold value is $c = 1$.

For $k \geq 3$, studies were done in three main directions.

Direction 2.1. Upper and lower bounds are established (and improved) for the values c_1 and c_2 in (1.7). Variants of the first and second moment methods are used and, for c_1, algorithms to search an element of S.

Direction 2.2. Statistical Physics Studies [11, 12] adapt methods and calculation techniques from spin-glasses phase-transitions and annealing, relying on a strong analogy between these phase transitions, coming from different domains. The methods are partly non-rigorous. Physicists trust them because they conform well to physical theory and extensive experimentations (numerical ones for k-SAT). Beyond threshold values, they study the ramified structure of S, its connectivity components near the threshold, and search algorithms for elements of S. It is well known and emphasized that 3-SAT decision problem is Cook's canonical NP problem to which all NP-complete problems reduce; in many cases the reductions are useful in practice. Convincing indications show that the high complexity of decision and search are concentrated at values of m/n near the threshold.

Direction 2.3. Friedgut [7, 8] introduced a sweeping method to prove existence of sharp thresholds in many phase-transition situations, including k-SAT. It is weaker than (1.7) since the scalars c_1, c_2 are replaced by functions $c_1(n), c_2(n)$ ("non-uniformity"). Indeed $c_2(n) - c_1(n) = o(1)$, but the diminishing intervals are not proved to be convergent.

Intermediate level thresholds relate to the events

$$\text{Card}(S) = |S| \text{ crosses below } 2^{bn},\ 0 < b \leq 1 \tag{2.4}$$

Thresholds here should also be on the linear scale $m = c \cdot n$. To our knowledge, nothing deep is known about (2.4). Our cursory attempts point out that methods of 2.1 and of 2.3 (Friedgut's), which work for $b = 0$ ($S = \emptyset$), are not easy to extend to $b > 0$. Methods of 2.2 (Physics) should work and yield intermediate threshold values. Since they are expected to give information and provide a picture how

S is approaching the crossing to $\emptyset$ (as m/n grows), it is surprising that these calculations were not carried out.

The big relaxation: as k grows slowly to ∞ with n, the relative size of a refutation cell is $o(1)$. Then the decay of S is slow, follows an NPD rule. Our first proof was done for the equivalent k-DNF cover process, presented in Gerlitz M.Sc. thesis [10] from 2002, and also had limited circulation. Independently, Frieze and Wormald [9] proved sharp threshold for random k-SAT if $k \gg \log n$.

Drastic closure of $c_2 - c_1$ gap. Perfect decay would place the threshold for $S = \emptyset$ around $2^k(\ln 2)$. In 2002 Achlioptas and Moore got it up to a factor of 2 [1]. In 2003, Achlioptas and Peres [2] got it within $O(k)$ of that value. This result (for the random ensemble) subsumes sharp threshold for $k = \gamma \cdot \log n$ and also smaller rates of k (growing to ∞ with n). These articles use ingenious adaptations of the second-moment method, but are not algorithmic.

3. Definitions, pseudo-random conditions

Our study of satisfiability decay for a k-SAT process is based on formula (1.5): The next constraint is a refutation cell $q = q(m+1)$

$$q\colon y_1 = 0, \ldots, y_k = 0, \quad y_i, 1 \le i \le k, \text{ literals "present in } q", \tag{3.1}$$

all other variables can be 0 or 1, so setwise q is a subspace of codimention k.

Definition 3.2. The Ladder of q is the skewed tree of subspaces, from $Q = Q(0,0)$ to $q = Q(k,0)$:

$$Q(i,\zeta) \text{ is } y_1 = 0, \ldots, y_{i-1} = 0, y_i = \zeta, \quad \zeta = 0,1,\ 1 \le i \le k. \tag{3.3}$$

At the ith step of the ladder

$$Q(i-1,0) = Q(i,0) \cup Q(i,1), \tag{3.4}$$

a split into two halves, the two sides of y_i.

