Phase Transitions in Random Constraint Satisfaction Problems

Konstantinos Panagiotou

(with Amin Coja-Oghlan)

k-SAT

- Given:
  - n Boolean (true/false) variables $x_1, x_2, \ldots, x_n$
  - a Boolean formula in k*-conjunctive normal form* (k-CNF)

$$F = \bigwedge_{i=1}^{m} C_i, \qquad C_i = \bigvee_{j=1}^{k} l_{i,j}$$

  where $l_{i,j}$ is a variable or the negation of a variable

- An assignment

$$\sigma : \{x_1, \ldots, x_n\} \to \{\text{true}, \text{false}\}$$

  is called *satisfying (for F)*, if it satisfies all clauses

- A clause is satisfied (by σ) if at least one literal in it is satisfied

Example ($k = 2$, 2-CNF)

$$F = (x_1 \vee x_2) \wedge (x_2 \vee \overline{x_3}) \wedge (\overline{x_1} \vee x_3)$$

- Assignment $\sigma_1 = (\text{true}, \text{true}, \text{true})$ is satisfying
- Assignment $\sigma_2 = (\text{false}, \text{false}, \text{true})$ is not

The k-SAT Problem

- Question: given F, compute a satisfying assignment or verify that there is none!
- This is a central problem in Computer Science
- If $k = 1$, then it is easy:
  - F is satisfiable iff no variable appears both negated and not negated
- If $k = 2$, then there is a linear time algorithm
  [Aspvall, Plass & Tarjan (1979)]
- If $k \geq 3$, then the problem is NP-complete
  [Cook & Levin (1971)]

General Setting: CSP

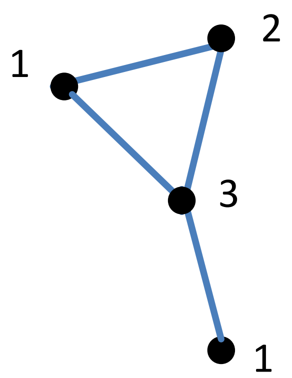
- Constraint Satisfaction Problems
- Given:
  - Set of *variables* $X = \{x_1, \dots, x_n\}$, finite domain D
  - Set of *constraints* $C = \{c_1, \dots, c_m\}$, where

$$c_i = (X_i, F_i) \quad \text{with} \quad X_i \subset X, F_i \in X_i \to D$$

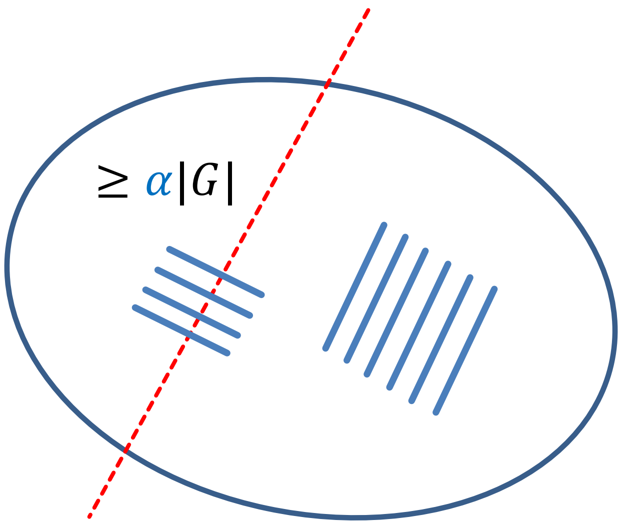
- F_i is a *forbidden assignment* to the variables in X_i
- Question: given (X, C), is there any assignment $\phi: X \to D$ such that all c_i are *satisfied*, that is, $\phi|_{X_i} \neq F_i, 1 \leq i \leq m$?

Other Examples

- k-COL
- Given: a graph G
- Question: is it possible to color the vertices of G with k colors such that *any two neighbors receive different colors*?



- α-ISET, where $\alpha \in (0,1)$
- Given: a graph G
- Question: is there an *independent set* that contains at least an α-fraction of the vertices?



Why are CSPs so hard?

Random Formulas

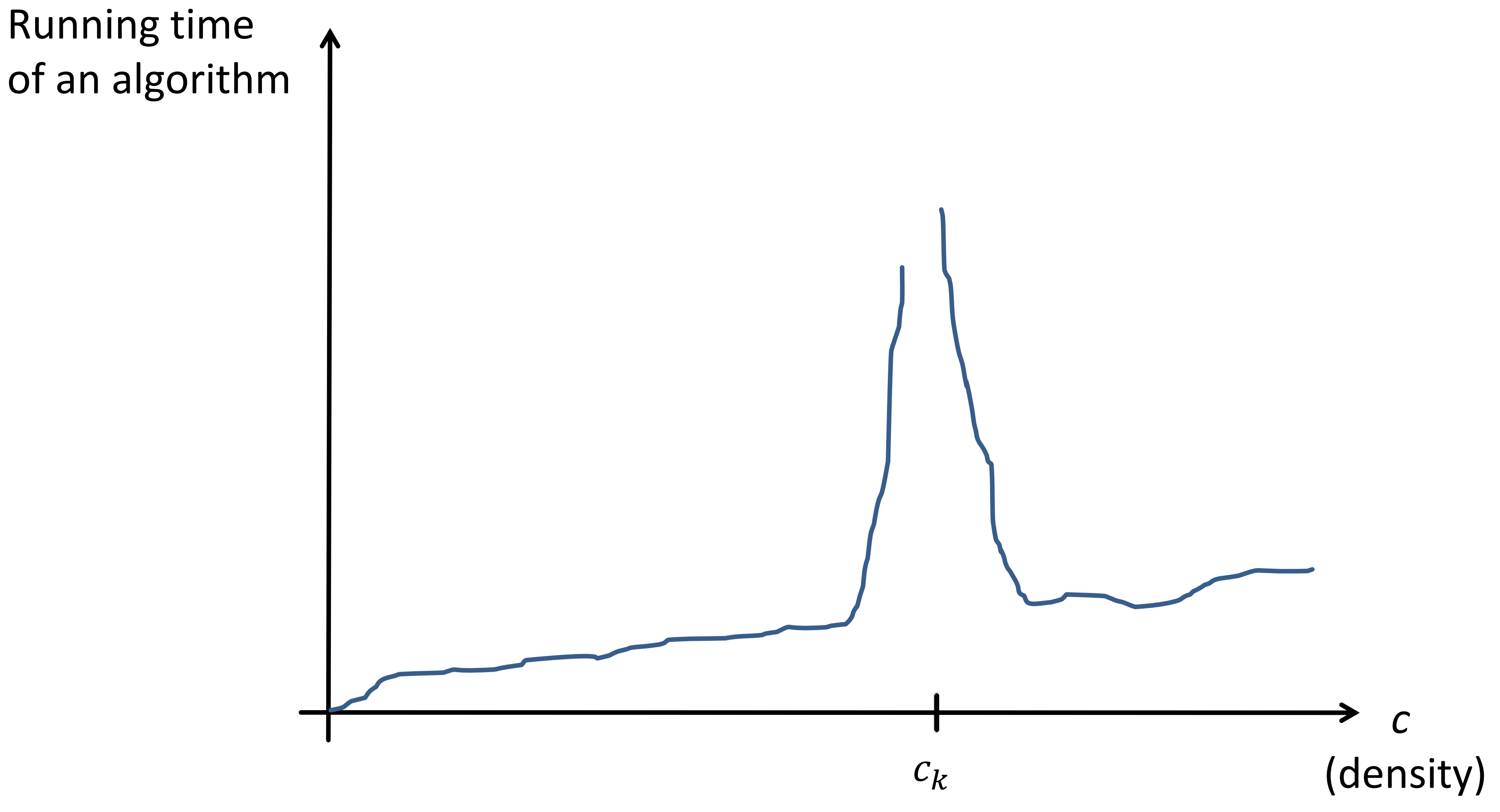
- Setup:
  - n Boolean variables $x_1, \ldots, x_n$
  - $m = \lfloor cn \rfloor, c > 0$
  - $F_{n,m}$ is a k-CNF with m clauses, where each clause is drawn uniformly at random from the set of all possible clauses
- We call c the *density* of the formula
- Initial motivation for studying random k-SAT: the „most difficult“ instances seem to be around a specific $c = c_k$

