Phase transitions and computational complexity

a physicist point of view

Federico Ricci-Tersenghi
Physics Department
Sapienza University, Roma

in collaboration over the years with
D. Achlioptas, A. Coja-Oghlan, F. Krzakala, M. Mézard, A. Montanari,
G. Parisi, G. Semerjian, M. Weigt, R. Zecchina, L. Zdeborova

Question

What makes a random constraint satisfaction problem hard to solve?

1 million dollars question ;-)
(P vs NP)

Answer

The structure of the solutions space

- Random CSP undergo phase transitions, that change drastically the solution space (proved)

- Connect behavior of solving algorithms to the structure of the solution space (first results...)

Random CSP

- random q-col

  q-coloring a random graph with N vertices and M links

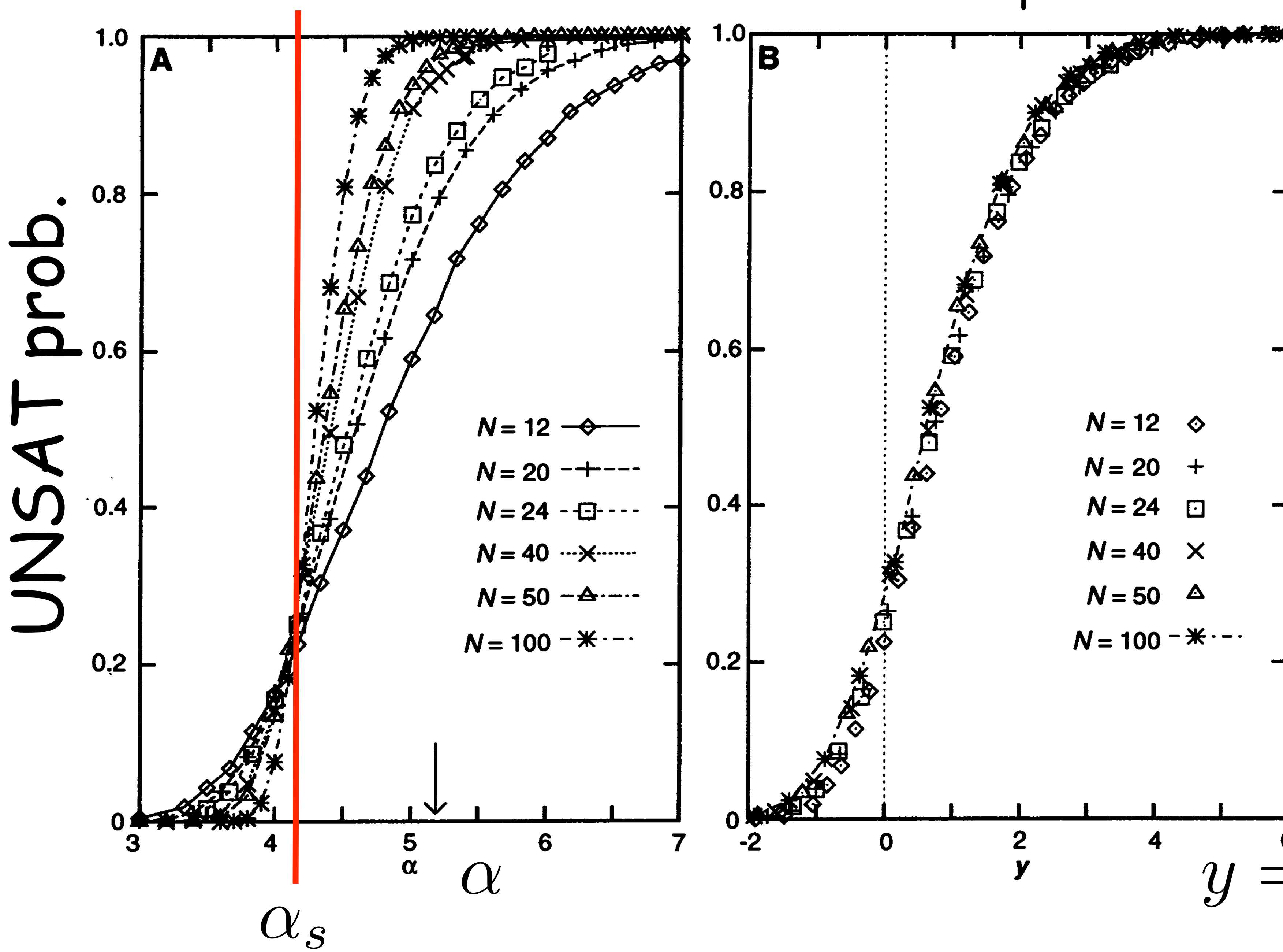
- random K-SAT

  N Boolean variables and M randomly generated clauses (constraints) of fixed length K

$$\boxed{\alpha = M/N}$$

SAT/UNSAT phase transition

Kirkpatrick & Selman, Science '94



random 3-SAT

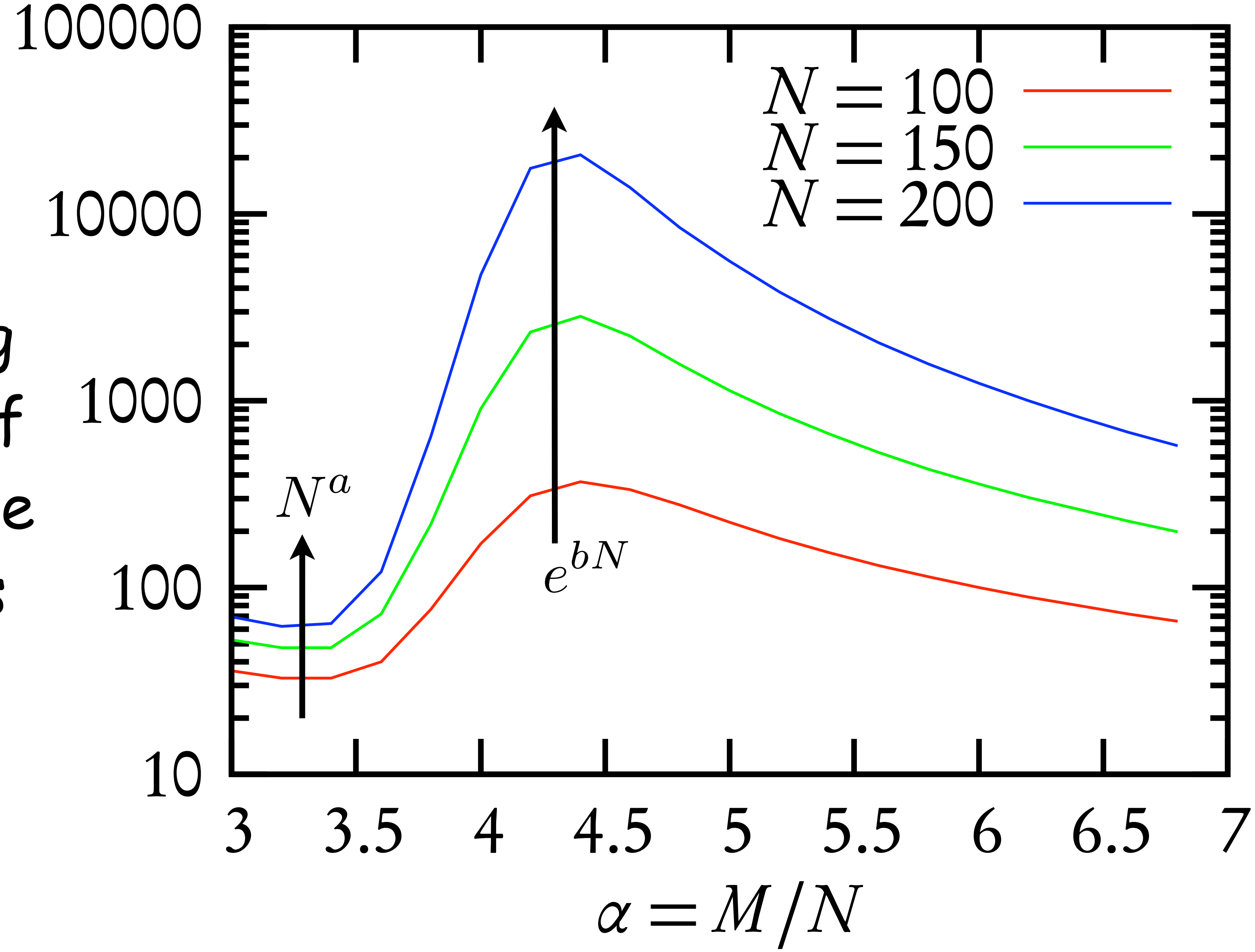
$$\alpha_s \sim 4.17$$

$$\nu \sim 1.5$$

$$y = N^{1/\nu}(\alpha - \alpha_s)$$

Connection to computational complexity

Running times of complete solvers (DPLL)



