Random CSPs: from Physics to Algorithms

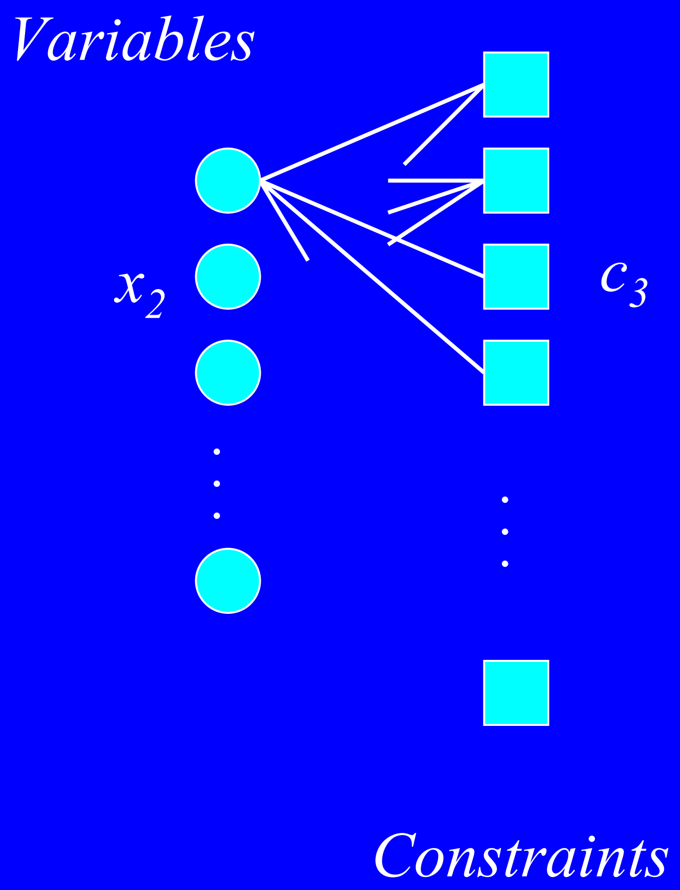
Dimitris Achlioptas
University of California
Santa Cruz

Federico Ricci-Tersenghi
University of Rome
La Sapienza

The Setting: Random CSPs

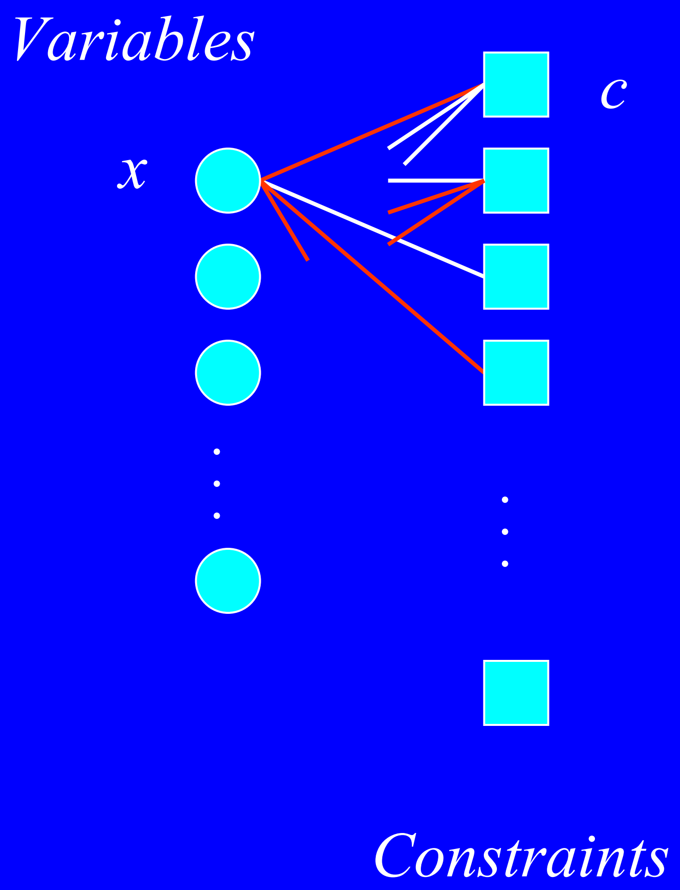
- n variables with small, discrete domains
- m competing constraints

- Random bipartite graph:
- Sparse graph, i.e. m=Θ(n)



“Diluted mean-field spin glasses”

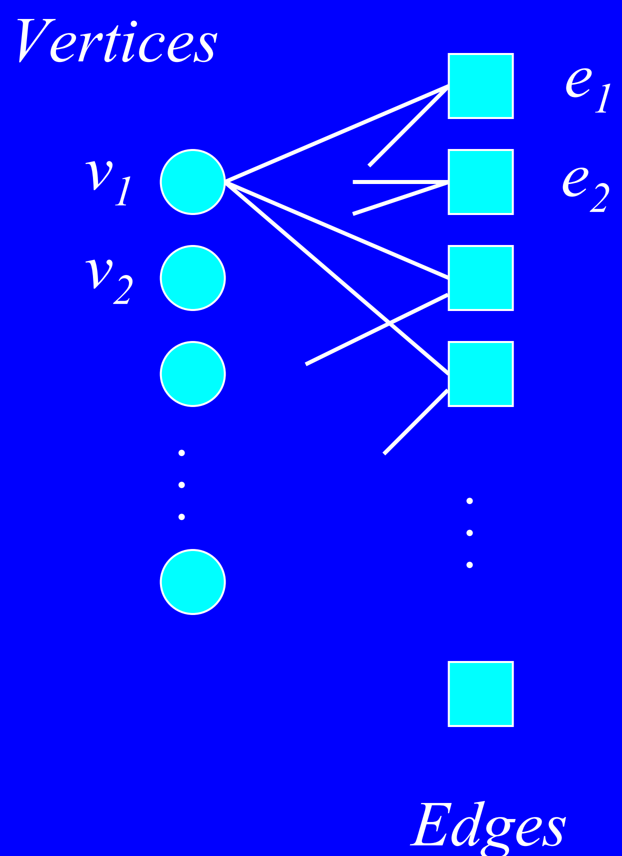
- Small, discrete domains: *spins*
- Conflicting constraints: *quenched disorder*
- Random bipartite graph: *lack of geometry, mean field*
- Hypergraph coloring, random XOR-SAT, error-correcting codes…



Random Graph k-coloring

- Each vertex is a variable with domain {1,2,…,k}
- Each edge is a "not-equal" constraint on two variables

- G(n,m) random graph: the two variables are chosen randomly
- Random r-regular: each variable is chosen r times



Random k-SAT

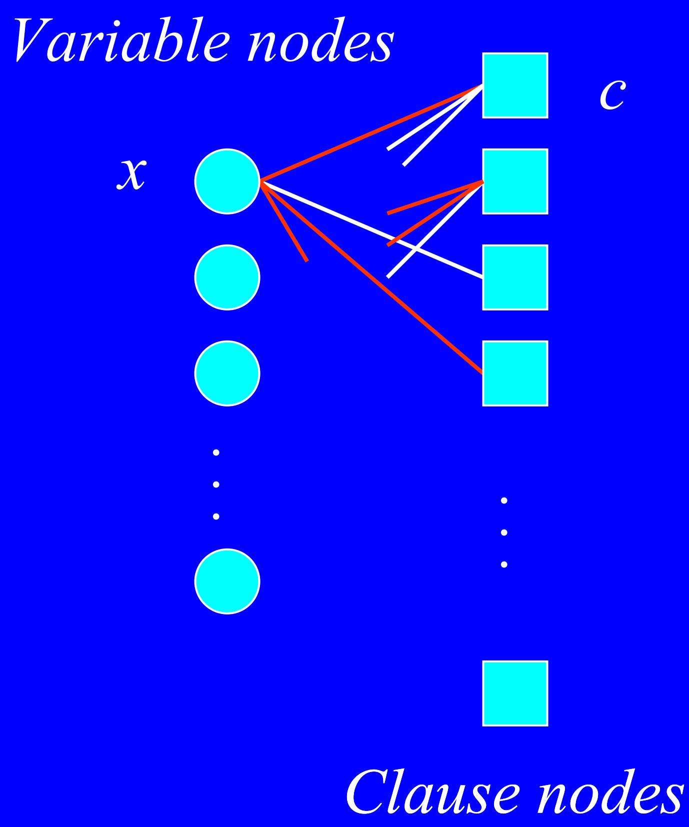
- Take n Boolean variables $X = \{x_1, x_2, \ldots, x_n\}$
- Among all $2^k \binom{n}{k}$ possible k-clauses select m uniformly and independently. Typically $m = rn$.
- Example ($k = 3$) :

$$(\overline{x}_{12} \vee x_5 \vee \overline{x}_9) \wedge (x_{34} \vee \overline{x}_{21} \vee x_5) \wedge \cdots\cdots \wedge (x_{21} \vee x_9 \vee \overline{x}_{13})$$

Random k-SAT

- Variables are binary.
- Every constraint (k-clause) binds k variables.
- Forbids exactly one of the 2^k possible joint values.

- Random k-SAT = each clause picks k random literals.



Similarly: NAE k-SAT, hypergraph 2-coloring, XOR-SAT…

Talk outline

- Part I
  - When do solutions exist?
  - When can known algorithms find them?
- Part II
  - Physics model of solution-space geometry
  - Rigorous results
- Part III
  - Algorithmic implications
  - Survey Propagation

Two Values

Theorem. For every $d > 0$, w.h.p. the chromatic number of $G(n, p = d/n)$

is either k or $k+1$

where k is the smallest integer s.t. $d < 2k \log k$.

[A., Naor '04]

Examples

- If $d = 7$, w.h.p. the chromatic number is 4 or 5 .