A Generative Procedure

- Generate $F_{n,m}$ as follows:
  - for $i = 1 \ldots m$ // Generate C_i - ith clause
  - for $j = 1 \ldots k$ // Generate jth literal in C_i
    - $l_{i,j} \coloneqq x_I$, where I is uar (uniformly at random) from $\{1, \ldots, n\}$
    - With probability ½ set $l_{i,j} \coloneqq \overline{l_{i,j}}$ (i.e. negate the occurrence of the variable)
- All random decisions are *independent*
  - Particularly, the choice of each variable occurence and of its „sign“ are distinct processes

Experimental Evaluation

- Anderson ’86, ’99, Cheesman et al ‘91



Many Questions…

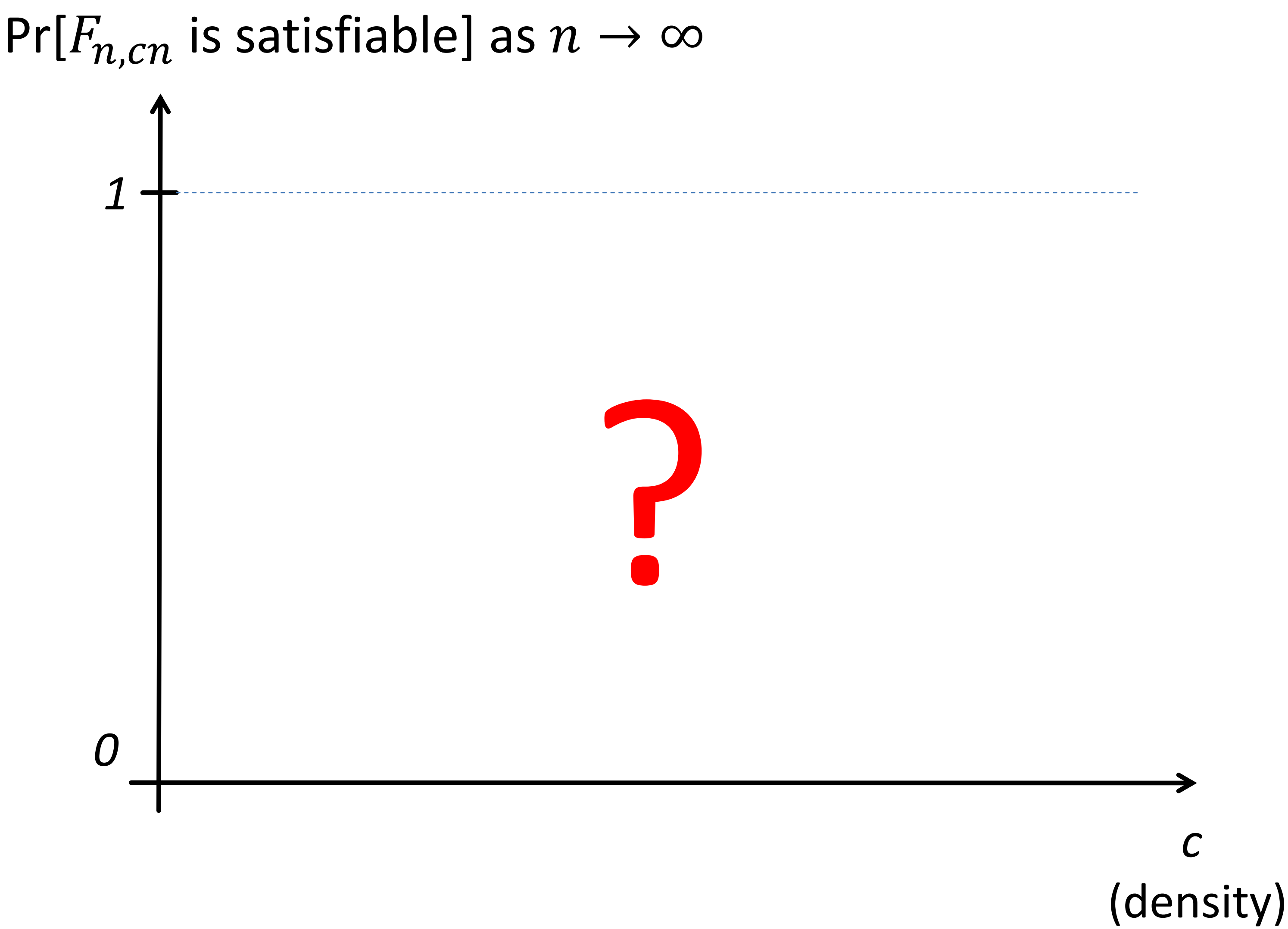
- For which densities c (# clauses $= m = cn$) is $F_{n,m}$ satisfiable whp (*with high probability*)?
- Other properties that hold whp?
- Algorithms?