Rigorous results

- Friedgut ('99): For any K there exist a sequence $\alpha_s(N)$ such that for $N \to \infty$

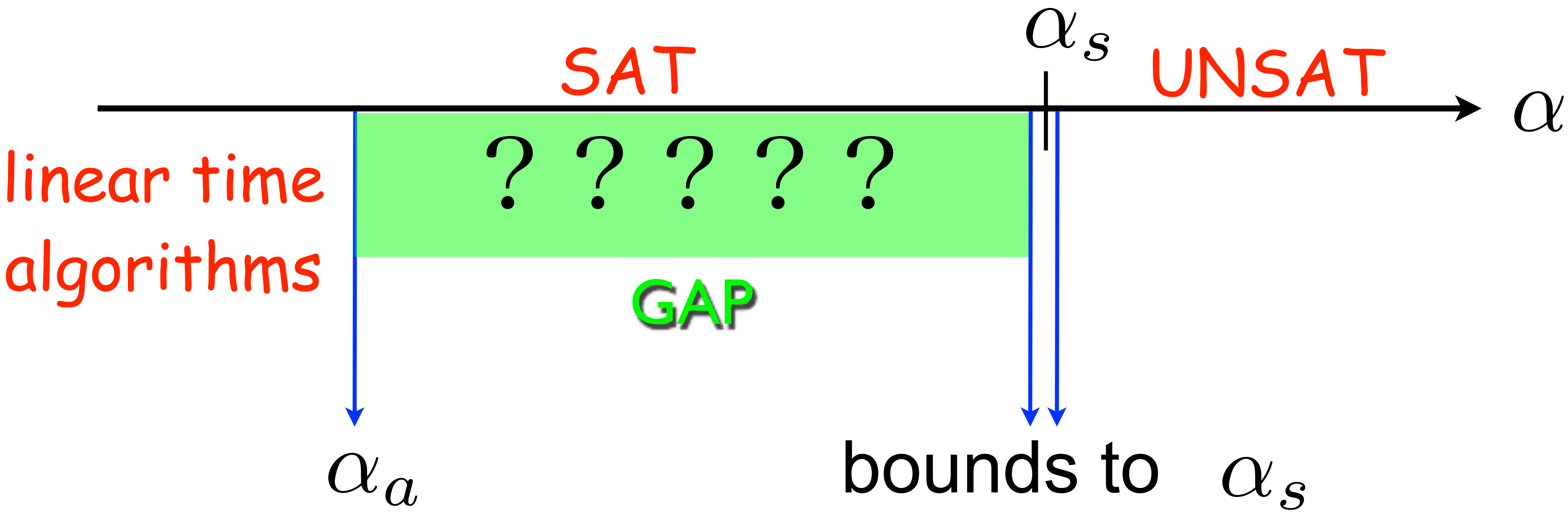
$$P_{\mathrm{SAT}}\big(M/N = \alpha_s(N) - \varepsilon\big) \to 1 \qquad \forall \varepsilon > 0$$
$$P_{\mathrm{SAT}}\big(M/N = \alpha_s(N) + \varepsilon\big) \to 0$$

  Numerically $\alpha_s(N) \to \alpha_s$
  Rigorously only bounds to α_s are known.

- All provably linear time convergent algorithms stop working at some α_a, well before α_s
  E.g. for large K

$$\alpha_a \leq \frac{\ln K}{K} 2^K \qquad \alpha_s \simeq 2^K$$

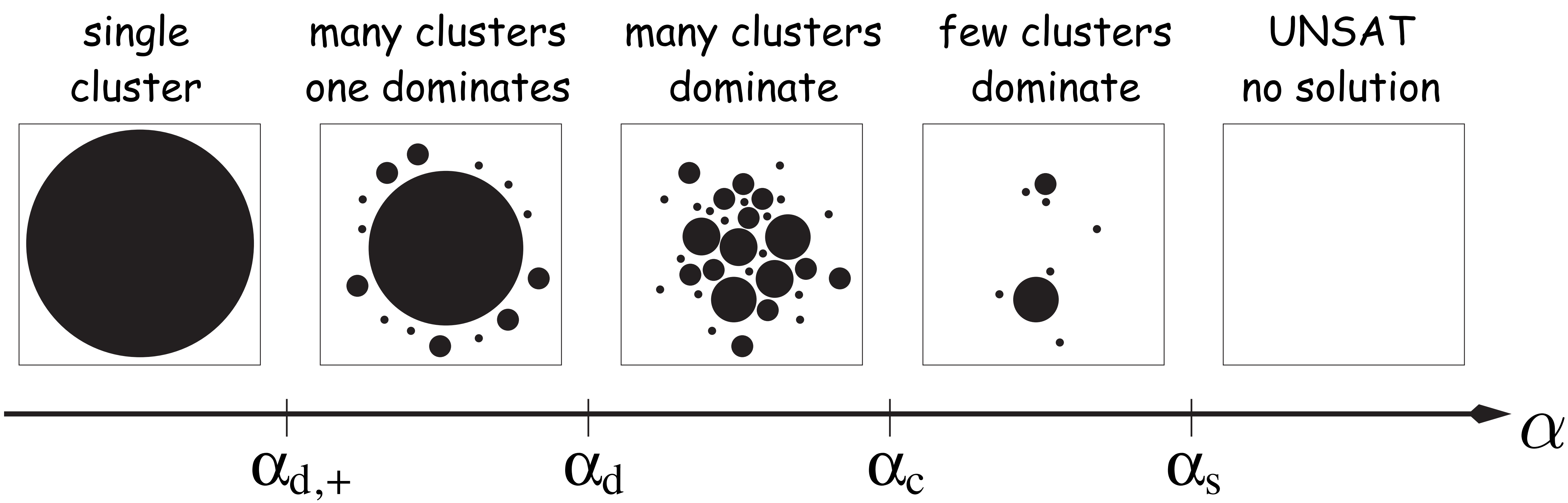
A big gap!



K	α_a	α_s
10	172.65	707 ± 2
20	95263	726813 ± 4

Solution space structure

Krzakala, Montanari, Ricci-Tersenghi, Semerjian, Zdeborova, PNAS '07



k	α_d	α_c	α_s
3	3.86	3.86	4.267
4	9.38	9.547	9.931
5	19.16	20.80	21.117
6	36.53	43.08	43.37

the largest for large K

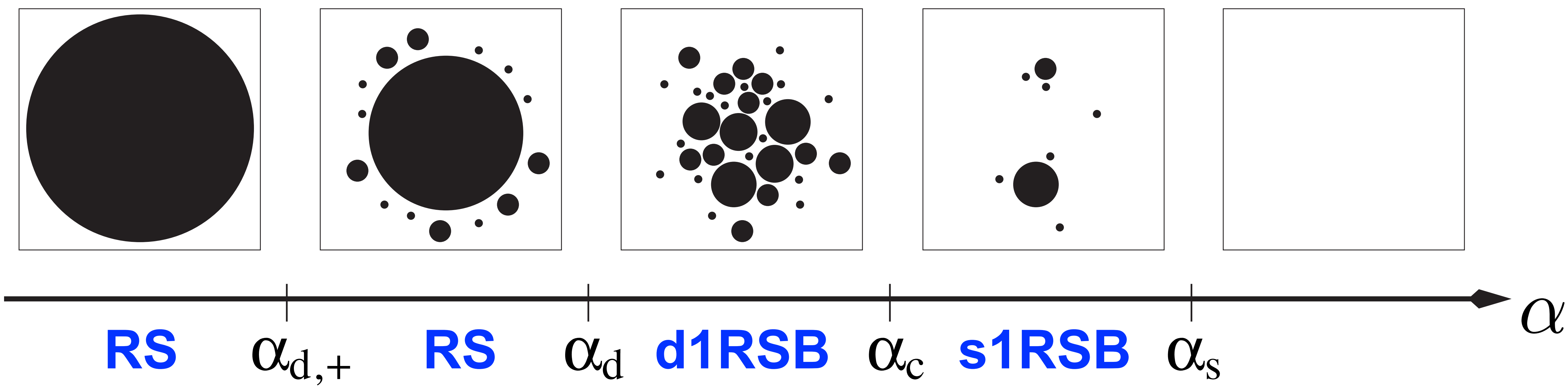
$$\alpha_d \sim \frac{\log(K)}{K} 2^K$$

$$\alpha_c \sim \alpha_s \sim 2^K$$

Solution space structure

Krzakala, Montanari, Ricci-Tersenghi, Semerjian, Zdeborova, PNAS '07

single cluster | many clusters one dominates | many clusters dominate | few clusters dominate | UNSAT no solution