- If $d = 10^{60}$, w.h.p. the chromatic number is

$$377145549067226075809014239493833600551612641764765068157\,5$$

or

$$377145549067226075809014239493833600551612641764765068157\,6$$

A simple k-coloring algorithm

- Repeat
  - Pick a random uncolored vertex
  - Assign it the lowest allowed number (color)

Works when $d \leq k \log k$

[Bollobás, Thomasson 84]
[McDiarmid 84]

- NOTHING is known to do better…

The satisfiability threshold conjecture

Conjecture: for every $k \geq 3$, there is r_k such that

$$\lim_{n \to \infty} \Pr[\mathcal{F}_k(n, rn) \text{ is satisfiable}] = \begin{cases} 1 & \text{if } r = r_k - \epsilon \\ 0 & \text{if } r = r_k + \epsilon \end{cases}$$

Since the 80s: for every $k \geq 3$,

$$c \; \frac{2^k}{k} < r_k < 2^k \ln 2$$

[Chvátal & Reed 92]

[Frieze & Suen 96]

Bounds for the k-SAT threshold

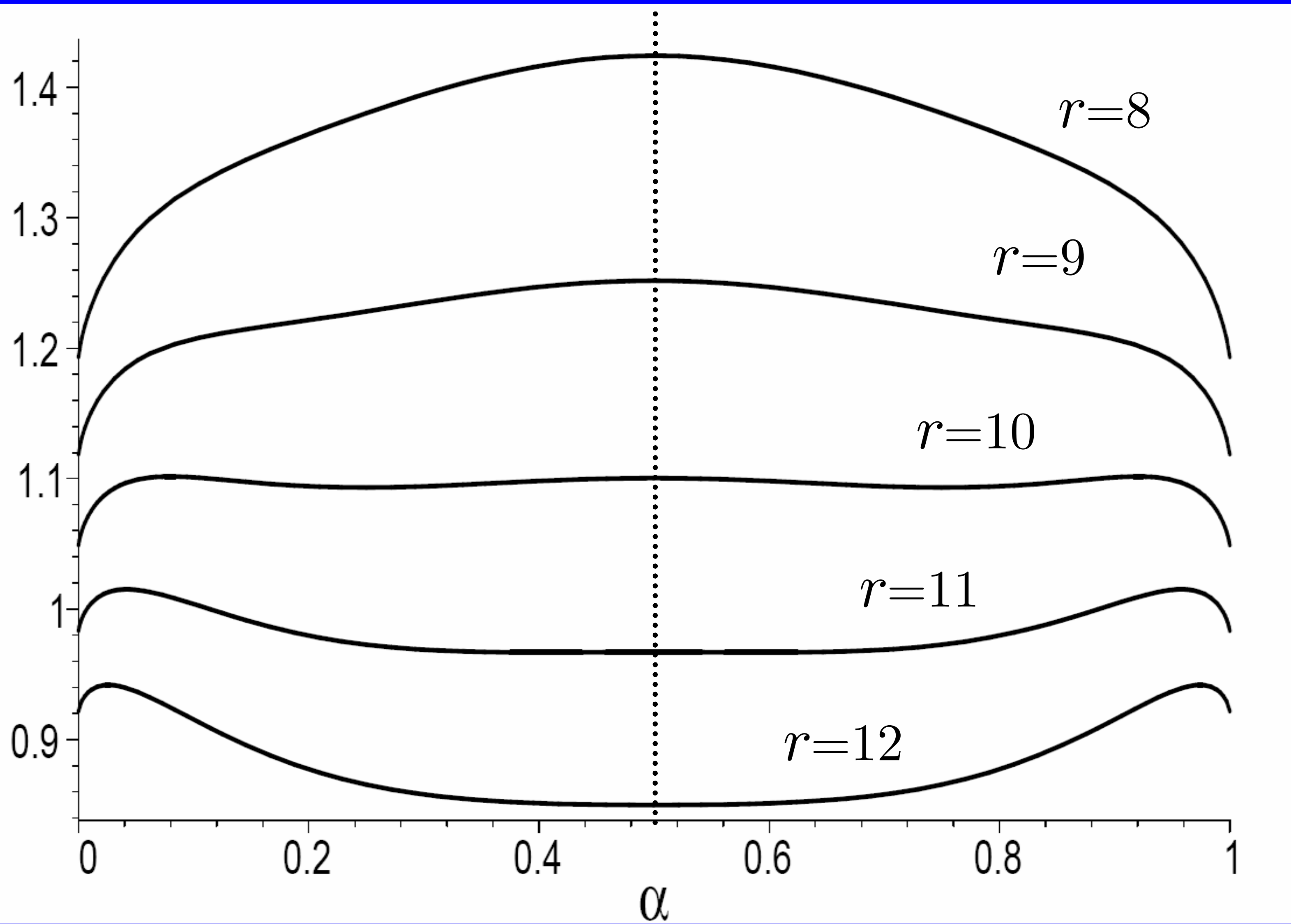
[A., Peres '04]

For all $k \geq 3$:

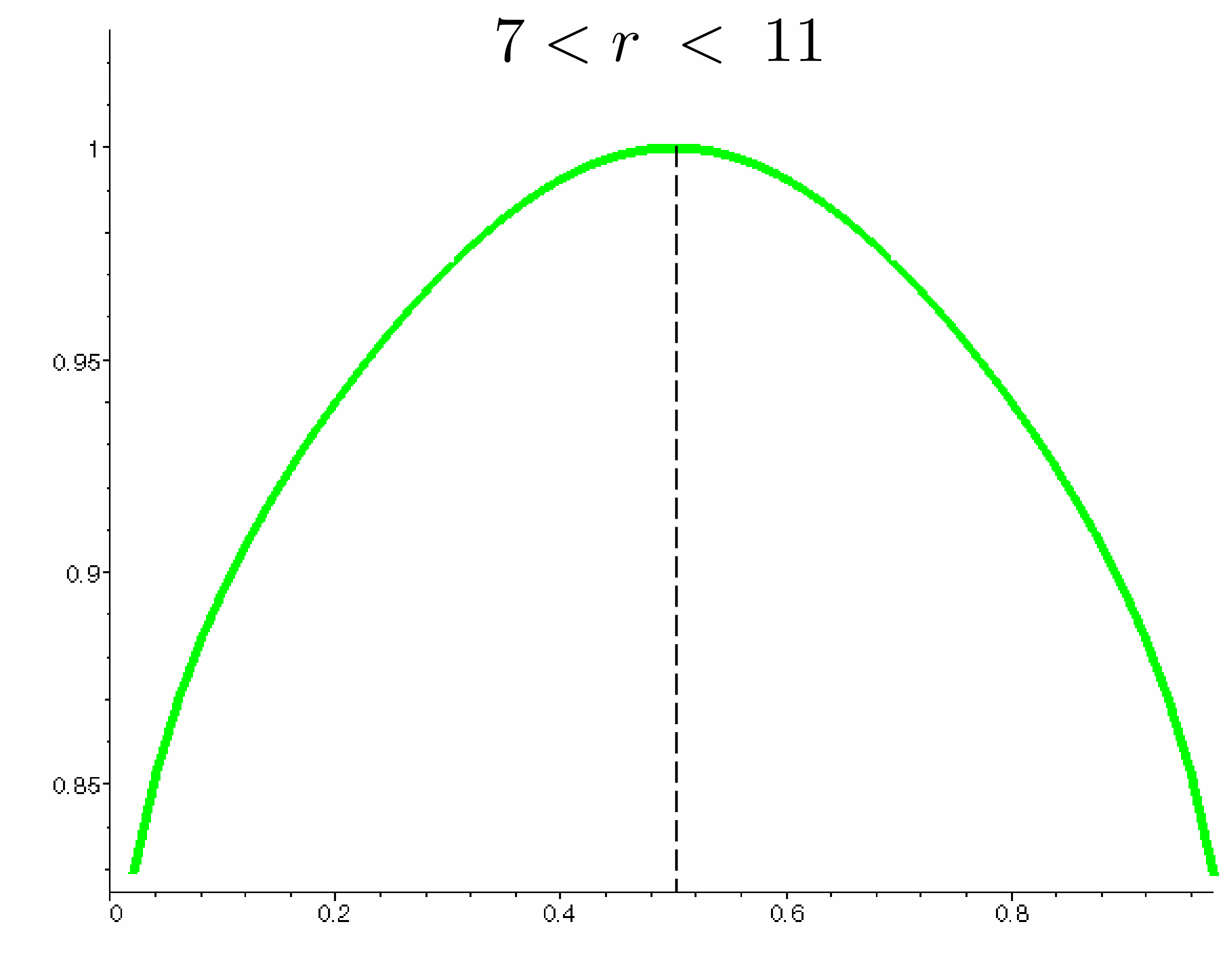
$$2^k \ln 2 - k < r_k < 2^k \ln 2$$

k	3	4	5	7	10	20	21
Upper bound	4.51	10.23	21.33	87.88	708.94	726,817	1,453,635
Lower bound	3.52	7.91	18.79	84.82	704.94	726,809	1,453,626
Best algorithm	3.52	5.54	9.63	33.23	172.65	95,263	181,453

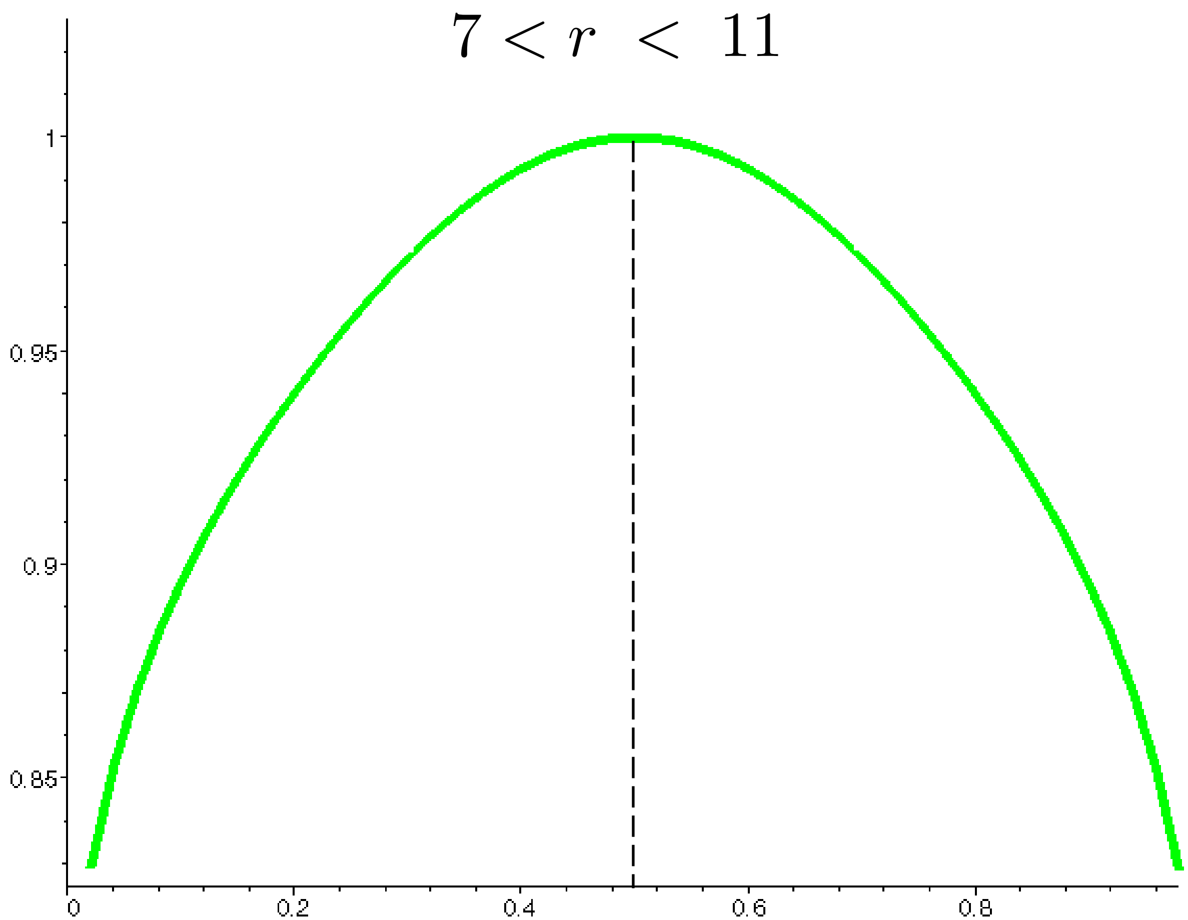
5-uniform hypergraphs



5-uniform hypergraphs



5-uniform hypergraphs



Natural question

Are there efficient algorithms
that work closer
to each problem's threshold?