- We will consider only the case $k \geq 3$ here.

Random CSPs

- Statistical physicists have developed sophisticated but non-rigorous techniques
  - detailed picture about the structural properties
  - several conjectures, algorithms
  - many papers: Krzakala, Montanari, Parisi, Ricci-Tersenghi, Semerjian, Zdeborova, Zecchina, …
- Most parts of the picture: beyond current capabilities of mathematics

Picture - Satisfiability



A First Bound

- Consider the obvious random variable

$$X = \text{\# of satisfying assignments of } F_{n,cn}$$

- If for the fixed value of c we can show

$$\mathbb{E}[X] \to 0 \text{ as } n \to \infty,$$

then $X = 0$ and $F_{n,cn}$ *is not satisfiable* whp.

- Let $X = \sum_{\sigma} X_{\sigma}$, where the sum is over all possible assignments in $\{\text{true}, \text{false}\}^n$ and

$$X_{\sigma} = \mathbf{1}[\sigma \text{ satisfies } F_{n,cn}]$$

A First Bound (cont.)

$$\mathbb{E}[X] = \sum_{\sigma} \Pr[\sigma \text{ satisfies } F_{n,cn}]$$

$$= \sum_{\sigma} \Pr[\forall 1 \leq i \leq cn\colon\ \sigma \text{ satisfies } C_i]$$

$$= \sum_{\sigma} \prod_{1 \leq i \leq cn} \Pr[\sigma \text{ satisfies } C_i]$$

$$= \sum_{\sigma} \left(1 - 2^{-k}\right)^{cn}$$

$$= 2^n \left(1 - 2^{-k}\right)^{cn}$$

$$\approx \exp(n(\ln 2 - 2^{-k} c))$$

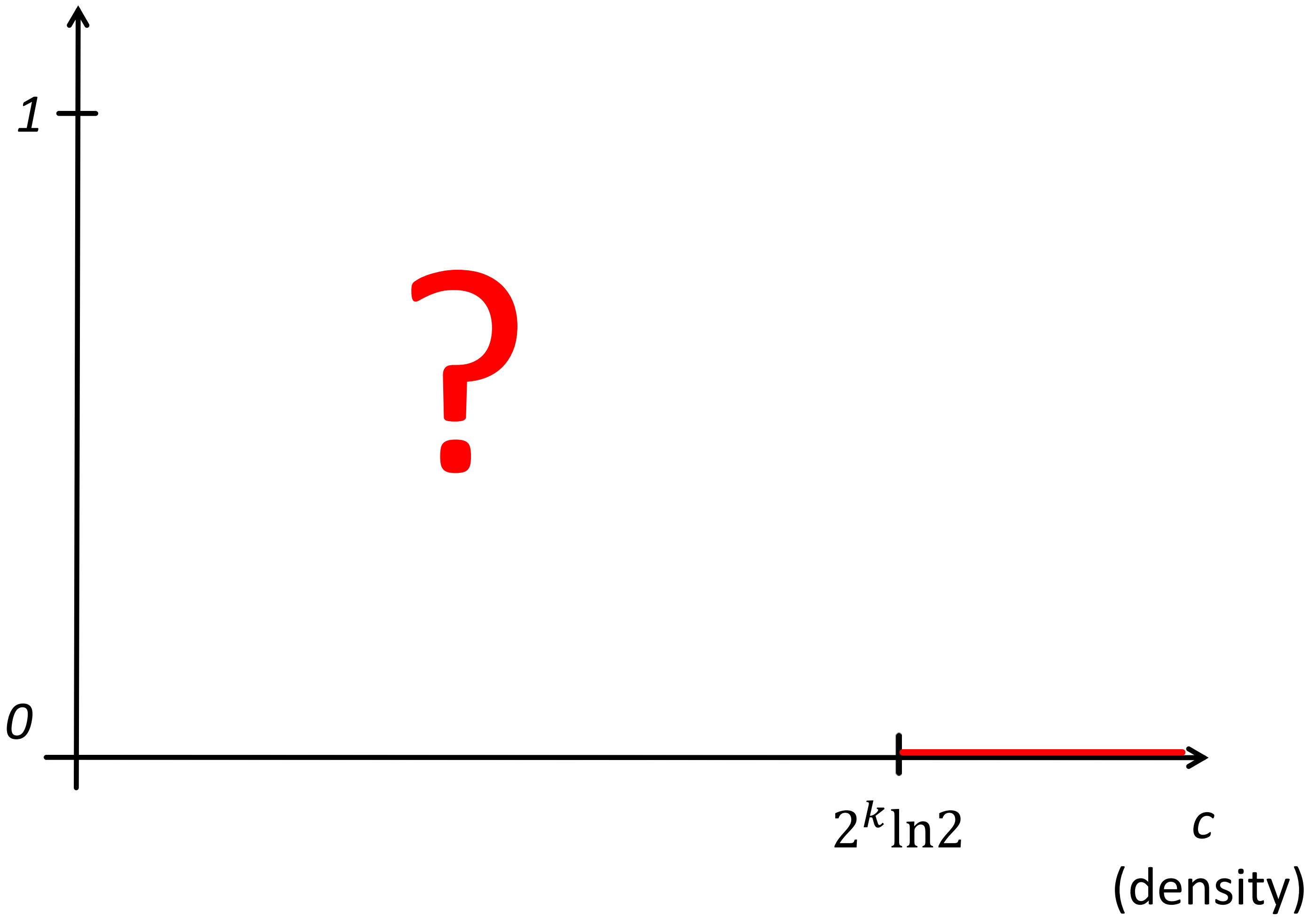
$C_i = \cdots \vee \cdots \vee \cdots \vee \cdots$