k	α_d	α_c	α_s
3	3.86	3.86	4.267
4	9.38	9.547	9.931
5	19.16	20.80	21.117
6	36.53	43.08	43.37

the largest for large K

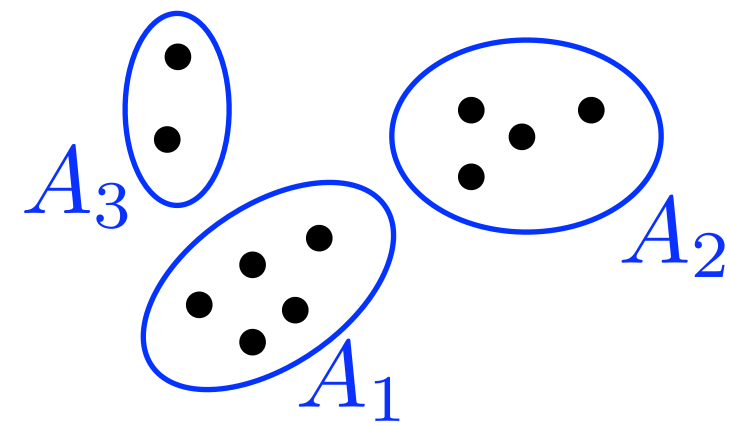
$$\alpha_d \sim \frac{\log(K)}{K} 2^K$$

$$\alpha_c \sim \alpha_s \sim 2^K$$

Solution space structure

$$\mu(\vec{\sigma}) = \frac{1}{Z} \prod_{a=1}^{M} \mathbb{I}_a \left(\sigma_{i_a(1)}, \ldots, \sigma_{i_a(k)}\right)$$

$$\boxed{w_\gamma = \sum_{\vec{\sigma} \in A_\gamma} \mu(\vec{\sigma})} \qquad w_1 > w_2 > w_3 > \ldots$$



- **RS**: most of the measure in a single cluster $\lim_{N\to\infty} w_1 = 1$
- **d1RSB**: the measure divides in $e^{N\Sigma^*}$ clusters
- **s1RSB**: the measure condensates in sub-exp number of clusters $\lim_{n\to\infty} \lim_{N\to\infty} \sum_{i=1}^{n} w_i = 1$

Random K-XORSAT

Ricci-Tersenghi, Zecchina & Weigt, PRE '01
Mézard, Ricci-Tersenghi & Zecchina, JSP '03
Cocco, Dubois, Mandler & Monasson, PRL '03

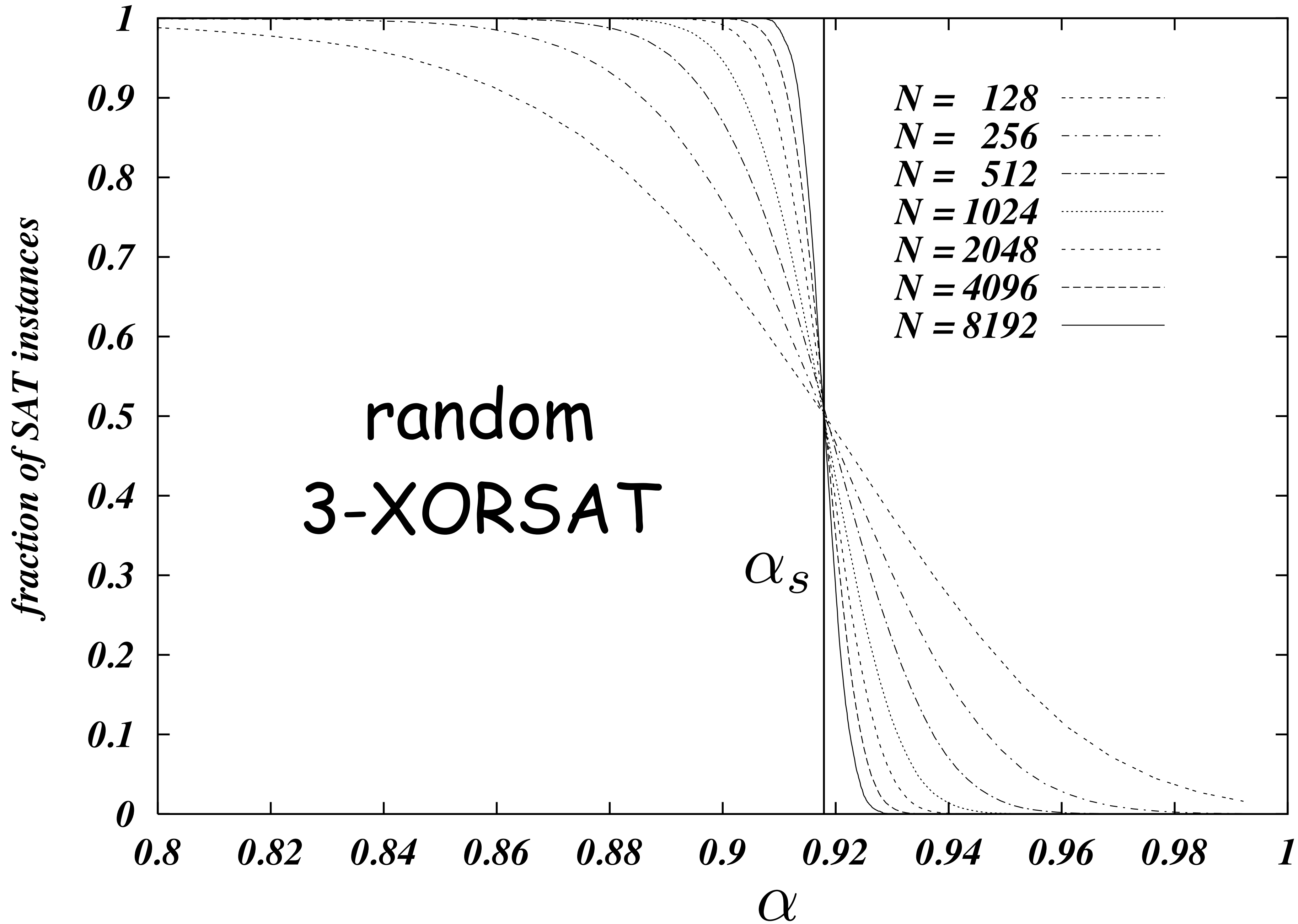
Like random K-SAT but replacing OR with XOR

$$(\sigma_7 \oplus \bar{\sigma}_4 \oplus \sigma_{13}) \wedge (\sigma_{10} \oplus \bar{\sigma}_{13} \oplus \bar{\sigma}_2) \wedge \ldots$$