Our Best Algorithms are Naive

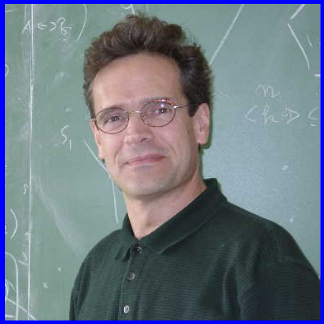
- Repeat
  - Pick a random uncolored vertex
  - Assign it the lowest available color

- Repeat
  - Pick a random variable and set it randomly
  - Satisfy 1-clauses if they exist (repeatedly)

Talk outline

- Part I
  - When do solutions exist?
  - When can known algorithms find them?
- Part II
  - Physics model of solution-space geometry
  - Rigorous results
- Part III
  - Algorithmic implications
  - Survey Propagation

In a parallel universe
(across the Atlantic)

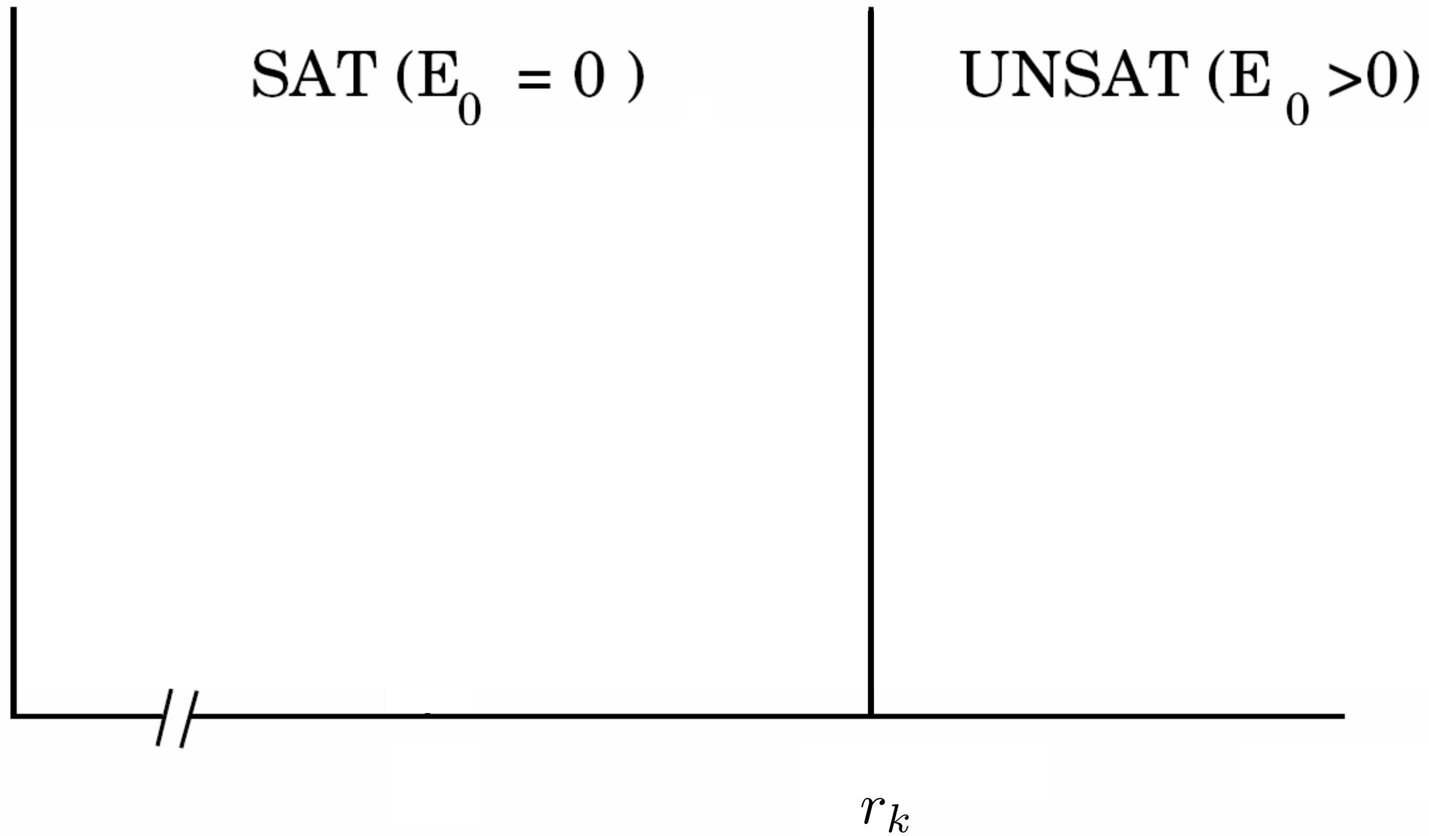