$C_i = x_{i_1} \vee \overline{x_{i_2}} \vee \overline{x_{i_3}} \vee x_{i_4}$

In σ: 0 1 1 0

Picture

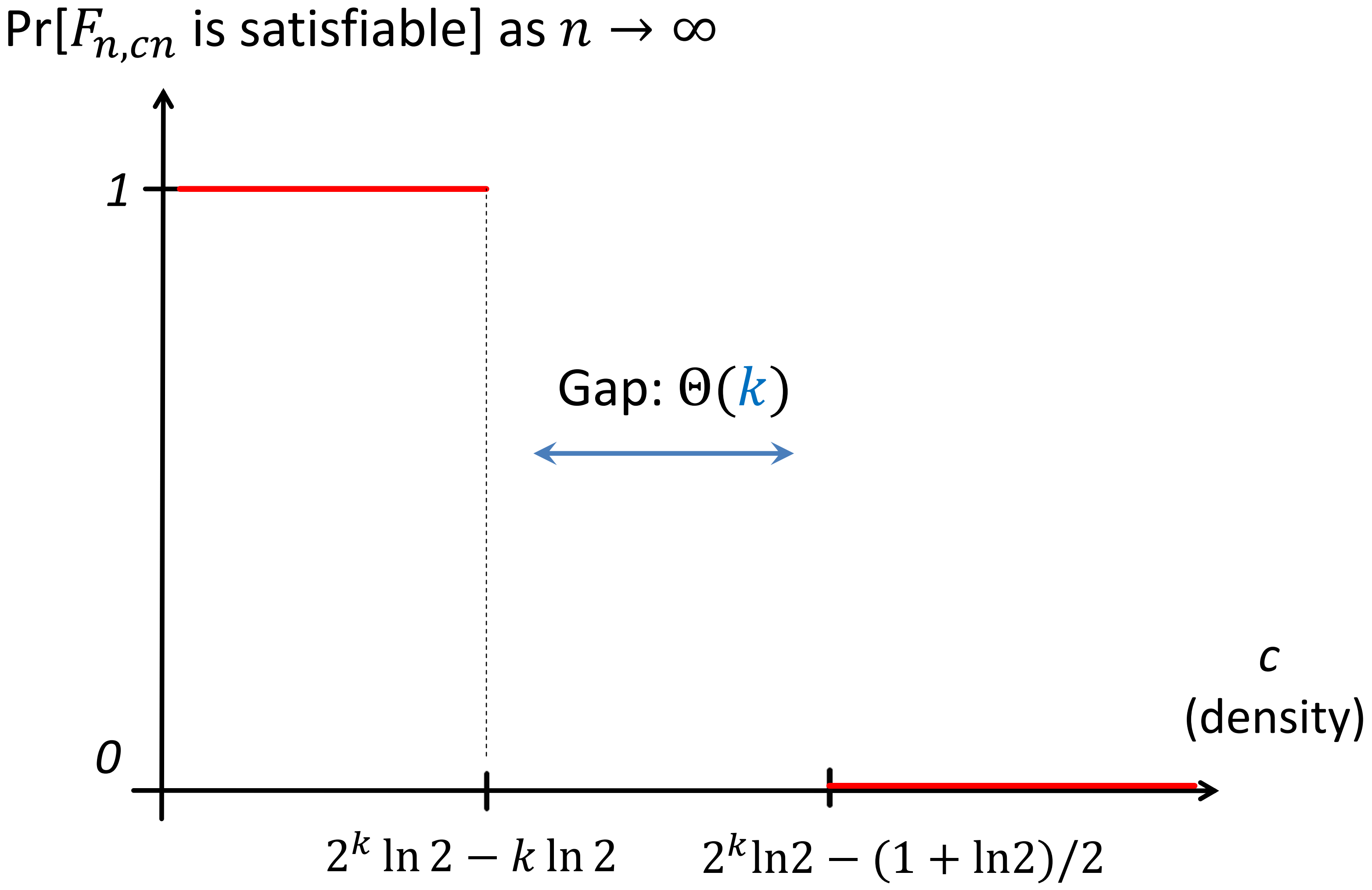
Pr[$F_{n,cn}$ is satisfiable] as $n \to \infty$



(Some) Previous Work

- Friedgut '05: There is a sharp threshold *sequence* $c_k(n)$:
  - If $c < c_k(n)$, then $F_{n,cn}$ is satisfiable whp
  - If $c > c_k(n)$, then it is not whp
- Kirousis et al. '98:
$$c_k(n) \leq 2^k \ln 2 - \frac{1}{2}(1 + \ln 2\,)$$
- Achlioptas and Peres '04:
$$c_k(n) \geq 2^k \ln 2 - k \ln 2$$