In our proofs we consider restrictions of the given process C to "ambient subspaces" like $Q(i,\zeta)$, in which i literals are **hardwired**. In the restricted process, cells, sets will be primed: C', q', S' etc. The relative size will be

$$\|A'\| = |A'|/2^{n-i} \text{ (since the ambient space has } 2^{n-i} \text{ vectors)}. \tag{3.5}$$

We proceed to define, for cells in the process **C**

$$T(q,y) = \{ \text{ cell } r \text{ preceding } q \text{ in } C, \text{ which share } y \text{ or } \bar{y} \text{ with } q\} \tag{3.6}$$

$$excess(q,y) = \textstyle\sum_r \eta(r,q,y), \text{ sum over } r \in T(q), \text{ where } \eta(r,q,y) = [-1]{+}1 \text{ if polarities of } y \text{ in } r \text{ and } q \text{ [dis]agree.} \tag{3.6$'$}$$

$$T(q) = \bigcup_{y \in q} T(q,y), \quad excess(q) = \sum_y excess(q,y) \tag{3.7}$$

(3.7′) $excess[T(q)] = \sum_r excess(r)$, sum over $r \in T(q)$.

The pseudo-random conditions. These conditions will hold with probability $1-o(1)$ in the standard sample space SP of processes with $m = Dn^{1+\gamma}$ clauses of width $k = \gamma \log n$, then we assume $0 < \gamma < 1/2$, $\epsilon > 0$ small and $q = q(m')$.

PR1 $|T(q,y)| = m'k/n(1 \pm \epsilon)$ $\left(\begin{smallmatrix}\text{Expected value in SP}\\ \text{is } m'k/n\end{smallmatrix}\right)$.
PR2 $\sum_{y,z\in q} |T(q,y) \cap T(q,z)| \leq 5 \log n$ $\left(\begin{smallmatrix}\text{Expected value}\\ \text{in SP}<<1\end{smallmatrix}\right)$.
PR3 Two cells share at most 3 variables.
PRW4 $|excess(q)| = |T(q)|^{1/2\pm\epsilon}$.
PRW5 $|excess \sim (T(q))| \leq |T(q)|^{1+\epsilon}$.

Reasoning for the SP claim: PR1 and PR2 follow from Chernoff bound; PR3 is a "birthday paradox" type applied to 4-tuples of variables which actually appear in clauses (or cells) of the process.

As for PRW 4–5, we first outline a way to form SP, in two stages. Stage 1 chooses m random k-tuples of variables. Stage 2 chooses $\pm$ polarities for the variables, **going backward** from the right-end of the process sequence. The "tail sets" $T(q)$ are determined in Stage 1. The polarities of the literals in q same as the positive reference frame for computing $\eta(r,q,y)$ and the excess-count $excess(q)$. Thus PRW4 is a standard simple random walk distance estimate [2]. Similarly for PRW5, the excess-count is over the cells of $T(T(q))$ but the positivity frame in $T(q)$ is modified as explained in Step 4 in the proof of Theorem 4.6.

4. Near-Perfect Decay Theorems

Our basic recursion scheme computes $decay(q)$ in terms of $decay(r)$, $r \in T(q)$:

$$decay(q) = \|S(m+1)\|/\|S(m)\|, \quad \text{upon joining } q = q(m+1); \tag{4.1}$$

the invariant form is

$$decay(q) = 1 - \|q\| \cdot [slack(q)] = 1 - \|q\| \cdot [1 \pm 2\|q\|(1+\rho)|excess(q)|], \tag{4.2}$$

thus $slack(q)$ [in brackets] is a small interval around 1, equality here (and also in PRW4) is an interval containment. We use $1 \pm \epsilon \approx \exp(\epsilon)$, when $\epsilon = O(n^{-\beta})$; this will not effect our estimates.