Marc Mezárd

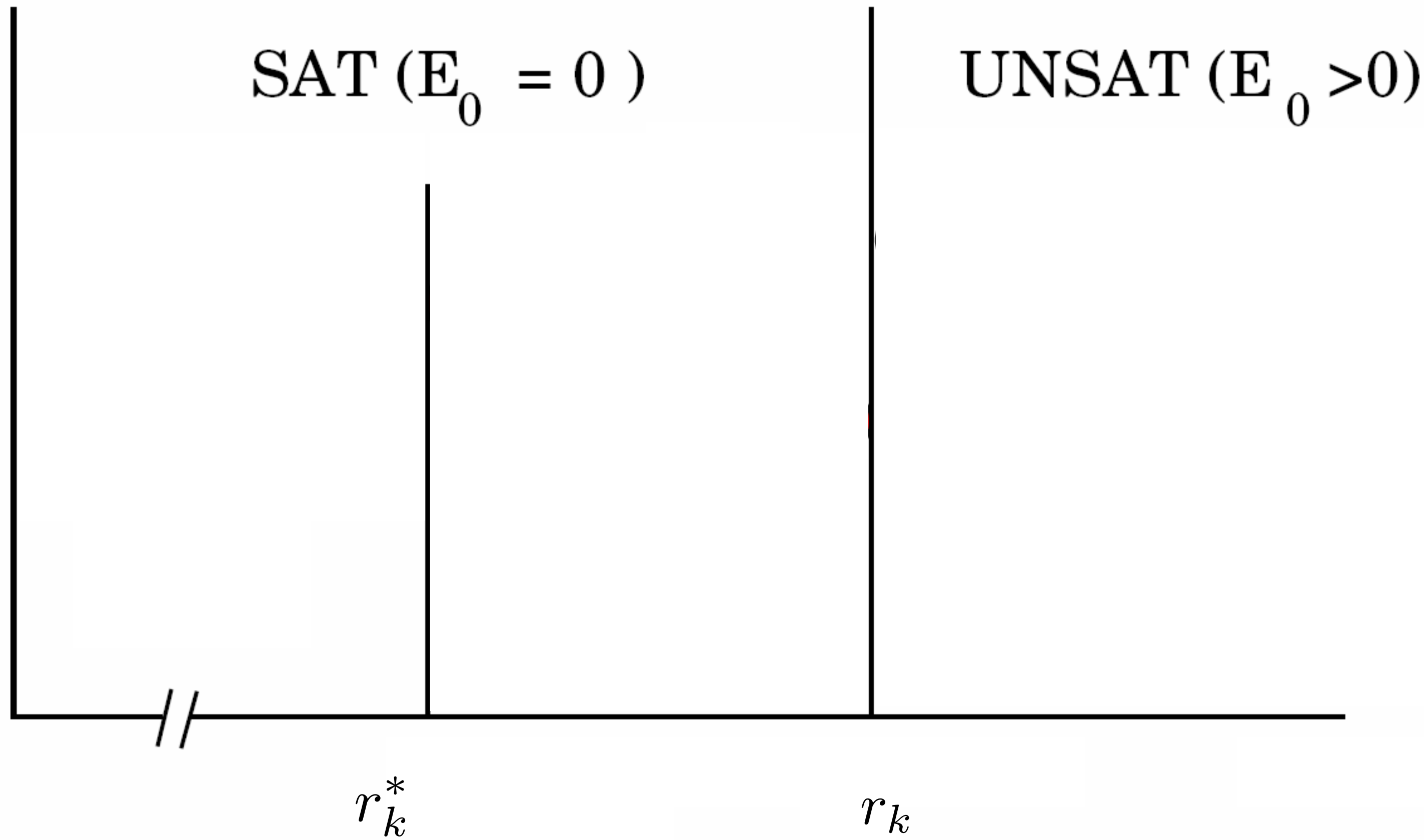
Giorgio Parisi

Riccardo Zecchina

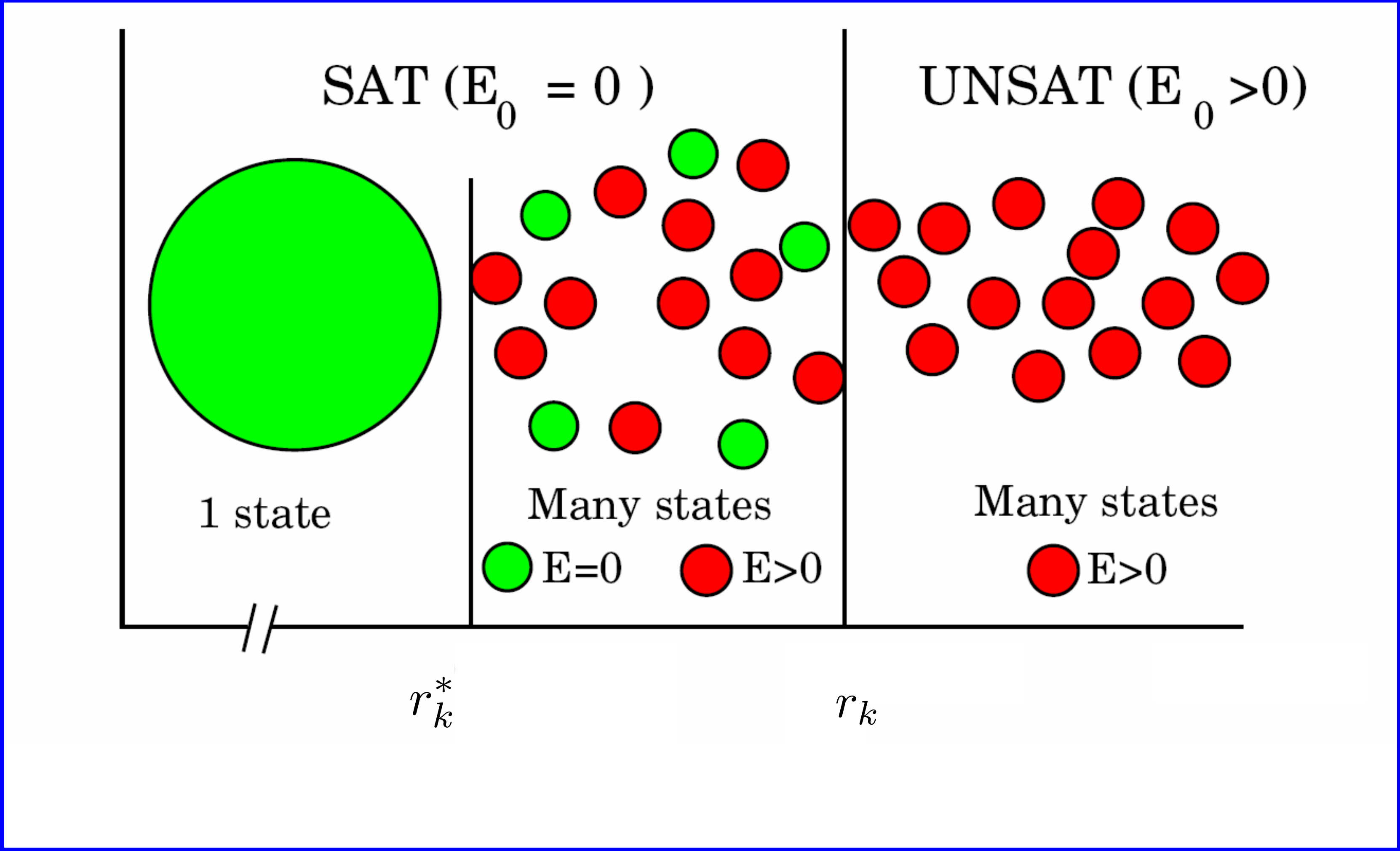
Phycisists say…



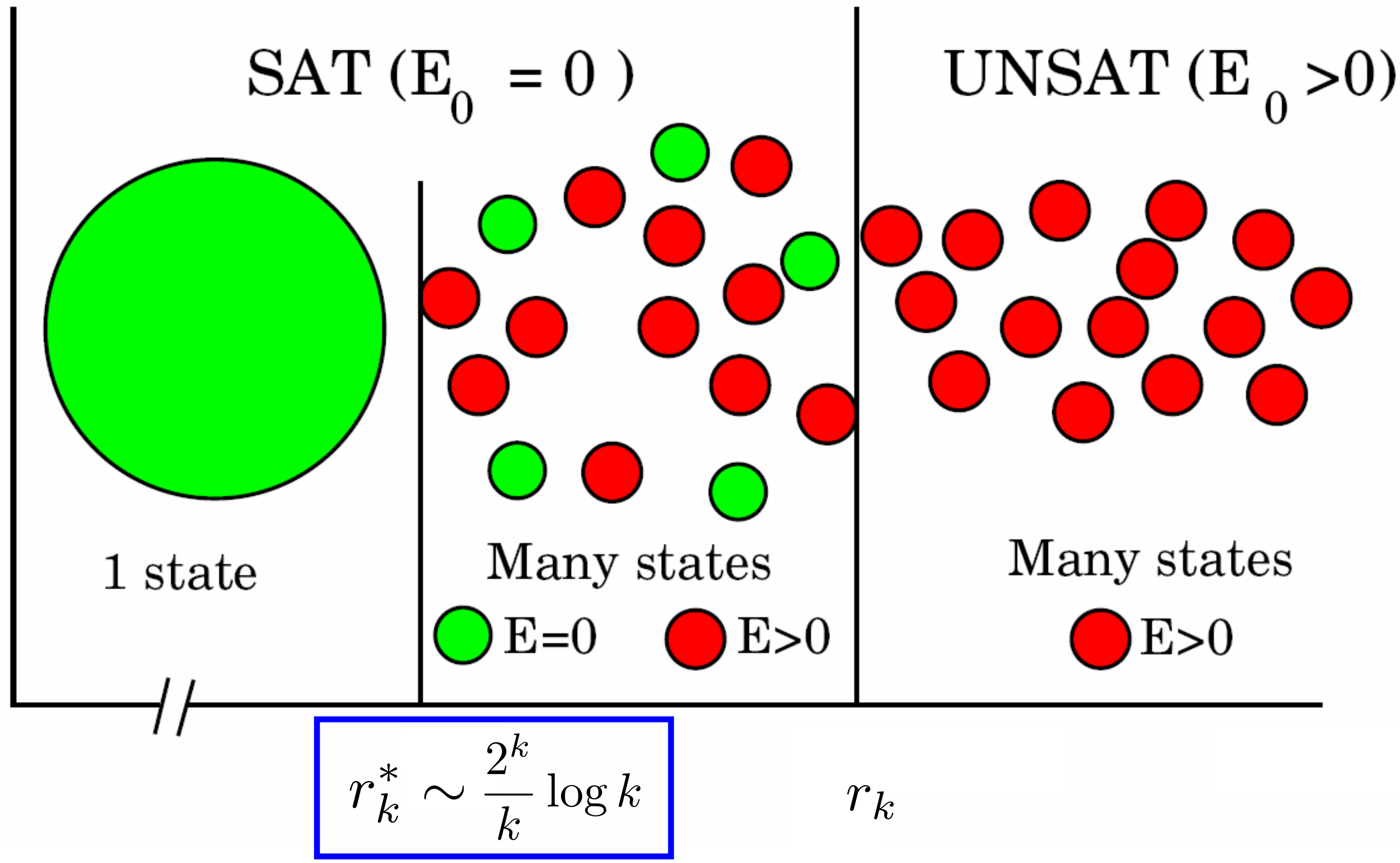
Phycisists say…



Phycisists say…



Phycisists say...



Clusters do exist!

For all $r > 2^{k-1} \ln 2$ we can prove:

- Exponentially many
- Far apart from one another
- Small diameter
- Most variables are frozen

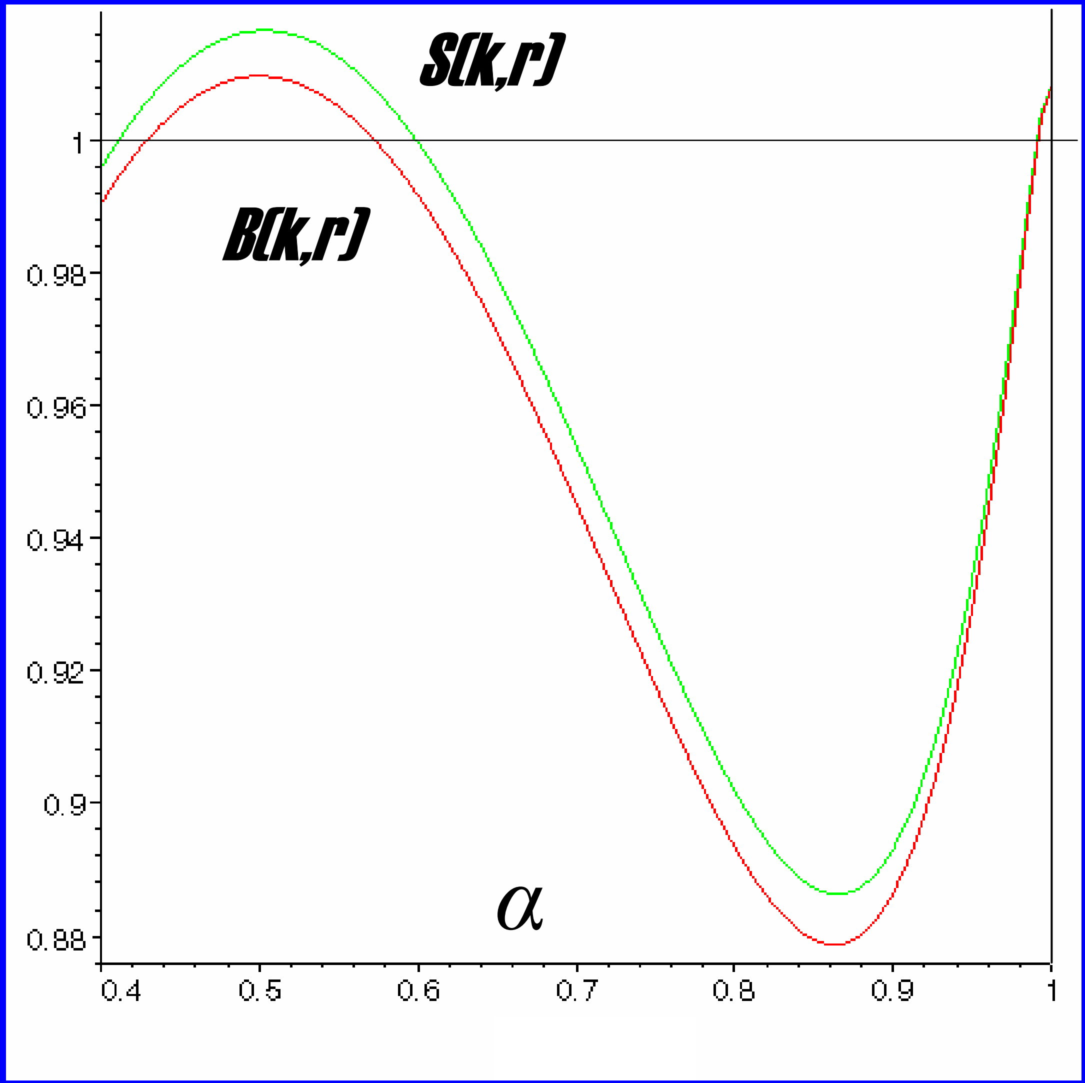
Random 10-SAT, r=700

The expected number of pairs of satisfying assignments having overlap αn is

$S(k,r)^n$ x poly(n)