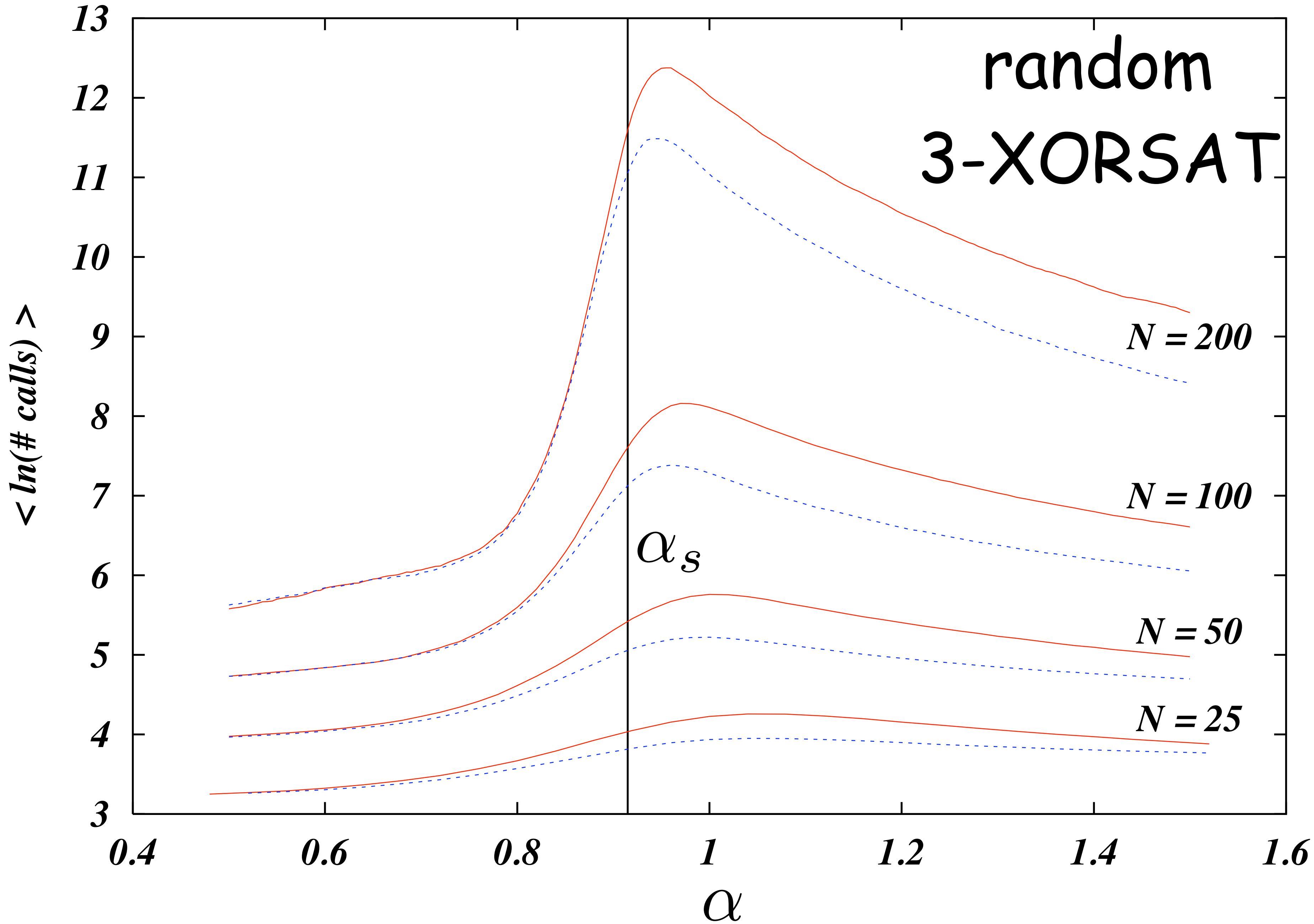
M parity checks over N variables

Equivalent to M linear equations in N binary variables

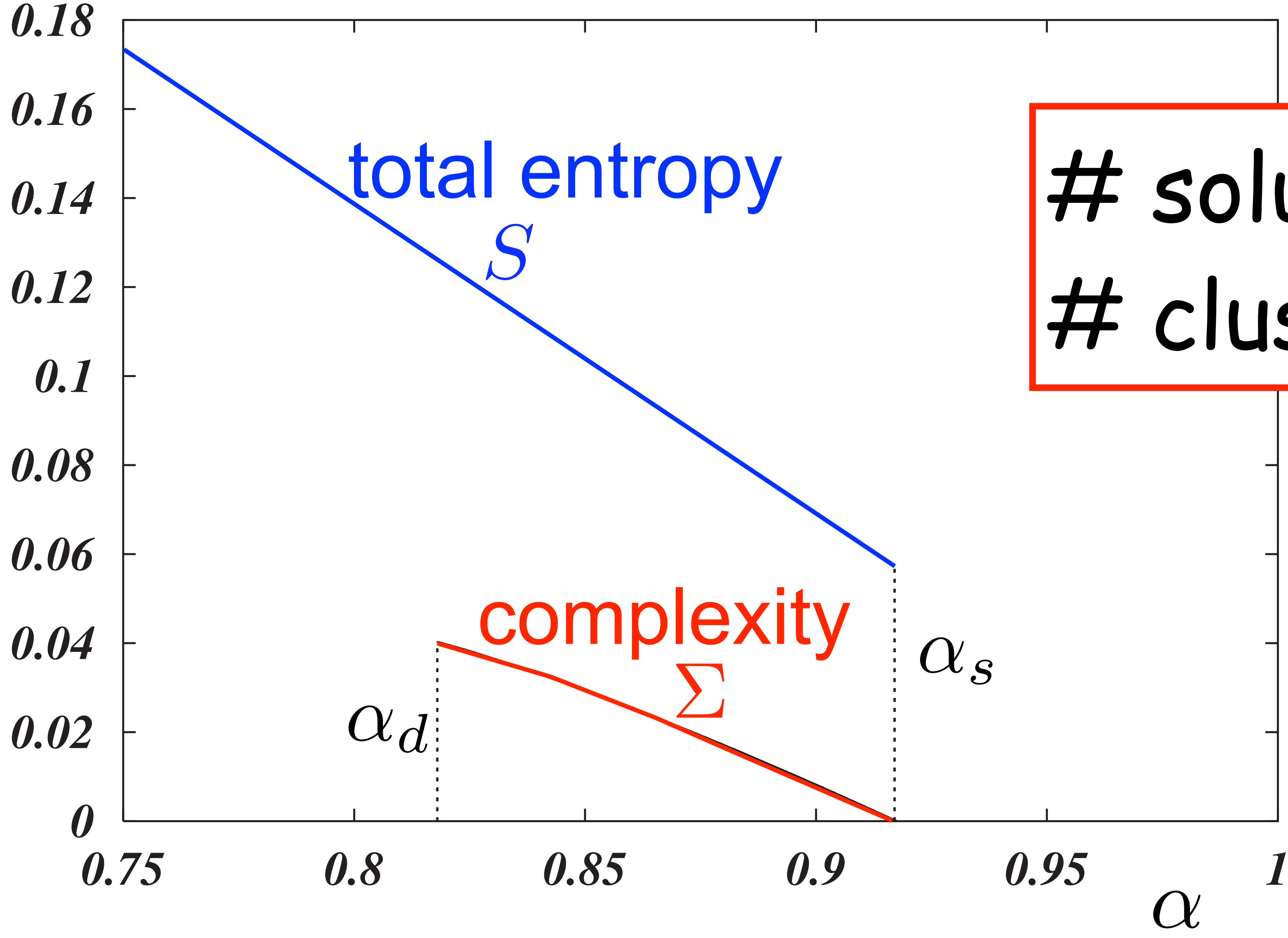
SAT/UNSAT phase transition in random K-XORSAT