Remark 4.3. If y ia a literal in $r \in T(q)$, then setwise r is disjoint from $y = 0$ and fully contained in $y = 1$; (this is reversed if $\bar{y}$ is in r). So upon hardwiring the process to $y = 0$, the induced r' drop out as a refutation cell, its decay factor on S' is 1, while in $y = 1$ the same r just loses the literal y and so its relative size $\|r'\|$ doubles to $2\|r\|$. Moreover, since clearly

$$decay(r) = \frac{1}{2}(1 + decay(r')), \quad \text{[due to the split by } y\text{]}; \tag{4.4}$$

$$decay(r) = 1 - \|r\| slack(r) \quad \textbf{iff} \quad decay(r') = 1 - 2\|r\| slack(r) \tag{4.5}$$

Theorem 4.6 (Basic NPD recursion scheme)**.** *Let $k = \gamma \log n$, $0 < \gamma' < \gamma < 1/2$, $n^{1+\gamma'} < m \leq Dn^{1+\gamma}$. Assume the k-SAT process up to index $(m+1)$ satisfies the PR conditions. Then the NPD relation for joining $q = q(m+1)$ follows from the NPD for $r \in T(q)$, with the invariant form* (4.2)*.*

Proof. From formula (1.5) we see that the $(m+1)$-slack value in $decay(q)$ coincides with the value of *slack* in

$$\|S(m) \cup q\| / \|S(m)\| \cdot \|q\| = slack(q) \quad (= 1 + \delta) \tag{4.7}$$

we first outline four steps in proving that (4.7) gives indeed $slack(q)$ as defined in (4.2)

Step 1: going down the ladder of q, we express 4.7 as product of ratios

$$\|S'(m) \cap Q(i,0)\| / \|S'(m) \cap Q(i,1)\| \quad (= 1 + \delta_i) \tag{4.8}$$

which compare sizes for $S'(m)$ on both sides of the i-th ladder stage.

Step 2: in the i-th ratio (4.8), we express both sides as product of $decay(r')$, $r \in T(q,i)$. Then taking again product over i (by Step 1), (4.7) is expressed as a "big product" (or exp-sum) over $decay(r')$ with coefficients $\eta(r,q,y_i)$, $r \in T(q) = \bigcup_i T(q,i)$.

Step 3: combines the main parts $1 - \|r'\|$ in the decay factors of the big product. It will give the main part in the formula (4.2) for $slack(q)$.

Step 4: combines the deviations in the big product, showing it is $o(n^{-\beta})$ times the main parts contribution in Step3, and so subsumed by a factor $1 + \rho$.

Now the details of the proof steps: For **Step 1**, (4.1) shows how $S(m)$ splits (successively) between the two halves $Q(i,\zeta)$, $\zeta = 0, 1$. If all the ratios (4.1) are 1 (i.e. $\delta_i = 0$) then S is also exactly halved at each stage i, and then the accumulated slack in (4.7) shrinks to 1. In any case the final slack is clearly the product of the ratios $1 + \delta_i$, over $i \leq k$.

Step 2: ladder stage i the literals before y_i were hardwired before. Now y_i is set to 0 and 1. We may assume all cells $r \in T(q, y_i)$ are placed at the end of the process sequence, from index $m - \lambda + 1$ to m. Then $S'(m - \lambda, \zeta)$ will be the same for both sides, $\zeta = 0, 1$, since all previous cells do not contain y_i. Now

$$\|S'(m,\zeta) \cap Q(i,\zeta)\| = \|S'(m-\lambda,\zeta)\| \prod decay_\zeta(r'), \quad r \in T(q,y_i),\ \zeta = 0,1. \tag{4.9}$$

Upon devision ($\zeta = 0$ over $\zeta = 1$) the common $(m - \lambda)$-factor cancels out, then taking the "big product" over i (Step 1) and passing to exp-sum form we get that 4.7 is

$$\exp\left\{ -\sum_r \eta(r,q) \cdot decay(r') \ \mid \ r \in T(q) = \bigcup T(q,y_i) \right\}, \tag{4.10}$$

indeed (see Remark 4.3 and formula (3.6′)), the same r contributes decay 1 to one side and 'real' $decay(r')$ to the other side according to the value of $\eta(r,q,y_i)$ (matching of polarities).