The expected number of pairs of balanced sat. assignments having overlap αn is

$B(k,r)^n$ x poly(n)



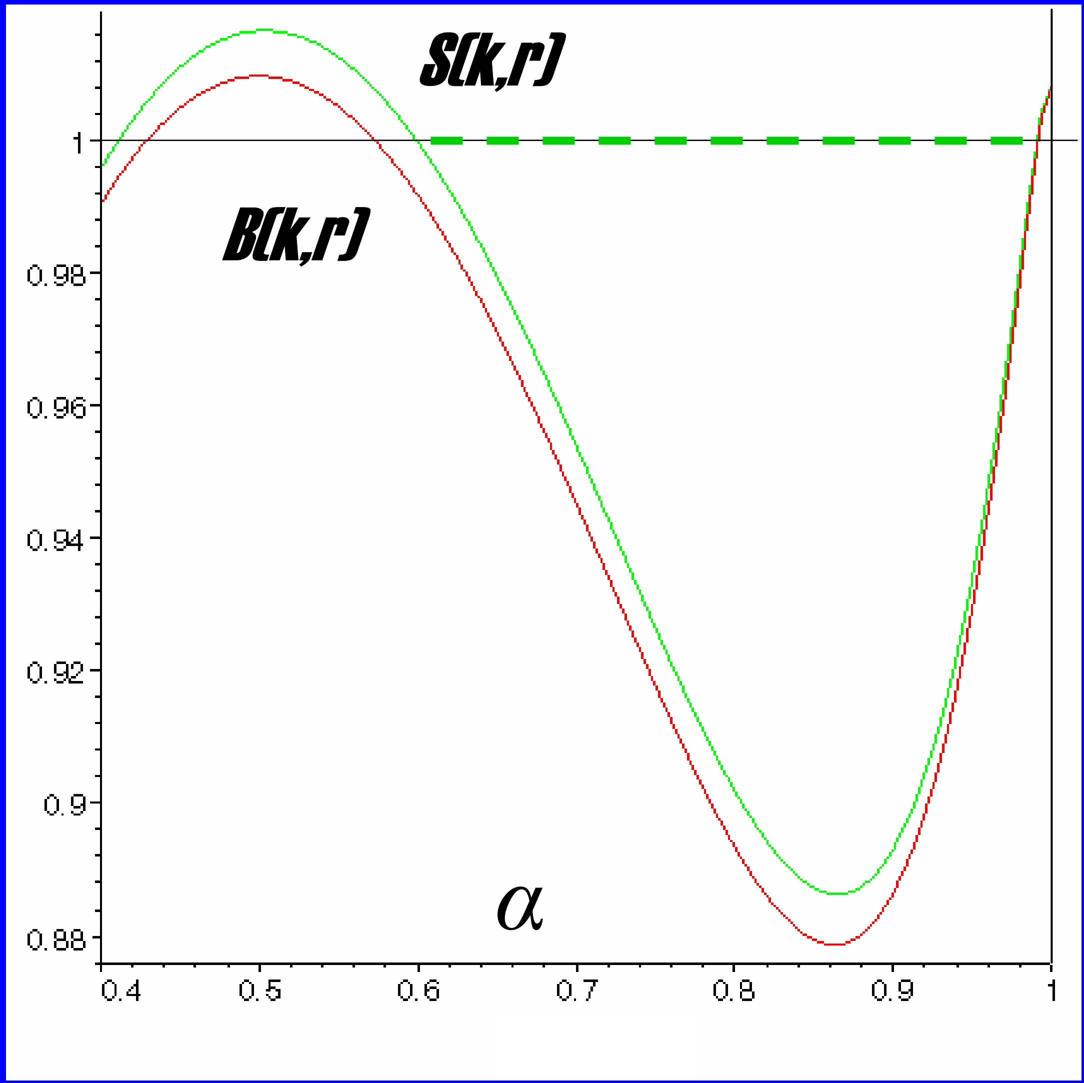
Random 10-SAT, r=700

The expected number of pairs of satisfying assignments having overlap αn is

$S(k,r)^n \times poly(n)$

The expected number of pairs of balanced sat. assignments having overlap αn is

$B(k,r)^n \times poly(n)$



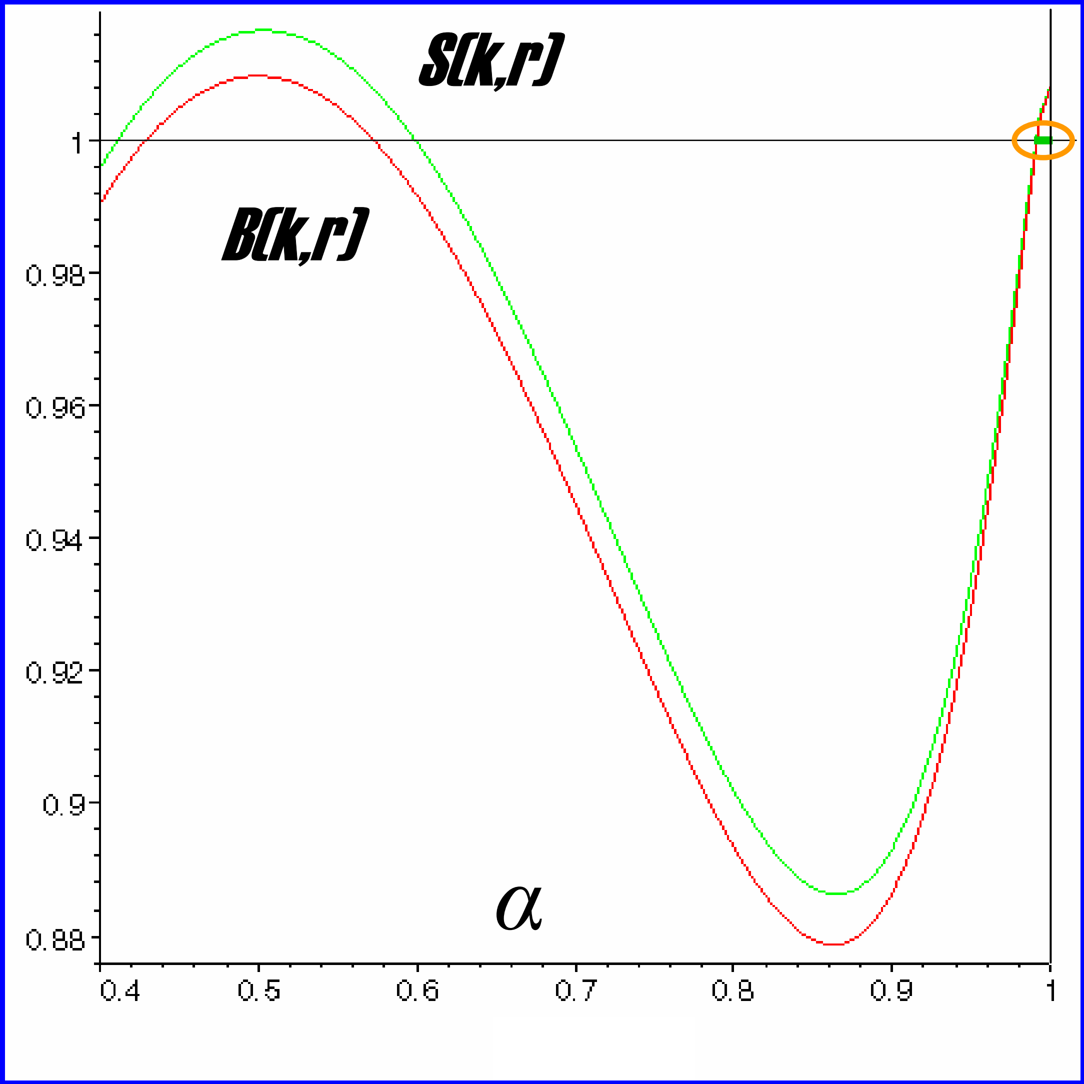
Random 10-SAT, r=700

The expected number of pairs of satisfying assignments having overlap αn is

$S(k,r)^n$ x poly(n)

The expected number of pairs of balanced sat. assignments having overlap αn is

$B(k,r)^n$ x poly(n)