Increase in computing times



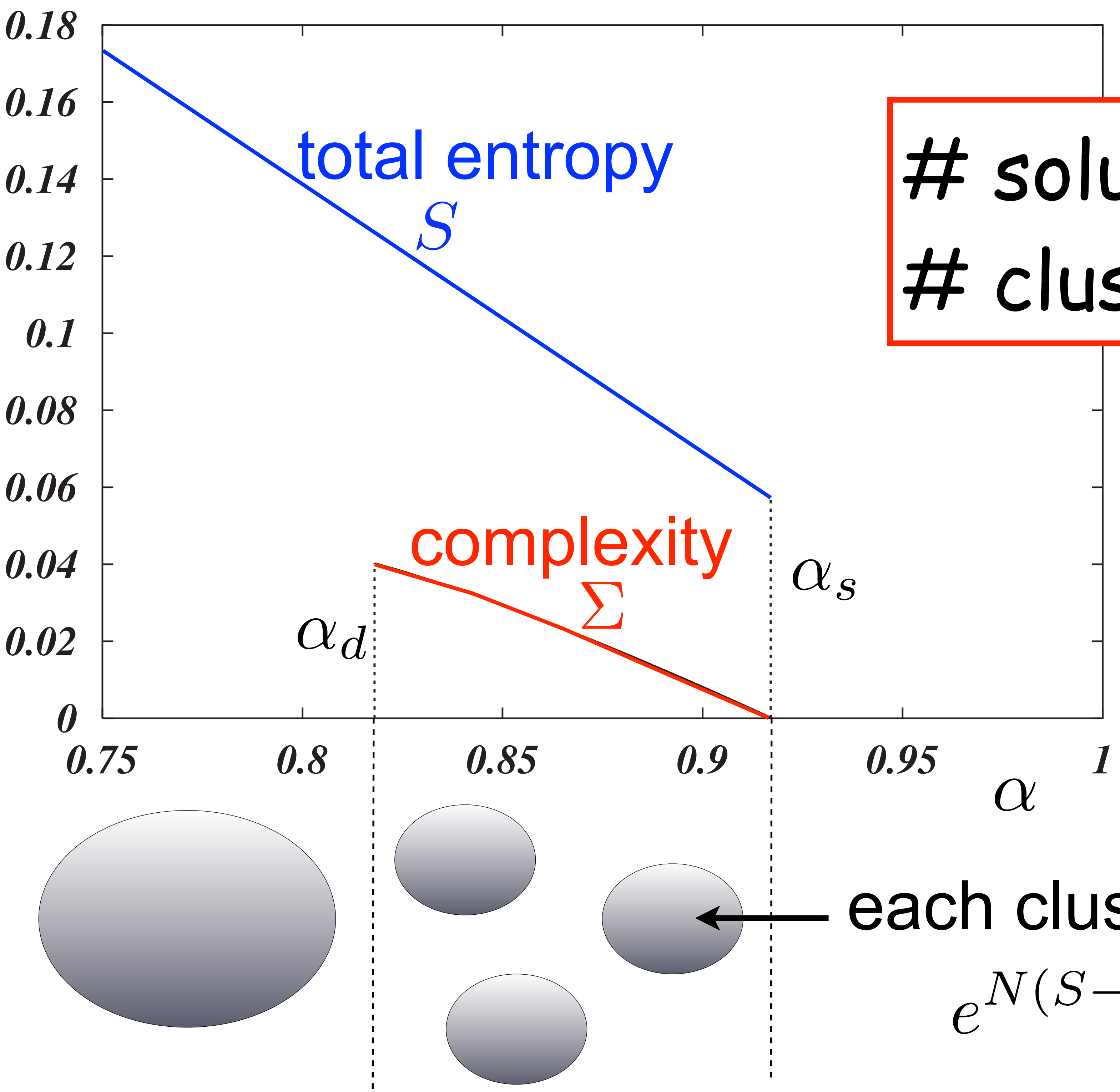
Solution space structure



\# solutions = $\exp(NS)$

\# clusters = $\exp(N\Sigma)$

Solution space structure

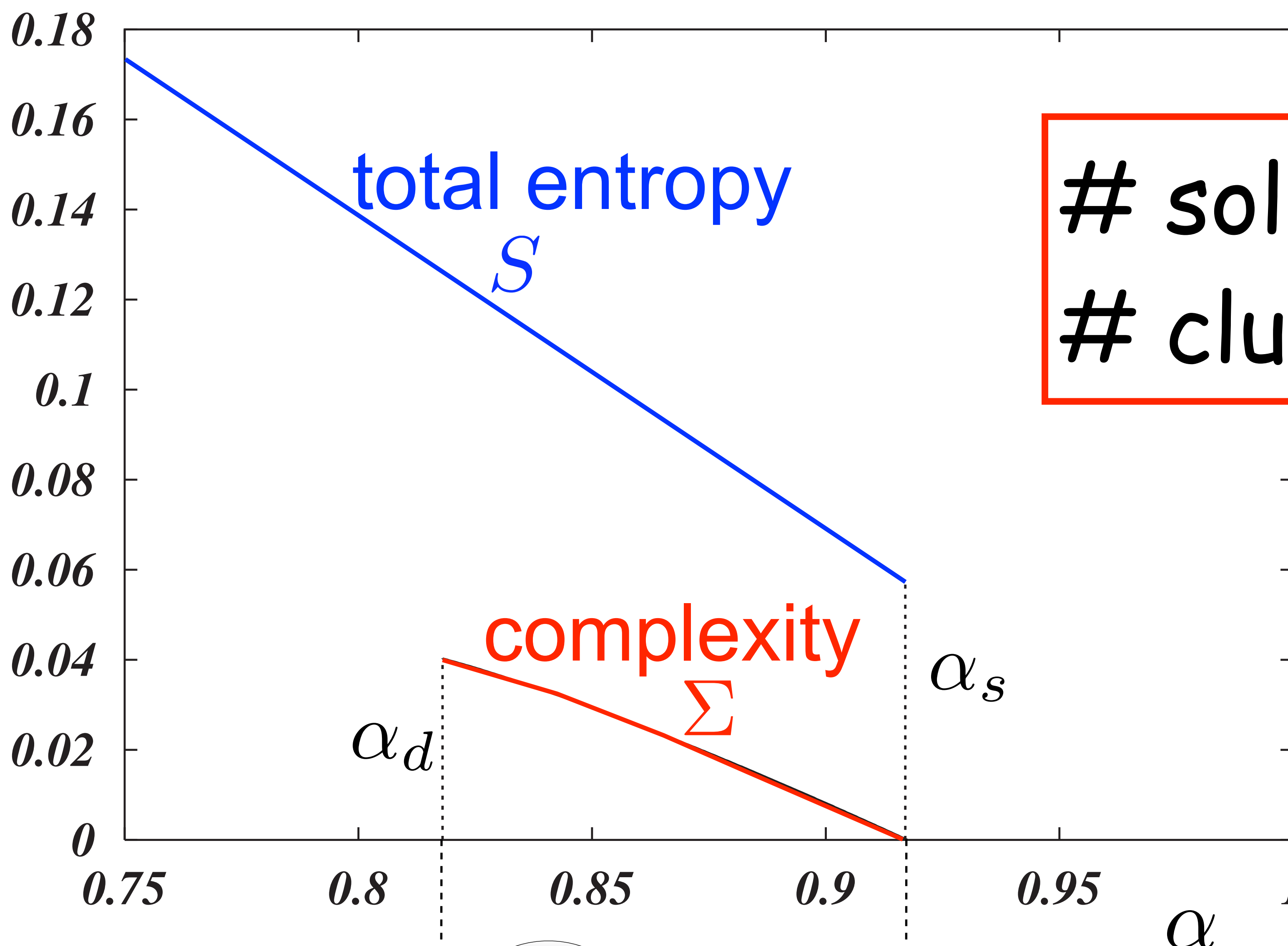


solutions = $\exp(NS)$

clusters = $\exp(N\Sigma)$

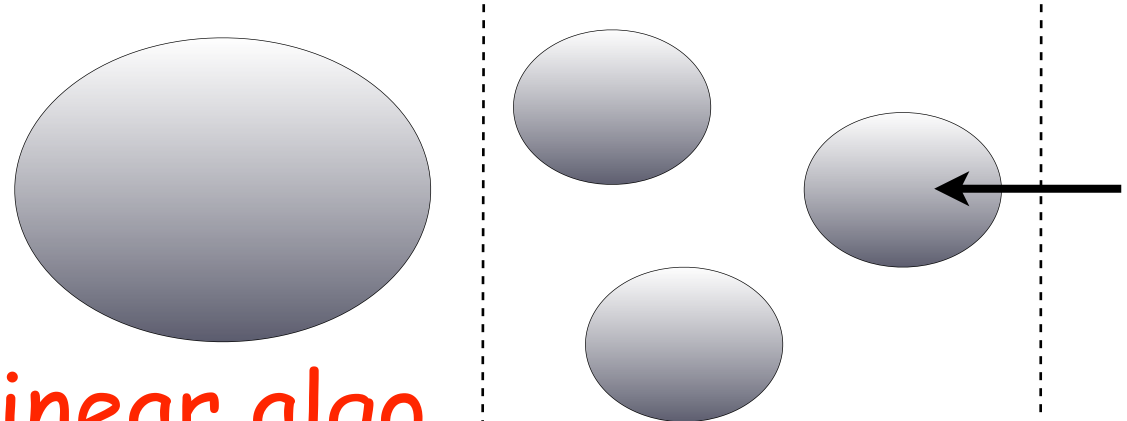
each cluster contains $e^{N(S-\Sigma)}$ solutions

Solution space structure



\# solutions = $\exp(NS)$

\# clusters = $\exp(N\Sigma)$



each cluster contains $e^{N(S-\Sigma)}$ solutions

Leaf removal algorithm

- while (there exists a vertex of degree 1)
  remove it and the clause it belongs to