Now by PR2 and PR3, at all stages of the ladder most r' cells get hardwired just once (when y_i is set to 0 or 1), except $5 \cdot \log n$ cells out of $|T(q,i)|$, which may get 2 or 3 hardwires and size growth up to $8\|r\|$. This creates a $(1 \pm n^{-\gamma})$

perturbations [accommodated by $(1+\rho/2)$-factor] to the argument of exp in (4.11), which comes from the vast majority of r cells, with $decay(r)$ from (4.5) and (4.2):

$$\exp - \left\{ \sum_r \eta(r,q,y_i) \|2r\| \, [1 \pm 4\|r\|(1+\rho)|excess(r)|] \right\} \tag{4.11}$$

Step 3: in the exponent, the main parts $2\|r\| \sum_r \eta(r,q,y_i)$ sum up to $A_3 = 2\|q\| excess(q)$ since $\|r\| = \|q\| = 2^{-k}$, (and there is an extra $1+\rho/2$ factor).

Step 4: for the $\pm$ parts in $\sum_r$ of (4.11), in the worst case all terms have the same sign, and again modules $(1 \pm n^{-\gamma})$ factor they compile to

$$(\pm 8\|r\|^2(1+\rho)) \sum_{r \in T(q)} \operatorname{sign} excess(r) \sum_{z \in s} \sum_{s \in T(r,z)} \eta(s,r,z), \tag{4.12}$$

indeed the double sun is an expression of $excess(r)$. Now the triple sum counts ± 1 for $S \in T(T(q))$ cells but the positivity frame coming from the literals $r \in T(q)$ in inverted if $sign\, excess(r) = -1$. Such triple sum is denoted $excess \sim (T(q))$ in PRW5 and so estimated by $|T(q)|^{1+\epsilon}$. Because of the factor $\|r\|^2 = \|q\|^2$, (4.12) is $o(n^{-\beta}(A_3))$, and so subsumed in another $(1+\rho/2)$ factor. altogether, we proved that (4.7) is contained in the invariant form (4.2) for $slack(q)$ and concluded the proof of Theorem 4.6.

We note that the main difficulty to overcome was controlling the slack deviation in the big product, which required going to pseudo-random walk on the cells y $TT(q)$. □

Now we iterate the recursive scheme in Theorem 4.6 until we hit cells with index $< n^{1+\gamma'}$ (we have flexibility in γ'; to optimize we take $\gamma' = \gamma/2$ since $\beta < \{\gamma - \gamma', \gamma/2\}$). Then we switch to:

Theorem 4.13 (The crude NPD recursion)**.** *For $m < n^{1+\gamma'}$, under the conditions PR 1,2,3 the NPD recursion from q to $T(q)$ holds with slack $= 1 \pm n^{-\beta}$.*

proof-in this case is much simpler. Since $|T(q)| << \|r\| = n^{-\gamma}$.Going down the ladder of q, Steps 1 and 2 are the same. Then in Step 3 and 4, we proceed as if all $r \in T(q)$ fall on one side and the deviation is still $\leq n^{-\gamma}|T(q)| = o(n^{-}\beta)$. □

Theorem 4.14. *Under the conditions of Theorem 4.6 the process* **C** *follows β-NPD as long as $|S| \geq n^{2\gamma}$.*

Proof. We recurse backward down to values of m where perfect decay holds. Indeed for $m = 0$ it is evident (also for hardwired process) and also for $m' << n$, where we can assume (or take as condition PRO) that the first m' cell are pairwise disjoint.

On the forward direction, how big can m grow?

The provision $|S| \geq n^{2\gamma}$ is needed to avoid a "coarse grain effect": The provision $|S| \geq n^{2\gamma}$ keeps the devision in splittings S down the ladder to be $\geq n^{\gamma}$, safely away from 1, so we do not have to hairsplit single points, which might damage our estimates.