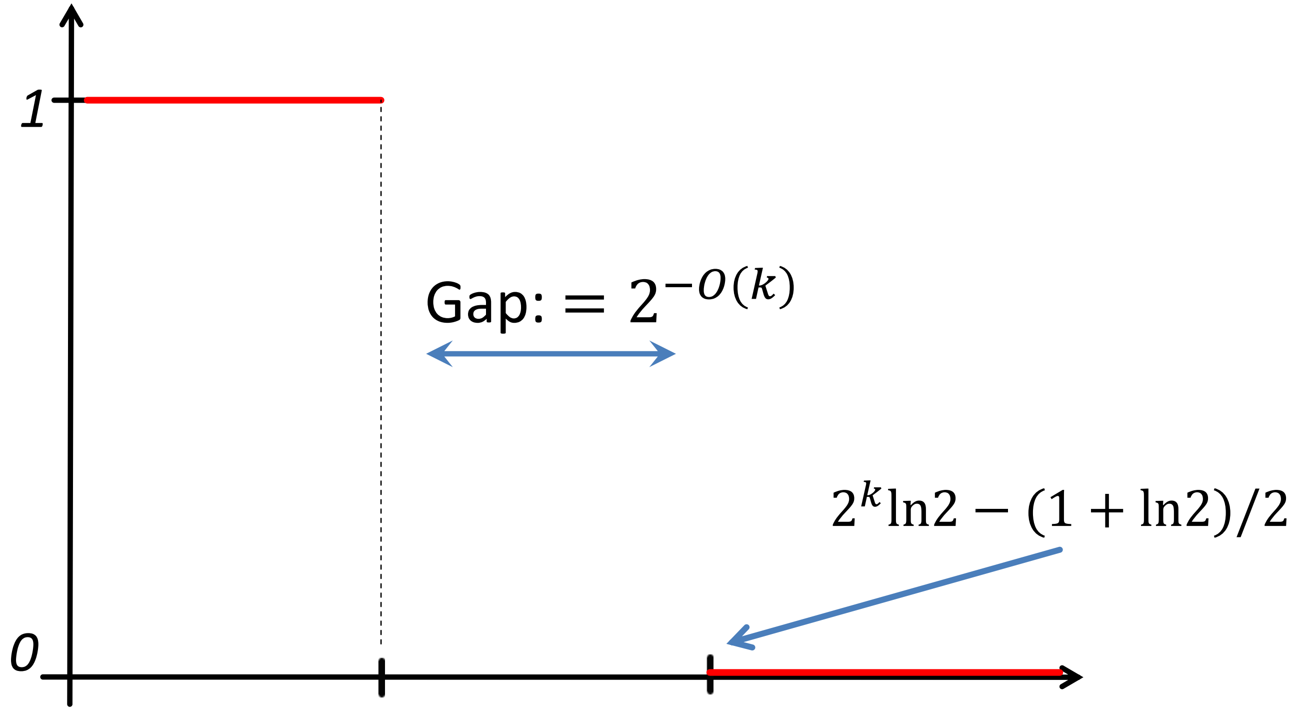
Rigorous Bounds



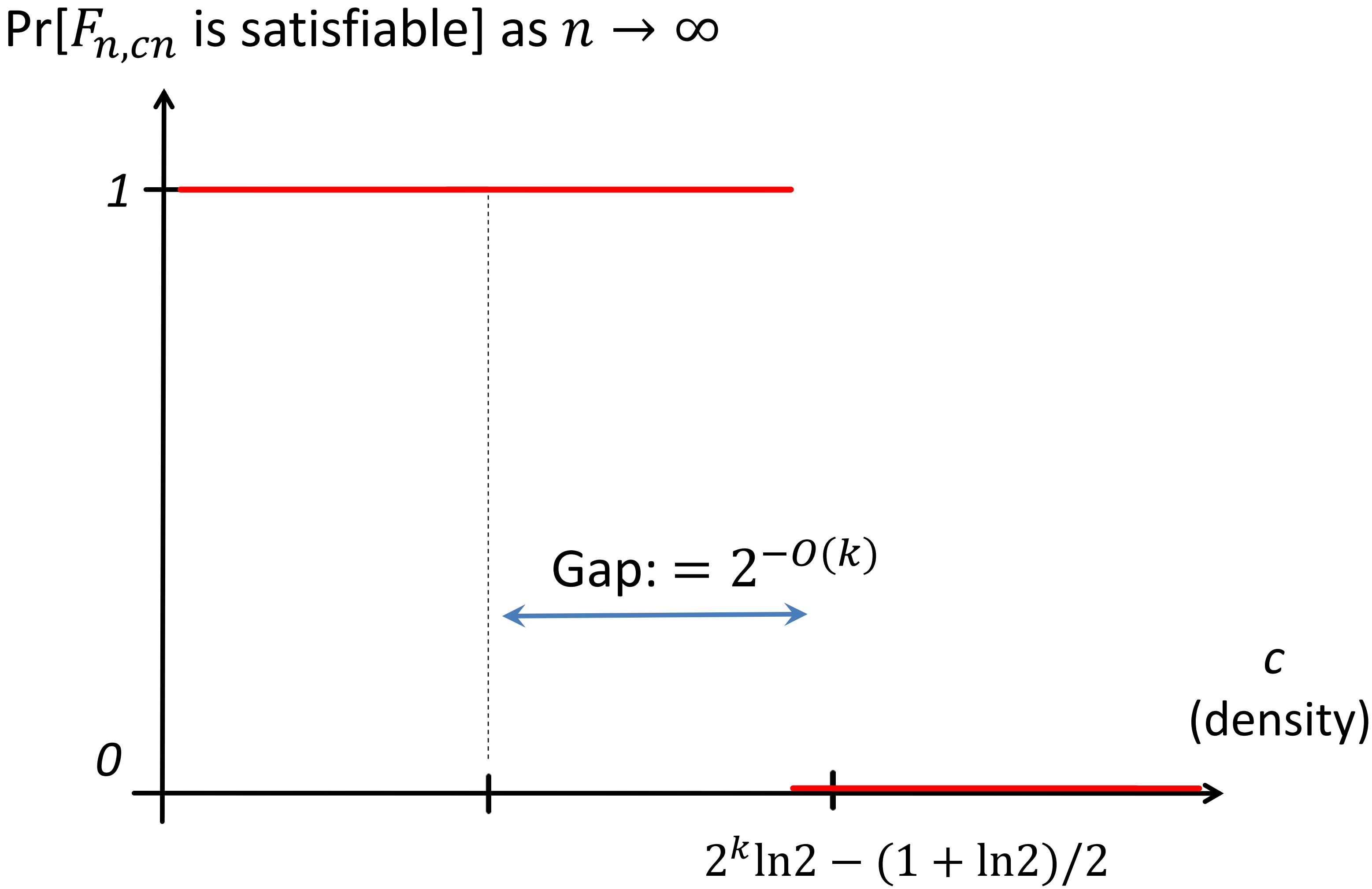
The Next Step

Coja-Oghlan, P. ’13, ’14, ‘16:

$$c_k(n) \geq 2^k \ln 2 - \frac{1 + \ln 2}{2} - 2^{-O(k)}$$



THE Conjecture for k-SAT



Satisfiability Conjecture for many CSPs

- There is a critical (problem specific) density c^* such that
  - Random instance of CSP is satisfiable if $c < c^*$
  - Is not if $c > c^*$
- Non-rigorous arguments even determine the value of c^* for several problems!

The Second Moment Problem

- If Z is a non-negative random variable

$$\Pr[Z > 0] \geq \frac{\mathbb{E}[Z]^2}{\mathbb{E}[Z^2]}$$

Paley-Zygmund Inequality
Second Moment Method

- We can apply this to X, the number of satisfying assignments of $F_{n,cn}$
- If $\mathbb{E}[X]^2 \approx \mathbb{E}[X^2]$ for the given c, then we are done!

Bound for 2nd Moment

$$\mathbb{E}[X^2] = \sum_{\sigma,\tau} \Pr[\sigma, \tau \text{ satisfy } F_{n,cn}]$$

$$= \cdots$$

$$\gg \mathbb{E}[X]\text{^}2$$

Problem: for all $c > 0$ we have that $\mathbb{E}[X^2]$ is *exponentially* larger than $\mathbb{E}[X]^2$!

Why?