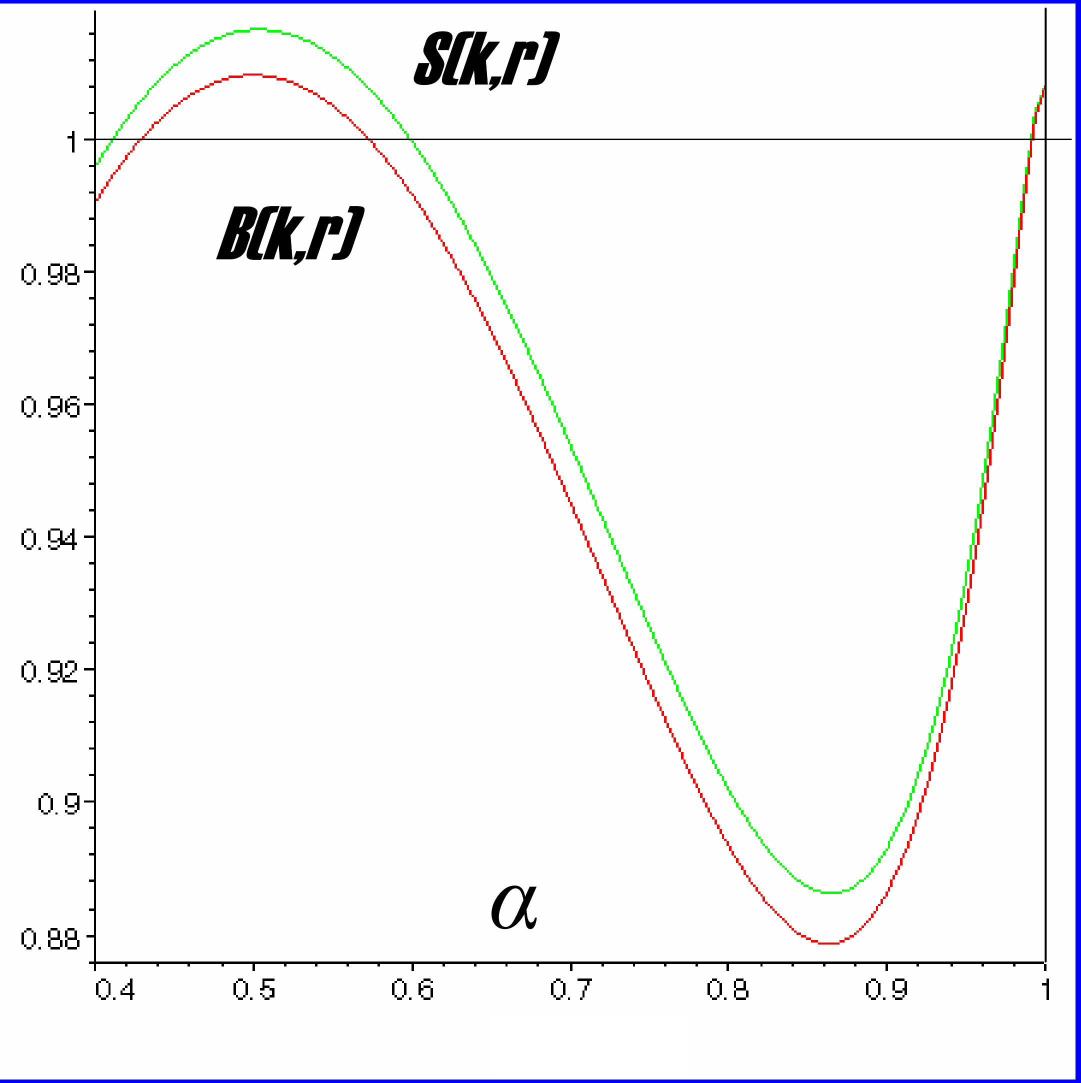
Random 10-SAT, r=700

The expected number of pairs of satisfying assignments having overlap αn is

$S(k,r)^n$ x poly(n)

The expected number of pairs of balanced sat. assignments having overlap αn is

$B(k,r)^n$ x poly(n)



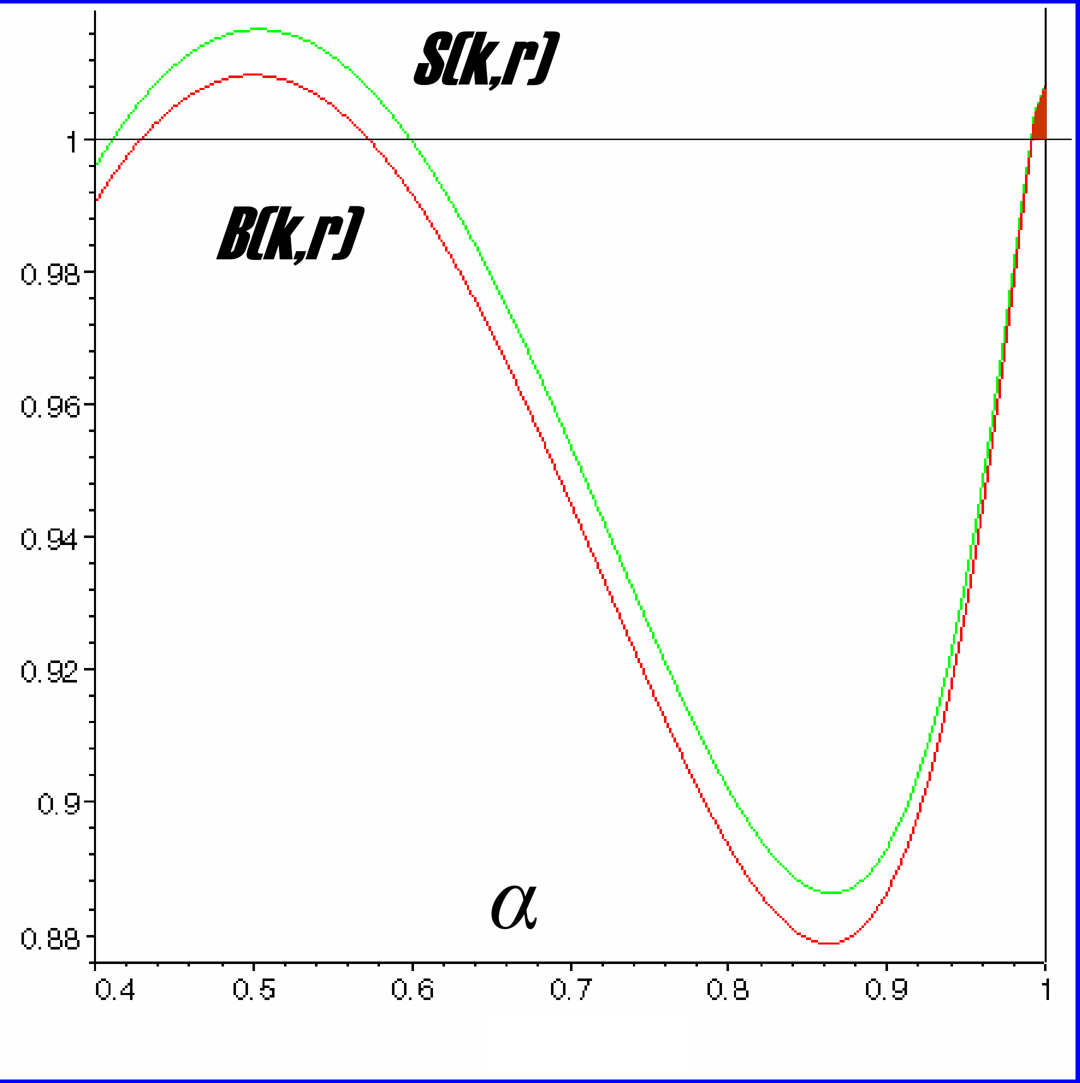
Random 10-SAT, r=700

The expected number of pairs of satisfying assignments having overlap αn is

$S(k,r)^n$ x poly(n)

The expected number of pairs of balanced sat. assignments having overlap αn is

$B(k,r)^n$ x poly(n)



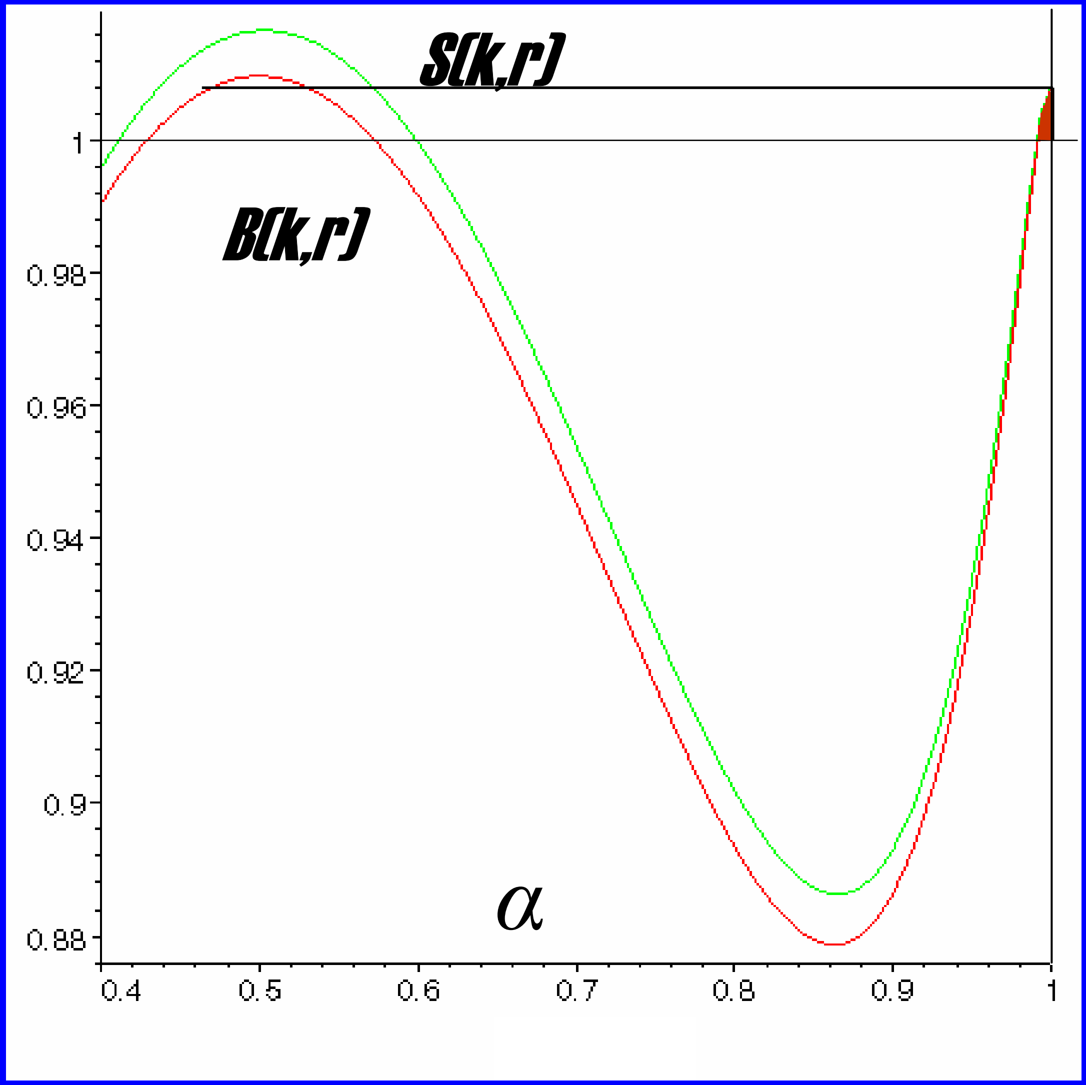
Random 10-SAT, r=700

The expected number of pairs of satisfying assignments having overlap αn is

$S(k,r)^n$ x poly(n)

The expected number of pairs of balanced sat. assignments having overlap αn is

$B(k,r)^n$ x poly(n)