In Section 5, to get down to $S = \emptyset$, we will use "coupre-collector" as an "end-game". □

Extensions For $\gamma > 1/2$, cells are even smaller the proofs are identical but require an adaptation of parameters (in the PR conditions) for smaller values of k, cells become larger. We can get down to $k = \theta \log\log n\, \theta > 1$, but omit details.

Remark 4.15. The crude NPD Theorem 4.13 suffices for the applications in Section 6, and it allows **C** to have considerable variability in width of clauses.

5. Time Analysis — Sharp Thresholds All The way

When a process S has a perfect decay rate θ^*, the time m for S to get down from relative size 1 to $1 - \alpha$ (intermediate level) solves

$$(\theta^*)^m = 1 - \alpha, \quad \text{so } m = \ln(1-\alpha)/\ln(\theta^*) \tag{5.1}$$

By the NPD theorem, the actual rate θ is near $1 - \|q\| = 1 - n^{-\gamma}$. We easily check that

$$\theta/\theta^* = 1 \pm Bn^{-\gamma-\beta}; \quad \ln(\theta/\theta^*) = \pm Bn^{-\gamma-\beta} \tag{5.2}$$

Going from θ^* to θ in (5.1) gives time

$$m = \ln(1-\alpha) \cdot n^{\gamma}(1 \pm Bn^{-\beta}) \tag{5.3}$$

the relative **width of the threshold** for m is $\tilde{O}(n^{-\beta})$.

If $\alpha = 1 - 2^{-an}$ ($a = 1$ for near empty)

$$m = n^{1+\gamma} a(\ln 2)(1 \pm \tilde{O}(n^{-\beta})), \quad \beta = \gamma/2 - \epsilon \tag{5.4}$$

However, by Theorem 4.14 the NPD works till S goes down to $n^{2\gamma}$ points. To "collect" those, we switch to an "end game": randomly choose some extra k-clauses. Notice that their refutation cells have relative size $n^{-\gamma}$.

Lemma 5.5 ("coupon collector")**.** *$N = n^{\gamma} \cdot 3\ln n$ random cell choices suffice to collect any specific set of $n^{2\gamma}$ points, with probability $1 - O(n^{-\gamma})$.*

Proof. The probability that a cell of size $n^{-\gamma}$ misses a specific point is $1 - n^{-\gamma}$. Thus, all N random cells miss it with probability $\leqq \exp(-3\gamma \ln n) = n^{-3\gamma}$. By the union bound, the probability these N cells miss **some** point is $O(n^{-\gamma})$. Since $\beta < \gamma$, the time (the number of cells) in the end-game is absorbed in the $\tilde{O}(n^{-\beta})$ term of (5.4).

Thus we proved a sharp threshold phase-transitions all the way down to $S = \emptyset$. □

6. Testable Unique Representations

Canonical forms of Boolean functions which admit the property of a unique representation within the class are useful, especially for algorithmic design. Ordered binary decision diagrams (OBDDs) have this property, but CNF, DNF do not have it. However, we show now that the crude NPD Theorem 4.13 provides us with a testable unique-representation classes.

For simplicity, we give detailed proof for the class $\{\mathbf{C}(m)\}$ with parameters

$$k = \gamma \log n, \quad \gamma < 1, \quad m = (\log 2) n^{\gamma}, \quad \text{so } \|S(m)\| \sim 1/2 \tag{6.1}$$

The proof extends to classes with other parameters, provided $\|S(m)\|$ is not too close to 0. By remark 4.15, the class can have some flexibility in m, and in width of clauses within one sequence $\mathbf{C}$.