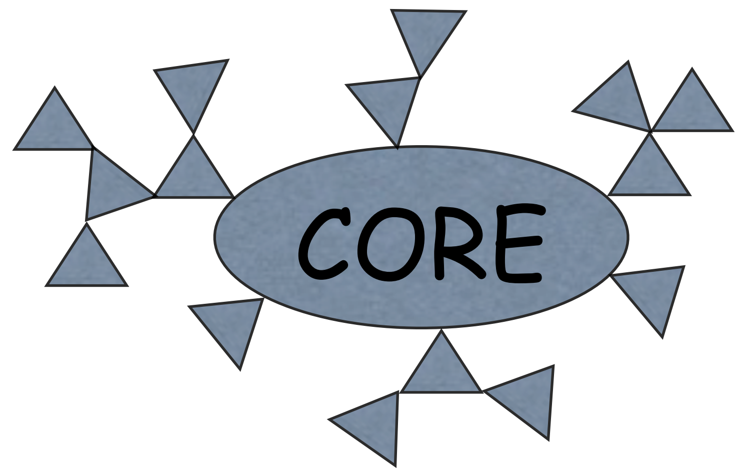
for $\alpha < \alpha_d$ $\qquad \mathcal{G} = (V,E) \to (V_c, \emptyset)$

for $\alpha \geq \alpha_d$ $\qquad \mathcal{G} = (V,E) \to (V_c, E_c)$

- reconstruction procedure for $\alpha < \alpha_d$
  - assign to any value the variables in V_c
  - add clauses in the reverse order and assign the newly added variable to satisfy the clause

The core

For $\alpha \geq \alpha_d$

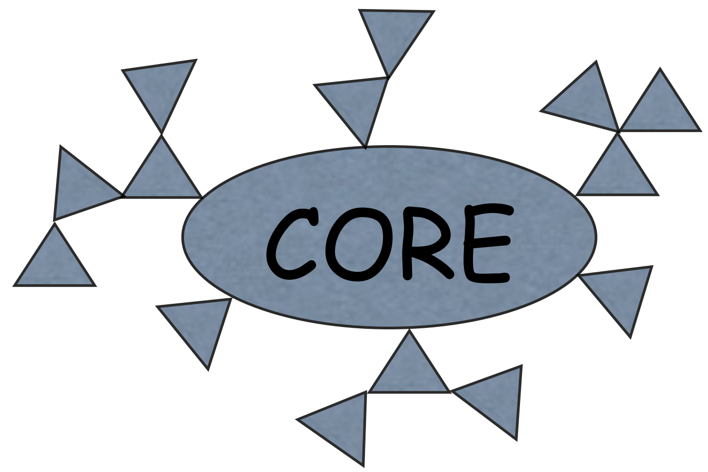


On the core:

- N_c variables, minimum degree 2, M_c clauses
- $\exp(N\Sigma)$ solutions at distance $O(N)$
- long range correlations: hard to find solutions
- solutions exist as long as $M_c \leq N_c$

The core

For $\alpha \geq \alpha_d$

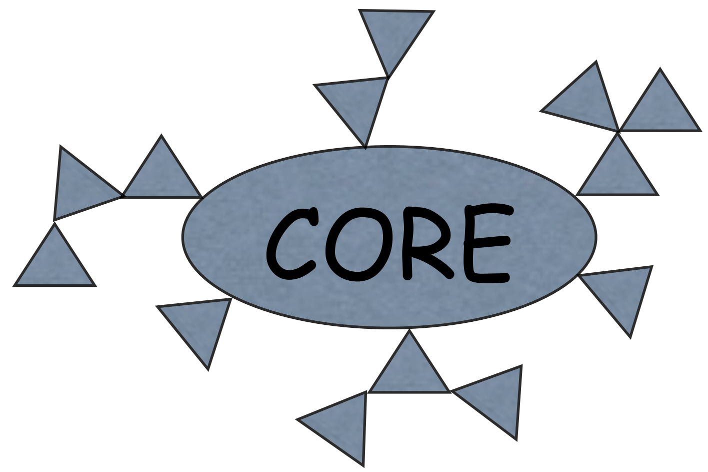


On the core:

- N_c variables, minimum degree 2, M_c clauses
- $\exp(N\Sigma)$ solutions at distance $O(N)$
- long range correlations: hard to find solutions
- solutions exist as long as $M_c \leq N_c$

The core

For $\alpha \geq \alpha_d$

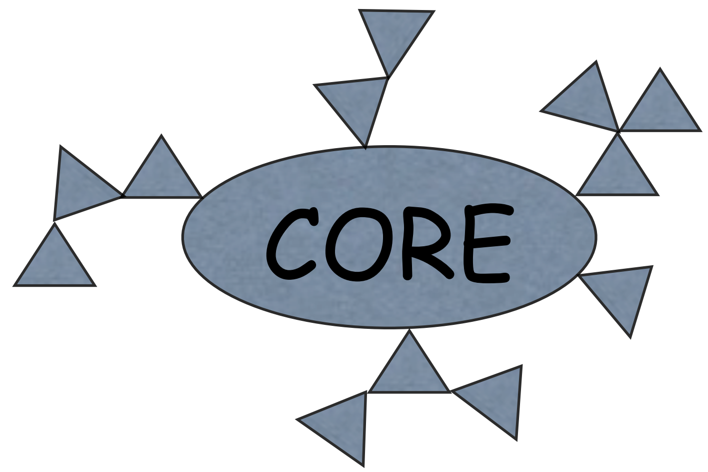


On the core:

- N_c variables, minimum degree 2, M_c clauses
- $\exp(N\Sigma)$ solutions at distance $O(N)$
- long range correlations: hard to find solutions
- solutions exist as long as $M_c \leq N_c$

The core

For $\alpha \geq \alpha_d$

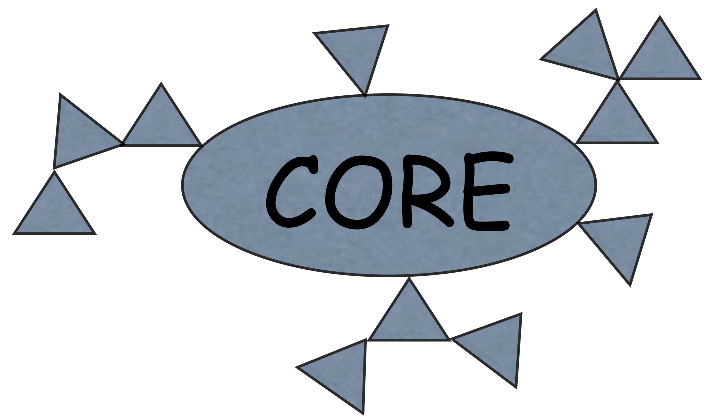


On the core:

- N_c variables, minimum degree 2, M_c clauses
- $\exp(N\Sigma)$ solutions at distance $O(N)$
- long range correlations: hard to find solutions
- solutions exist as long as $M_c \leq N_c$

The core

For $\alpha \geq \alpha_d$

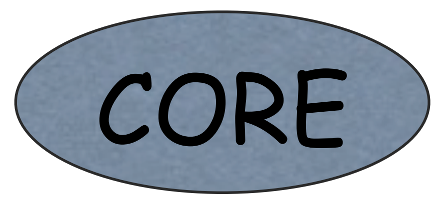


On the core:

- N_c variables, minimum degree 2, M_c clauses
- $\exp(N\Sigma)$ solutions at distance $O(N)$
- long range correlations: hard to find solutions
- solutions exist as long as $M_c \leq N_c$