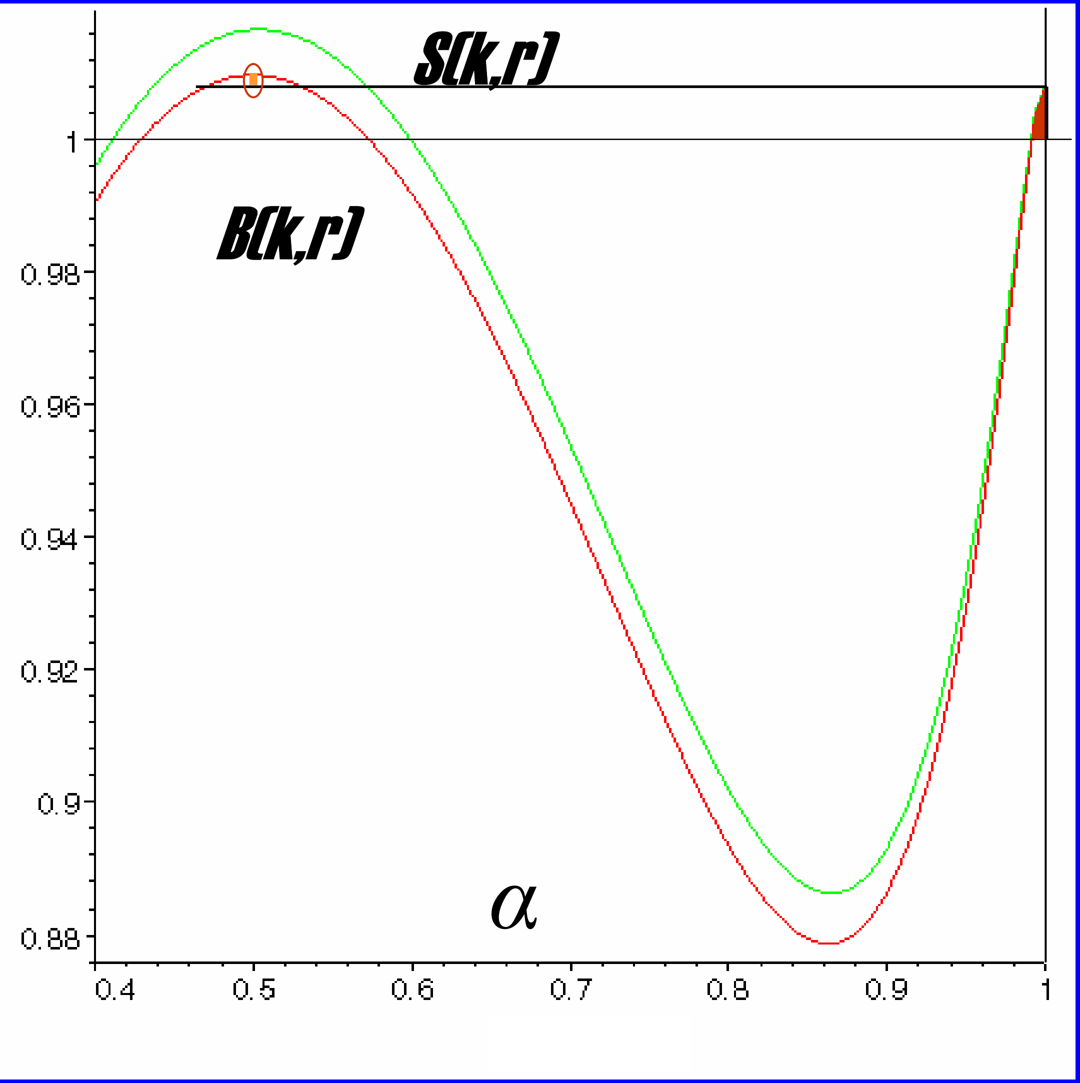
Random 10-SAT, r=700

The expected number of pairs of satisfying assignments having overlap αn is

$S(k,r)^n$ x poly(n)

The expected number of pairs of balanced sat. assignments having overlap αn is

$B(k,r)^n$ x poly(n)



Looking Inside
(Main Result)

Physics prediction: clusters have frozen variables

Looking Inside
(Main Result)

Physics prediction: clusters have frozen variables

Theorem. *For every* $k \geq 9$ *and*

$$r > c_k = \frac{4}{5}\, 2^k \ln 2\, (1 + o(1)),$$

w.h.p. in every cluster the majority of variables are frozen.

Nearly everything freezes

Theorem. *For every* $\epsilon > 0$ *and all* $k \geq k_0(\epsilon)$, *there exists* $c_k^\epsilon < r_k$, *such that w.h.p. in every cluster at least* $(1-\epsilon) \cdot n$ *variables are frozen.*

Nearly everything freezes

Theorem. *For every* $\epsilon > 0$ *and all* $k \geq k_0(\epsilon)$, *there exists* $c_k^\epsilon < r_k$, *such that w.h.p. in every cluster at least* $(1-\epsilon) \cdot n$ *variables are frozen.*

As k *grows,*

$$\frac{c_k^\epsilon}{2^k \ln 2} \to \frac{1}{1+\epsilon(1-\epsilon)}$$

Talk outline

- Part I
  - When do solutions exist?
  - When can known algorithms find them?
- Part II
  - Physics model of solution-space geometry
  - Rigorous results
- Part III
  - Algorithmic implications
  - Survey Propagation

Sampling satisfying assignments

(thought experiment)

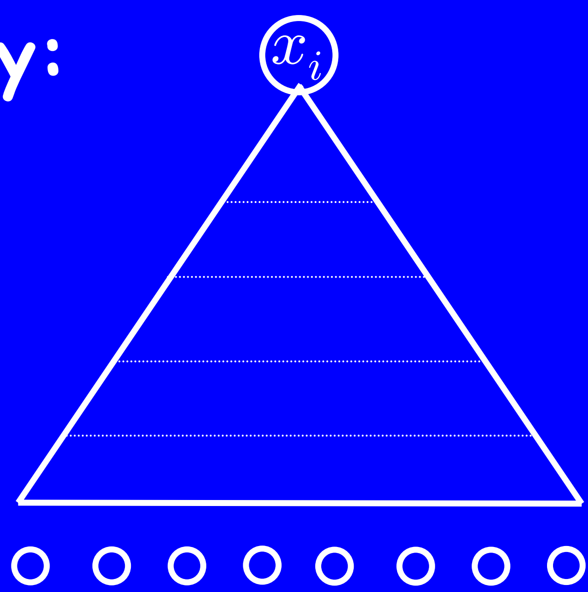
- Approximate the fraction p_i of satisfying truth assignments in which variable x_i takes value 1.
- Set x_i to 1 with probability p_i and simplify.

Sampling satisfying assignments

(thought experiment)

- Approximate the fraction p_i of satisfying truth assignments in which variable x_i takes value 1.
- Set x_i to 1 with probability p_i and simplify.

Locally:

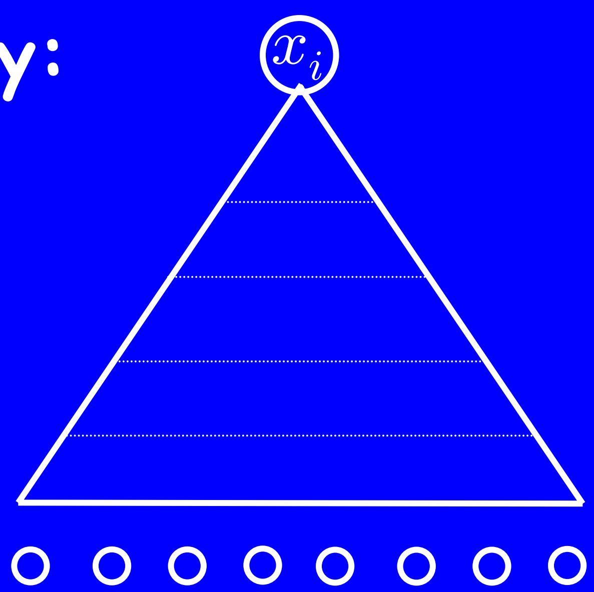


Sampling satisfying assignments

(thought experiment)

- Approximate the fraction p_i of satisfying truth assignments in which variable x_i takes value 1.
- Set x_i to 1 with probability p_i and simplify.

Locally:



Given boundary Λ:

compute p_Λ

$$\boxed{p_i = \sum_{\Lambda} p_\Lambda \times \mathrm{Ext}(\Lambda)}$$

Hope

- The variables in the boundary of the tree are probably "far apart" (if we remove the tree).
- Therefore, they should be uncorrelated, in which case (for $k>2$) "we can cope".

e.g., LDPC codes

Hope

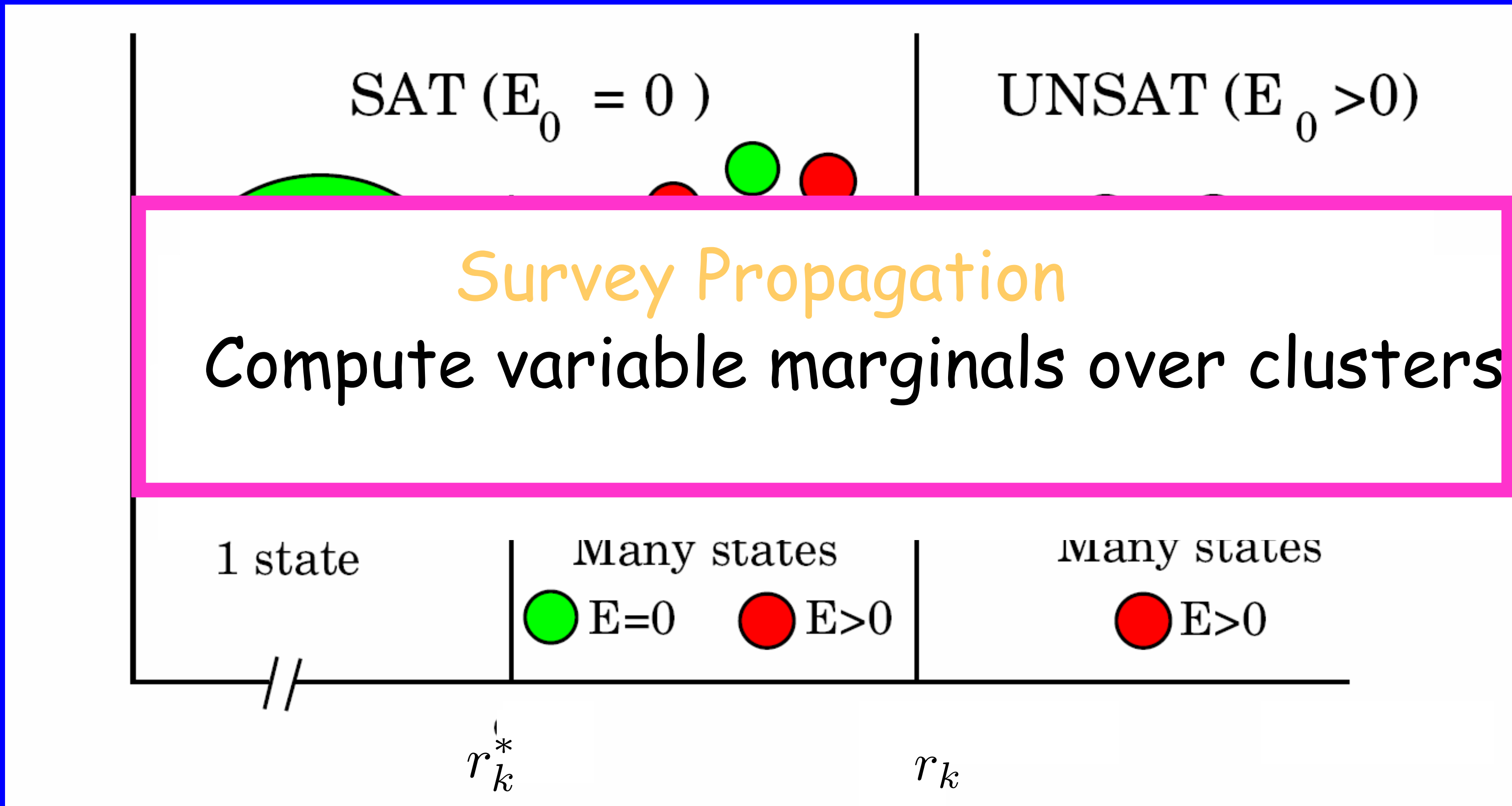
- The variables in the boundary of the tree are probably "far apart" (if we remove the tree).
- Therefore, they should be uncorrelated, in which case (for k>2) "we can cope".