Non-uniqueness for f in a k-CNF class implies that there is some clause (with refutation cell q) which partakes in another representation of f but is **external** to the given representation of f. Clearly then

$$S(f) \cap q = \emptyset \tag{6.2}$$

However, we shall prove, for any external clause

$$\|S(f) \cap q\| \geqq 1/2 \|S(f)\| \cdot \|q\| = 1/4 \|q\| \tag{6.3}$$

Theorem 6.4 (uniqueness)**.** *Consider the class* $\{\mathbf{C}(m)\}$ *with parameters given in* (6.1)*, and subject to*

$$\text{PR1. } \textit{For each } 1 \leq i \leq n, \ [\sharp \textit{cells containing } x_i] \leq 5 \log n \tag{6.5}$$

$$\text{PR2. } \textit{Two clauses share at most one index} \tag{6.6}$$

Then (6.3) *holds for any k-clause (with refutation cell q) which is external to the given representation.*

Proof. Assume q is given by $x_1 = 1, \ldots, x_k = 1$ and let $\mathbf{C}'$ be the process hardwired to q (which is $n - k$ dimensional). Clauses of $\mathbf{C}$ with x_j, $1 \leq j \leq k$, are satisfied and drop out. By PR1 this decreases the length m by $\leq 5k \log n$. In clauses with $\bar{x_j}$, $1 \leq j \leq k$, we can simply erase this literal. After hardwiring, the width of clauses in C' ranges from 1 to k (a clause in $\mathbf{C}$ is not fully refuted in q unless it has q as its refutation cell, in which case it is **not** external).

Case I. There is a clause y or $y \vee z$ in $\mathbf{C}'$. Then no other clause of $\mathbf{C}'$ has width $\leq k-3$, else the pre-image of the two will share two indices from $\{1, \ldots, k\}$, contrary to PR2 in (6.6) form. Now we hardwire $y = 1$ and apply Theorem 4.13 (see Remark 4.15), and obtain a satisfying set of size about $1/2$ in $q \cap (y = 1)$, about $1/4$ of q.

Case II. All Clauses of $\mathbf{C}'$ have width $l_i \geq 3$. In this case we study the complement $q \setminus S$, which is the union of the refutation cells of C' (i.e., it's a cover process).

Let A be the class of cells with $3 \leq l_i \leq k/3$. $|A| \leq 1$ since for two such clauses, their pre-image in $\mathbf{C}$ will have $\geq k/3$ indices in common, contrary to (6.6). The relative size of this unique cell is $\leq 1/8$. Next let B be the class of cells with $k/3 < l \leq k-3$, each pair of indices from $\{1, \ldots, k\}$ can be in at most one cell in the preimage of B (again by (6.6)) hence $|B| \leq \binom{k}{2}$ and the cover size (relative to q) of the union of B cells is $\leq \binom{k}{2} 2^{-k/3} \leq 1/8$ (for $n \geq 20$).

The remaining $m' = m - \binom{k}{2} - 1$ clauses have width $\geq k - 2$, and applying Theorem 4.13 their union cover has size about $1/2$ in q.

So in Case II the total cover (of $q \setminus S$) is $\leq 1/2 + 1/8 + 1/8 = 3/4$ and the satisfying set S in q is at least $1/4$ of q, proving (6.3) and Theorem 6.4. $\square$

Efficient tests for external clauses. Random samples of f's values on q give, by (6.3), an efficient test if this candidate clause is external or not. This is a good starting point for designing a Learning algorithm, reconstructing the internal clauses of a "hidden" f from enough random valuations. This is described in [10], where Fourier coefficients of f are used to construct a reasonably small set of candidate clauses and then filter the wrong ones out, by the above test. Actually, the algorithm design in [10] is for k-DNF, which is an equivalent dual.

7. Extensions, Concluding Remarks

In the Boolean domain $\{0,1\}^n$, one can consider conjunction of other types of constraints. Best known is

0–1 **coloring of a hypergraph** G**.** A valuation gives a color, 0 or 1, to all n vertices of G. The constraints are given by the (say k-uniform) hyperedges; they should not be monochromatic. Thus the refutation cell of an edge $(z_1, \ldots, z_k)$ is

$$z_1 = 0, \ldots, z_k = 0 \vee z_1 = 1, \ldots, z_k = 1 \tag{7.1}$$

it has relative size 2^{-k+1}. Other types of dually-closed constraints (refutation cells) can be considered.