An Asymmetry

- Consider a thought experiment
- Suppose that somebody makes the promise

  „x_1 appears in $F_{n,cn}$ exactly d_1 times …

  … and *all* these appearances are *positive*“
- What value do we assign to x_1?
- Other promise:

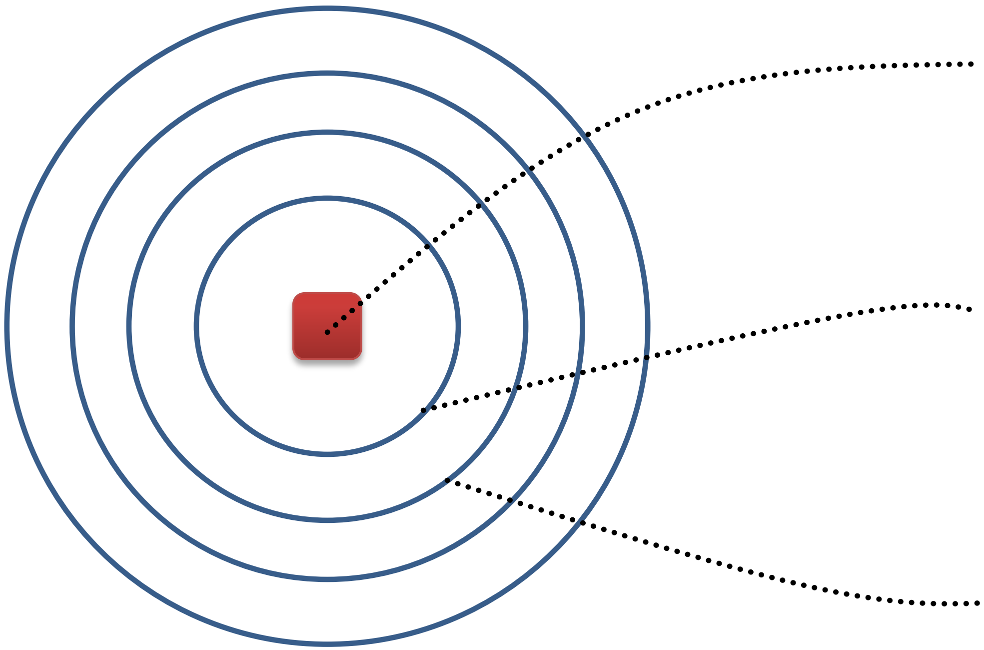
  „x_1 appears in $F_{n,cn}$ exactly d_1 times …

  … and 51% of the appearances are *positive*“
- We (should) set again x_1 to true

The Majority

- Our „best guess“ for a satisfying assignment is the *majority vote*:
  - Somebody tells us how often each variable appears positively and negatively, and nothing else
  - If x_i appears more often positively, assign it to true, and otherwise to false
- This assignment *maximizes* the probability that $F_{n,cn}$ is satisfied
- Even more: assignments that are „close“ to the majority vote have a *larger probability* of being satisfying

Picture of the Situation



- Majority assignment
- Largest probability of being satisfiable

- Distance 1
- Less probability of being satisfiable

- Distance 2
- Even smaller probability of being satisfiable

$\rightarrow$ The satisfying assignments *correlate*!

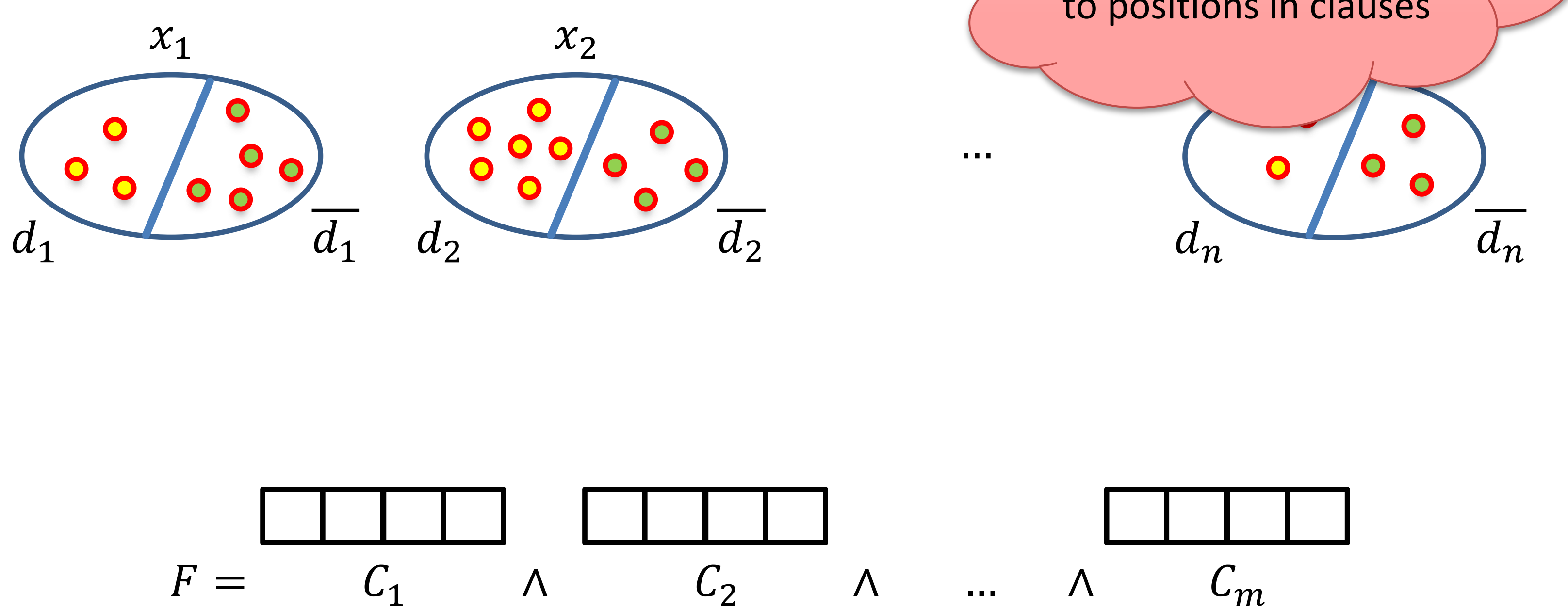
Getting a Grip on the Majority

- Generate $F'_{n,m}$ in *two steps* as follows:
  1. For *each variable* x_i choose randomly the number d_i of *positive* occurences and the number $\overline{d_i}$ of *negative* occurences.
  2. Choose randomly a *formula* where each variable x_i appears d_i times positively and $\overline{d_i}$ times negatively.
- Want: distributions of $F'_{n,m}, F_{n,m}$ are the same.
- Step 1
  – It is easy to see in $F_{n,m}$ that d_i and $\overline{d_i}$ are distributed like $\mathrm{Po}(kc/2)$, and they are almost independent