The core

For $\alpha \geq \alpha_d$

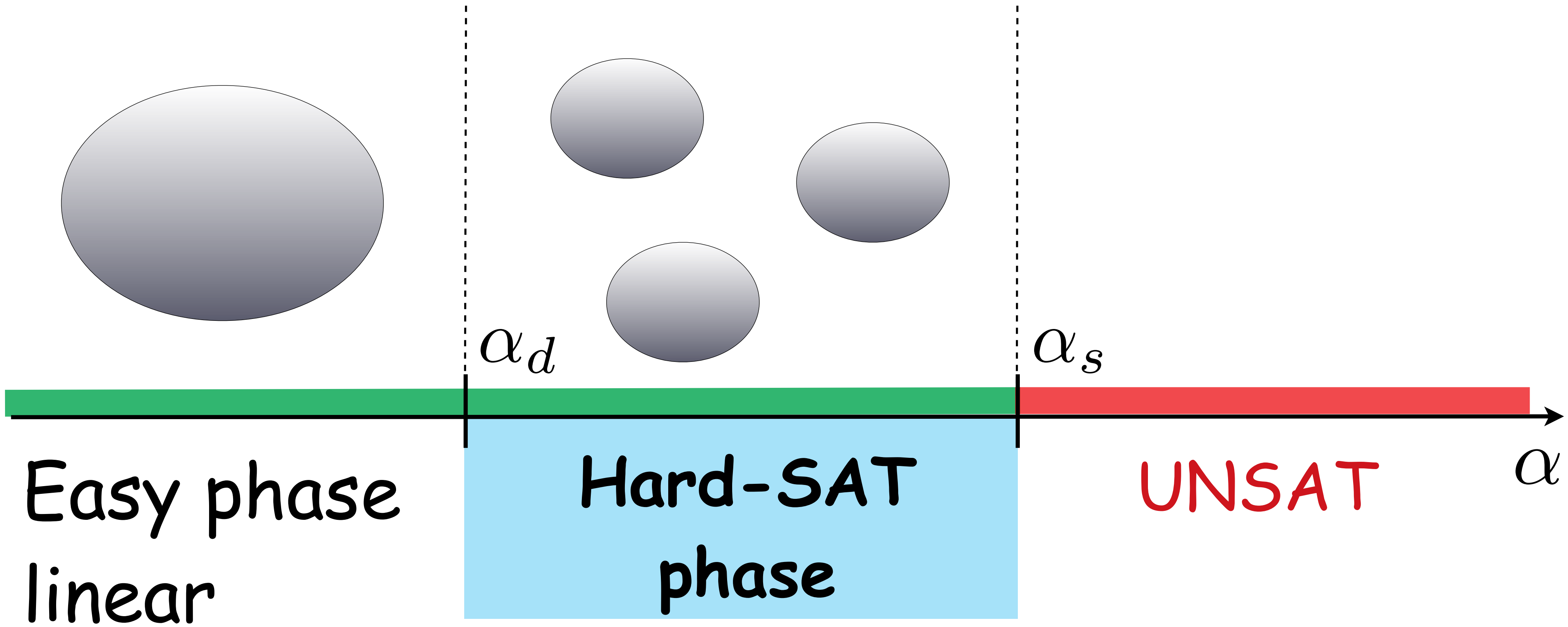


On the core:

- N_c variables, minimum degree 2, M_c clauses
- $\exp(N\Sigma)$ solutions at distance $O(N)$
- long range correlations: hard to find solutions
- solutions exist as long as $M_c \leq N_c$

Where are hard instances?

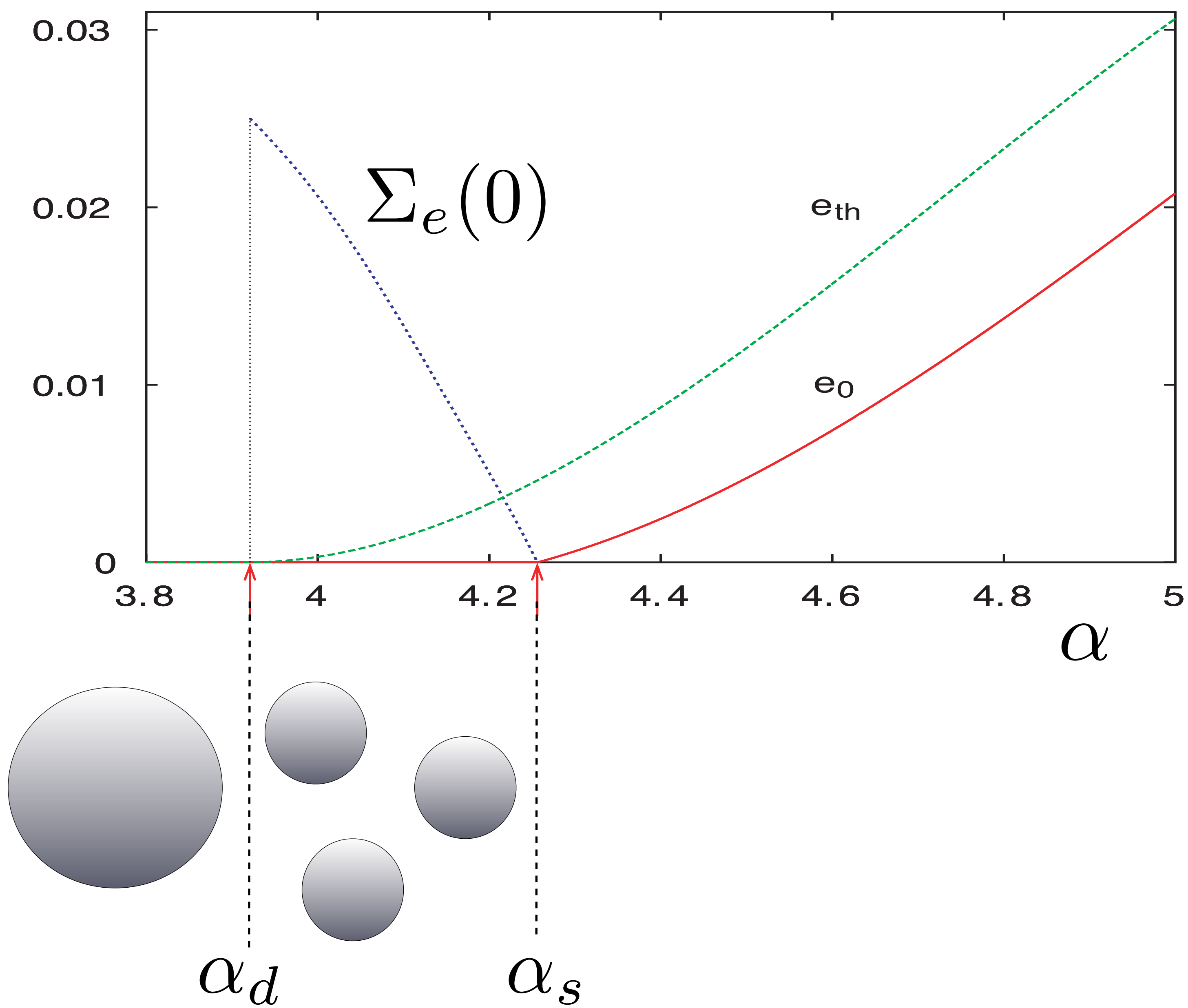
(random K-XORSAT)



Easy phase
linear
algorithms
exist

Solutions space structure
(random 3-SAT)