e.g., LDPC codes

Clusters: few or many?

- The marginals might NOT be uncorrelated.
- Clusters with many frozen variables may induce "long-range" correlations.

Frozen Variables -> Long Range Correlations



Survey Propagation

Compute variable marginals over clusters

Definitions

For any formula F:

-Let $\mathcal{S}(F)$ be the set of satisfying assignments of F.

-Let $C_1, C_2, \ldots$ be the connected components (clusters) of $\mathcal{S}(F)$. (Adjacent = Hamming distance 1)

-Let the label of C be its projection $\ell(C) \in \{0, 1, *\}^n$.

-If $\ell_i(C) \in \{0, 1\}$ we say that x_i is frozen in C.

Two quick observations:

- Labels are "lossless" for cubes.
- The label of C can be "all-stars" already with $|C| = n$.

Surveys

Definition. *A variable* x_i *is* **free** *in* $x \in \{0, 1, *\}^n$ *if in every clause containing* $x_i, \overline{x}_i$ *there is some other satisfied literal or* $*$.

Surveys

Definition. *A variable* x_i *is* **free** *in* $x \in \{0, 1, *\}^n$ *if in every clause containing* $x_i, \overline{x}_i$ *there is some other satisfied literal or* $*$.

Repeat until fixed point: set all free variables to $*$.

More Precisely

Definition. *A variable x_i is* **free** *in $x \in \{0, 1, *\}^n$ if in every clause containing $x_i, \overline{x}_i$ there is some other satisfied literal or $*$.*

Repeat until fixed point: set all free variables to * .

1. All σ in C have the same fixed point, called cover(C).
2. label(C) $\preceq$ cover(C) deterministically.
3. "Being a fixed-point" <-> locally tree-like factor graph G_φ.
4. Attempt to marginalize G_φ by local info.

∪B∇⊣⊓\∫⊔⌉⟩\⇔←Z⌉⌊⌋⟨⟩\⊣← ′△⊎∪M⊣\⌉⊑⊣⇔←
M≀∫∫⌉↕⇔←W⊣⟩\⊒∇⟩}⟨⊔←′△⊎

Surveys

Definition. *A variable* x_i *is* **free** *in* $x \in \{0,1,*\}^n$ *if in every clause containing* $x_i, \overline{x}_i$ *there is some other satisfied literal or* $*$.

Repeat until fixed point: set all free variables to * .

Question: do covers retain useful information?

e.g. are there fixed-points other than “all-*“ ?

Surveys

Definition. *A variable x_i is* **free** *in $x \in \{0, 1, *\}^n$ if in every clause containing $x_i, \overline{x}_i$ there is some other satisfied literal or $*$.*

Repeat until fixed point: set all free variables to $*$.

Question: do covers retain useful information?

e.g. are there fixed-points other than "all-$*$" ?

Answer: Yes. That's how we actually prove the existence of frozen variables.

Proof

- Let X be the number of satisfying assignments whose cover (fixed point) is “all-*”. (Call them “coreless”.)

Proof

- Let X be the number of satisfying assignments whose cover (fixed point) is "all-*". (Call them "coreless".)

Proof

- Let X be the number of satisfying assignments whose cover (fixed point) is "all-*". (Call them "coreless".)

$$
\begin{aligned}
\mathbf{E}[X] &= \sum_{\sigma} \Pr[\sigma \ \textit{is coreless} \mid \sigma \ \textit{is satisfying}] \times \Pr[\sigma \ \textit{is satisfying}] \\
&= 2^n \cdot \left(1 - \frac{1}{2^k}\right)^{rn} \cdot \Pr[\mathbf{0} \ \textit{is coreless} \mid \mathbf{0} \ \textit{is satisfying}]
\end{aligned}
$$

Proof

- Let X be the number of satisfying assignments whose cover (fixed point) is "all-*". (Call them "coreless".)

$$\begin{aligned}\mathbf{E}[X] &= \sum_{\sigma} \Pr[\sigma \ \textit{is coreless} \mid \sigma \ \textit{is satisfying}] \times \Pr[\sigma \ \textit{is satisfying}] \\ &= 2^n \cdot \left(1 - \frac{1}{2^k}\right)^{rn} \cdot \Pr[\mathbf{0} \ \textit{is coreless} \mid \mathbf{0} \ \textit{is satisfying}]\end{aligned}$$

- Conditioning on "**0** *is satisfying*" is easy
- Relevant clauses = uniquely-satisfied clauses
- Similar to hypergraph core computation

Proof

- Let X be the number of satisfying assignments whose cover (fixed point) is "all-*". (Call them "coreless".)

$$
\begin{aligned}
\mathbf{E}[X] &= \sum_{\sigma} \Pr[\sigma \ \mathit{is\ coreless} \mid \sigma \ \mathit{is\ satisfying}] \times \Pr[\sigma \ \mathit{is\ satisfying}] \\
&= 2^n \cdot \left(1 - \frac{1}{2^k}\right)^{rn} \cdot \Pr[\mathbf{0} \ \mathit{is\ coreless} \mid \mathbf{0} \ \mathit{is\ satisfying}] \\
&< \left[2 \cdot \left(1 - \frac{1}{2^k}\right)^{r} \cdot e^{-f(r)}\right]^n
\end{aligned}
$$

Proof

- Let X be the number of satisfying assignments whose cover (fixed point) is "all-*". (Call them "coreless".)

$$
\begin{aligned}
\mathbf{E}[X] &= \sum_{\sigma} \Pr[\sigma \ \textit{is coreless} \mid \sigma \ \textit{is satisfying}] \times \Pr[\sigma \ \textit{is satisfying}] \\
&= 2^n \cdot \left(1 - \frac{1}{2^k}\right)^{rn} \cdot \Pr[\mathbf{0} \ \textit{is coreless} \mid \mathbf{0} \ \textit{is satisfying}] \\
&< \left[2 \cdot \left(1 - \frac{1}{2^k}\right)^{r} \cdot e^{-f(r)}\right]^n
\end{aligned}
$$

$$
\Pr[\mathbf{0} \text{ is coreless} \mid \mathbf{0} \text{ is satisfying}] = \begin{cases} 1 - o(1) & \text{if } r < t_k \\ o(1) & \text{if } r > t_k \end{cases}
$$

Proof

- Let X be the number of satisfying assignments whose cover (fixed point) is "all-*". (Call them "coreless".)

$$
\begin{aligned}
\mathbf{E}[X] &= \sum_{\sigma} \Pr[\sigma \textit{ is coreless} \mid \sigma \textit{ is satisfying}] \times \Pr[\sigma \textit{ is satisfying}] \\
&= 2^n \cdot \left(1 - \frac{1}{2^k}\right)^{rn} \cdot \Pr[\mathbf{0} \textit{ is coreless} \mid \mathbf{0} \textit{ is satisfying}] \\
&< \left[2 \cdot \left(1 - \frac{1}{2^k}\right)^{r} \cdot e^{-f(r)}\right]^n
\end{aligned}
$$

$$t_k \sim \frac{2^k}{k} \log k$$

Summary

- Much before disappearing solutions form clusters:
  - Relatively small
  - Far apart
  - Exponentially many
- "Error-correcting-code with fuzz"

Summary

- Much before disappearing solutions form clusters:
  - Relatively small
  - Far apart
  - Exponentially many
- "Error-correcting-code with fuzz"

- Frozen variables -> long range correlations -> cause naive local algorithms to fail.

Summary

- Much before disappearing solutions form clusters:
  - Relatively small
  - Far apart
  - Exponentially many
- "Error-correcting-code with fuzz"

- Frozen variables -> long range correlations -> cause naive local algorithms to fail.
- Physicists say frozen variables are the main source of long range correlations (1-step RSB hypothesis).

Summary

- Much before disappearing solutions form clusters:
  - Relatively small
  - Far apart
  - Exponentially many
- "Error-correcting-code with fuzz"

- Frozen variables -> long range correlations -> cause naive local algorithms to fail.
- Physicists say frozen variables are the main source of long range correlations (1-step RSB hypothesis).
- Indeed, cover approximation is good.

Summary

- Much before disappearing solutions form clusters:
  - Relatively small
  - Far apart
  - Exponentially many
- "Error-correcting-code with fuzz"

- Frozen variables -> long range correlations -> cause naive local algorithms to fail.
- Physicists say frozen variables are the main source of long range correlations (1-step RSB hypothesis).
- Indeed, cover approximation is rigorously good.

- Survey Propagation works extremely well in practice

Summary

- Much before disappearing solutions form clusters:
  - Relatively small
  - Far apart
  - E
- “Er

Thank you!

- Fro
  naive local algorithms to fail.
- Physicists say frozen variables are the main source of long range correlations (1-step RSB hypothesis).
- Indeed, cover approximation is rigorously good.

- Survey Propagation works extremely well in practice