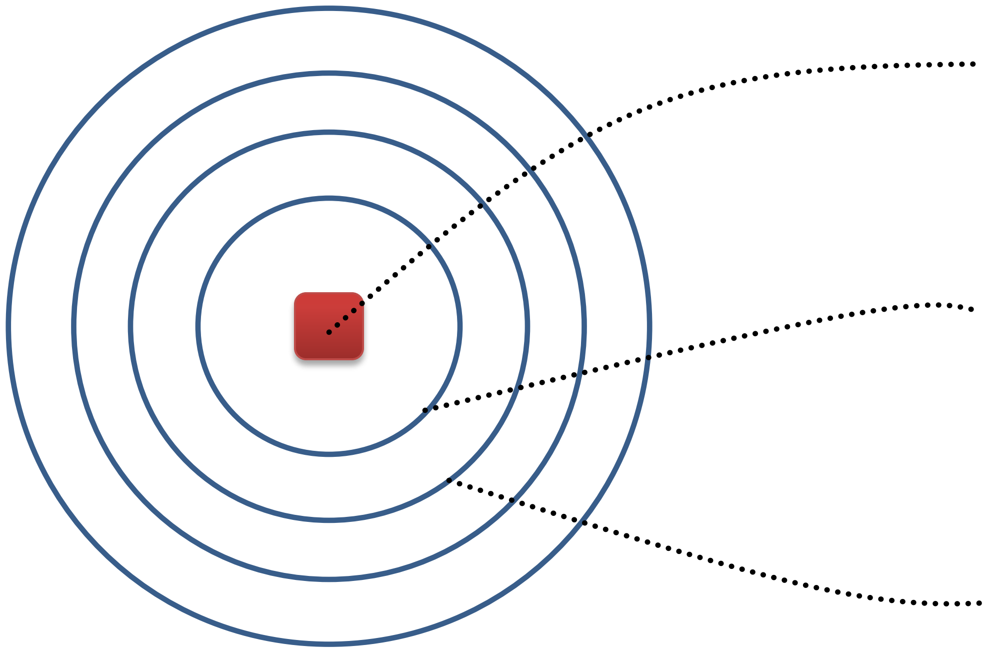
Step 2

- How do we choose a formula where each variable x_i appears d_i times positively and $\overline{d_i}$ times negatively?
- Configuration model:

Random Matching:
variable occurences
to positions in clauses



Recall the Situation



- Majority assignment
- Largest probability of being satisfiable

- Distance 1
- Less probability of being satisfiable

- Distance 2
- Even smaller probability of being satisfiable

Getting a Grip on the Majority

- Consider only *specific* satisfying assignments!
- *Intuition*: if a variable appears d times positively and $\bar{d}$ times negatively, then assign it to true *with some probability* that depends on $d, \bar{d}$ only.
- Map $p: \mathbb{Z} \to [0,1]$
- Set also $p(x_i) \ = \ p(d_i - \bar{d}_i), p(\overline{x_i}) = 1 - p(x_i)$
- Meaning: a p-fraction of the literals is satisfied under the assignments that we consider.

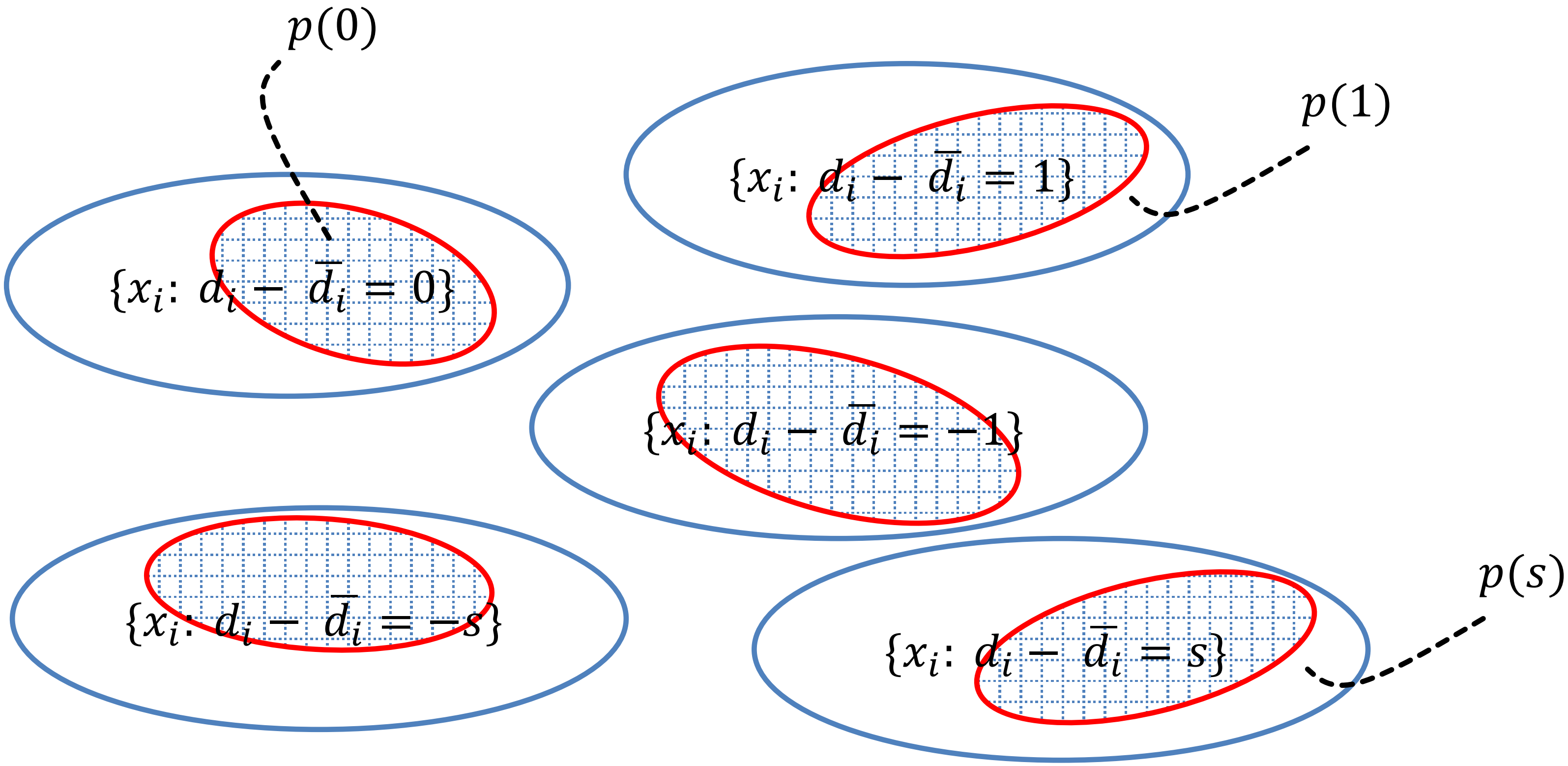
More formally

- Set $T = \{(p(x_i), 1 - p(x_i)) : 1 \le i \le n\}$
- This is the set of different „types“ of variable occurences (equivalent → $d_i - \overline{d_i} = const$)
- We say that $\sigma: \{x_1, \dots, x_n\} \to \{\text{true}, \text{false}\}$ has *p-marginals* if for all $(t, 1-t) \in T$