Again, Friedgut's method [8] proves existence of a sharp threshold for $S = \emptyset$. Lower and upper bounds for its value were derived [1]. We can now prove NPD decay for 2-coloring of pseudo-random k-uniform hypergraphs, which entails sharp thresholds. Details are forthcoming [15].

Geometric cover problems. People studied processes of covering domains by sequences of (random) cells from prescribed families. Covering a sphere by spherical caps is an example. One can construct analogues of k-DNF covers, by cells constrained in just a few dimensions (k out of n). This seems quite artificial so we do not give details.

Back to k**-SAT.** We noted that our NPD pseudo-random analysis works for k as small as $r \log\log n$, $r > 1$. The real hard challenge is to come up with a pseudo-random analysis for the decay process of S, for k constant. It stands to reason that some unknown combinatorial conditions (explicitly known for $k = 2$) determine the phase transition to $S = \emptyset$. Discovering them will give a better picture of the behavior of S near the threshold and may improve the search methods for this ubiquitous constraint satisfaction problem.

Acknowledgement. This work owes a lot to my M.Sc. student Or Gerlitz, who pressed me hard to write down proofs of my intuitions, offered useful comments and ran numerical experiments which corroborated the NPD analysis. I thank Alex Samorodnitsky for careful critical reading of the revised proofs.

References

[1] D. Achlioptas and C. Moore, Two Moments Suffice to Cross a Sharp Threshold, *SIAM J. on Computing* (to appear)

[2] D. Achlioptas and Y. Peres, The threshold for random k-SAT is $2^k \log 2 - O(k)$, *J. of the AMS* **17** 2004, 947–973.

[3] N. Alon and J. Spencer, *The Probabalistic Method*, Wiley, NY, 1992.

[4] V. Chvatal and B. Reed, Mick gets some (the odds are given on his side), in *Proc. 33rd Annual Symposium on Foundations of Computer Science*, 1992, 620–627.
[5] P. Clote and E. Kranakis, *Boolean Functions and Computation Models*, Springer, Berlin, 2002.
[6] R. Dechter, *Constraint Processing*, Morgan Kaufmann, SF, 2003.
[7] E. Friedgut, Sharp thresholds of graph properties and the k-SAT problem, *J. Amer. Math Soc.* **12(4)**, 1999, 1017–1054.
[8] E. Friedgut, Hunting for sharp thresholds, *Random Structures and Algorithms* **26(1-2)**, 2005, 37–51.
[9] A. Frieze and N. C. Wormald, Random k-SAT: a tight threshold for moderately growing k, in *Proc. of 5th Int. Symp. on Theory and Application of Satisfiability Testing*, 2002, 1–6.
[10] O. Gerlitz, Cover Processes and Fourier Learning Algorithm related to k-DNF Formulas, M.Sc. Thesis, Hebrew University, Computer Science Dept., 2002.
[11] M. Mezard, G. Parisi and R.Zecchina, Analysis and algorithmic solutions of random satisfiability problems, *Science* **297**, 2002, 812–815.
[12] M. Mezard and R. Zecchina, Random k-satisfiability: from an analytic solution to a new efficient algorithm, *Phys. Rev. E* **66** 056126, 2002.
[13] M. Mitzenmacher and E. Upfal, *Probability and Computing*, Cambridge Univ. Press, NY, 2005.
[14] M. Molloy, Thresholds for colourability and satisfiability in random graphs and Boolean formulae, *Surveys in Combinatorics 2001*, Cambridge University Press, 165–197.
[15] E. Shamir, Sharp thresholds for hypergraph 2-coloring, the pseudo-random case, in preparation.
[16] W. Stadje, The collector's problem with group drawings, *Advanced Applied Probability* **22**, 1990, 866–882.

Institute of Mathematics and School of Computer Science, Hebrew University of Jerusalem