$$\sum_{i: p(x_i) = t} d_i 1[\sigma(x_i) = \text{true}] = t \sum_{i: p(x_i) = t} d_i$$

- That is, a t-fraction of the variable occurences is set to true, for all $t \in T$
- Question: how do we choose p?

Pictorially



Detour: Physics

- For x_i let $\mu(x_i, F)$ be the fraction of satisfying assignments in which x_i is set to true in F
- It is *NP*-hard to compute $\mu(x_i, F)$
- According to physicists: $\mu\left(x_i, F_{n,m}\right)$ can be computed by a *message passing algorithm* called *Belief Propagation* [Montanari et al '07]

Conjecture

$$\mu\left(x_i, F_{n,cn}\right) = \frac{1}{2} + \frac{d_i - \overline{d_i}}{2^{k+1}} + O\left(\frac{(d_i - \overline{d_i})^2}{2^{2k}}\right)$$

- Belief Propagation leads to a stronger prediction
  - Conjecture for μ up to an error of $o(1)$ as $n \to \infty$
  - it does depend on many parameters

Our Choice

$$p(z) = \begin{cases} \frac{1}{2} + \frac{z}{2^{k+1}}, \text{if } |z| < 10\sqrt{k2^k \ln k} \\ \frac{1}{2}, \text{otherwise} \end{cases}$$

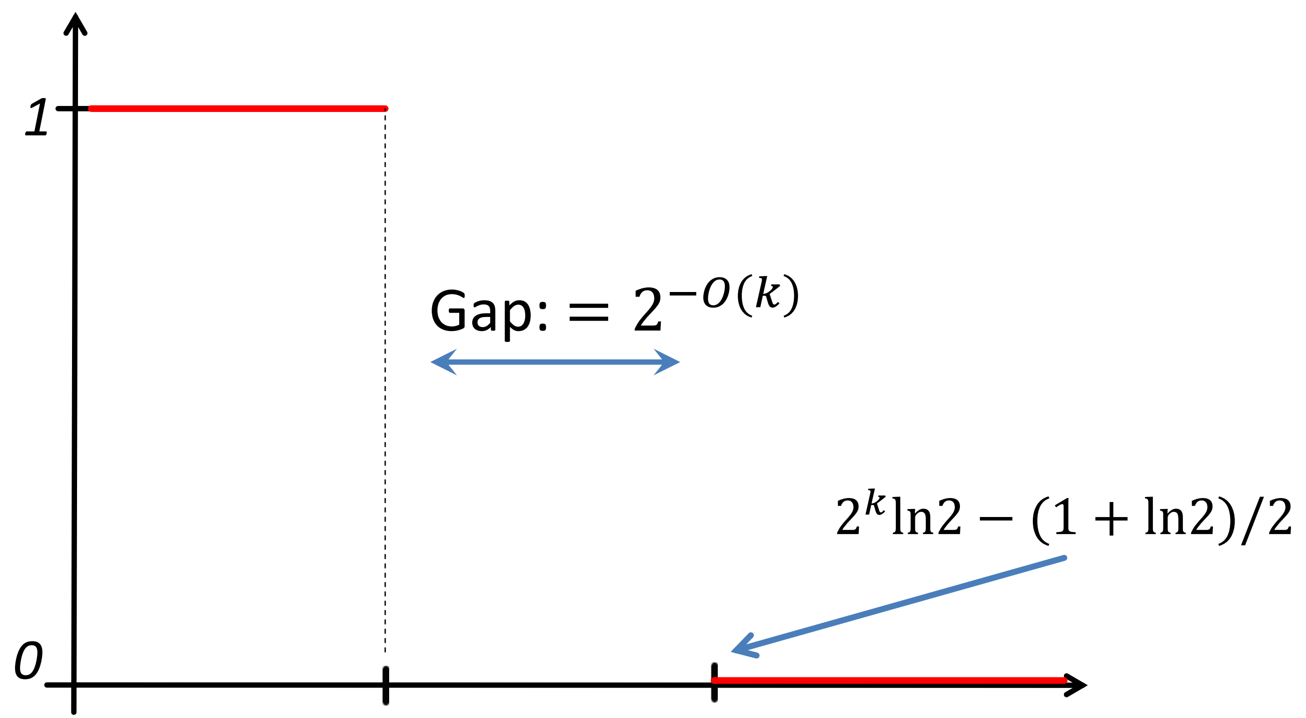
- This matches the conjecture on the „bulk" of the variables
  - Recall that $d_i, \bar{d_i} \sim \mathrm{Po}\left(\frac{kc}{2}\right) \approx \mathrm{Po}(k2^{k+1})$
  - Except of a very small fraction, all other variables have the property

$$\left|d_i - \bar{d_i}\right| = O(\sqrt{k2^k})$$

This yields...

Coja-Oghlan, P. ’13, ’14, ‘16:

$$c_k(n) \geq 2^k \ln 2 - \frac{1 + \ln 2}{2} - 2^{-O(k)}$$



Better?

- Yes!
- Not so long ago on arxiv by Ding, Sly, Sun: satisfiability conjecture is true for k-SAT, for k sufficiently large.

- Approach:
  - Work with the *correct value* for $\mu(x_i, F)$
  - This depends not only the appearances of x_i, but on the local neighborhood in F
  - Infinitely many parameters

Summary & Outlook

- The quest for the k-SAT threshold has (almost) ended
- This is only the tip of the iceberg
  - What can we say about other CSPs?
  - Algorithms for random instances?
- Rigorous translation of replica method?

Thank you!

+