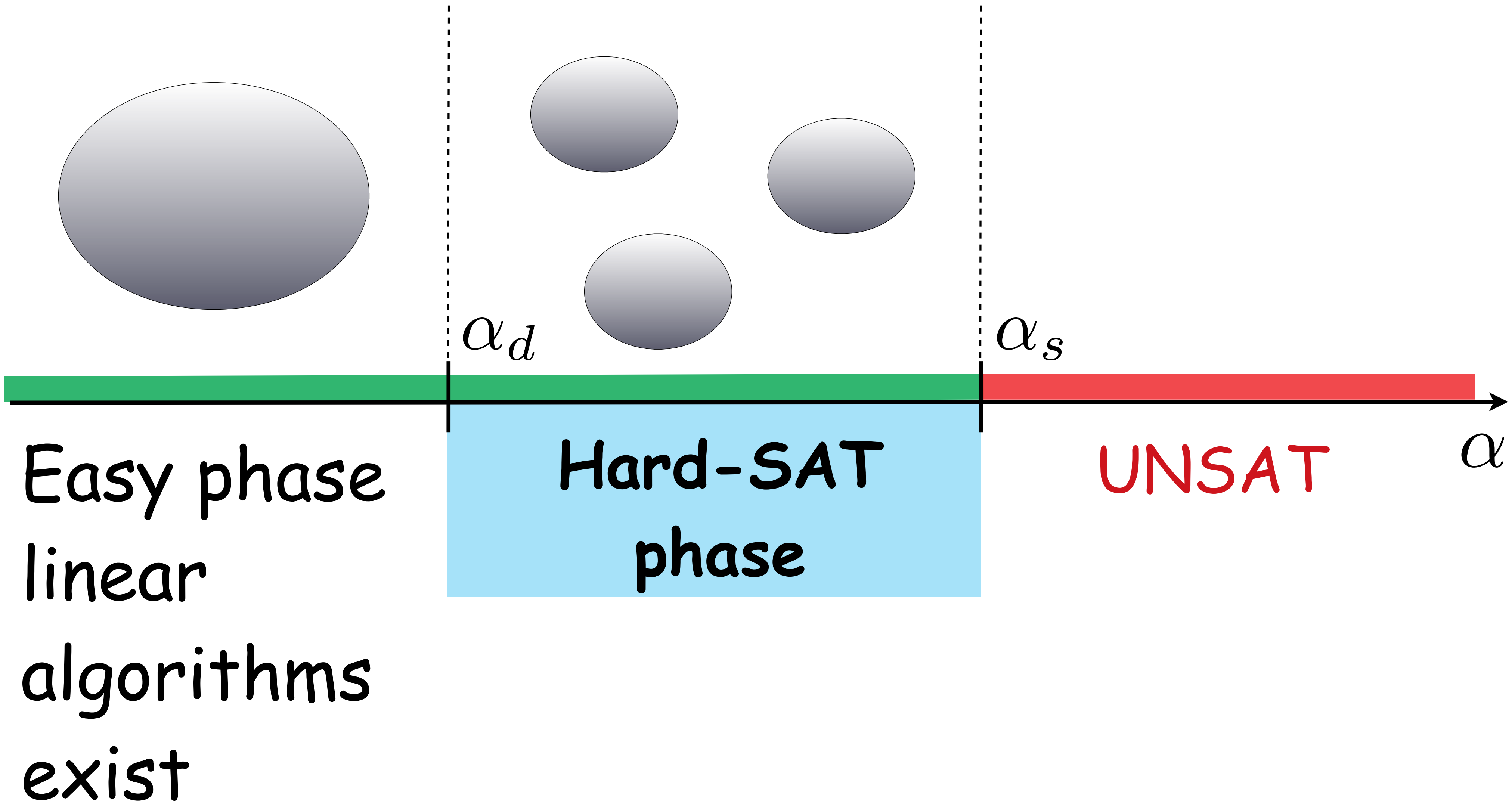
Mézard, Parisi & Zecchina, Science '02



Where are hard instances?

(random 3-SAT)

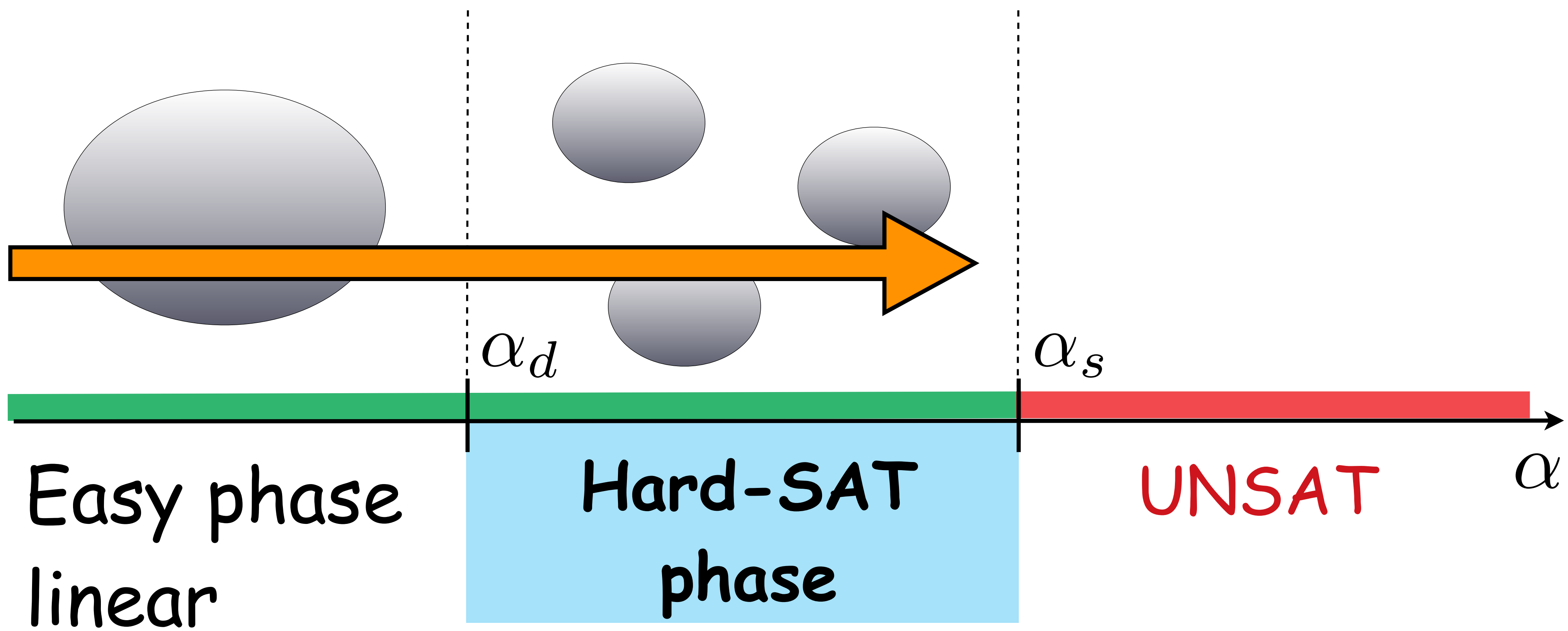
Mézard, Parisi & Zecchina, Science '02



Where are hard instances?

(random 3-SAT)

Mézard, Parisi & Zecchina, Science '02



Easy phase
linear
algorithms
exist

Counting solutions clusters

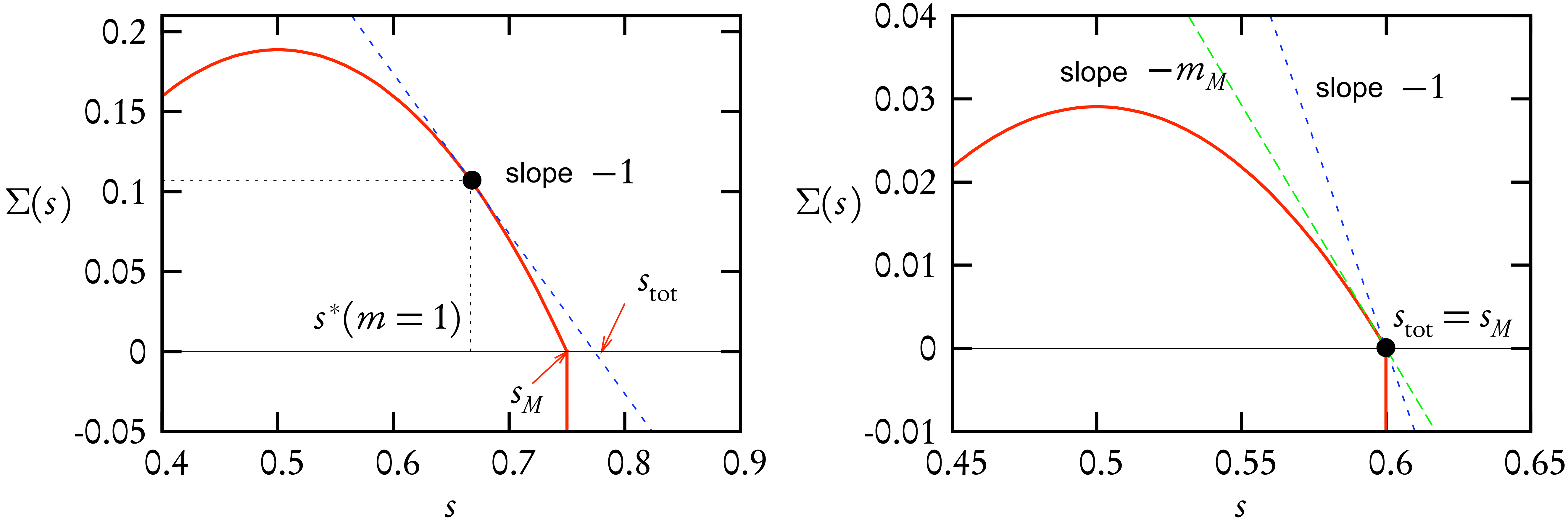
$e^{N\Sigma(s)}$ = # clusters of size e^{Ns}

$$\sum_s e^{N[\Sigma(s)+s]} \simeq \exp\left(N \max_{s:\Sigma(s)\geq 0}[\Sigma(s)+s]\right)$$

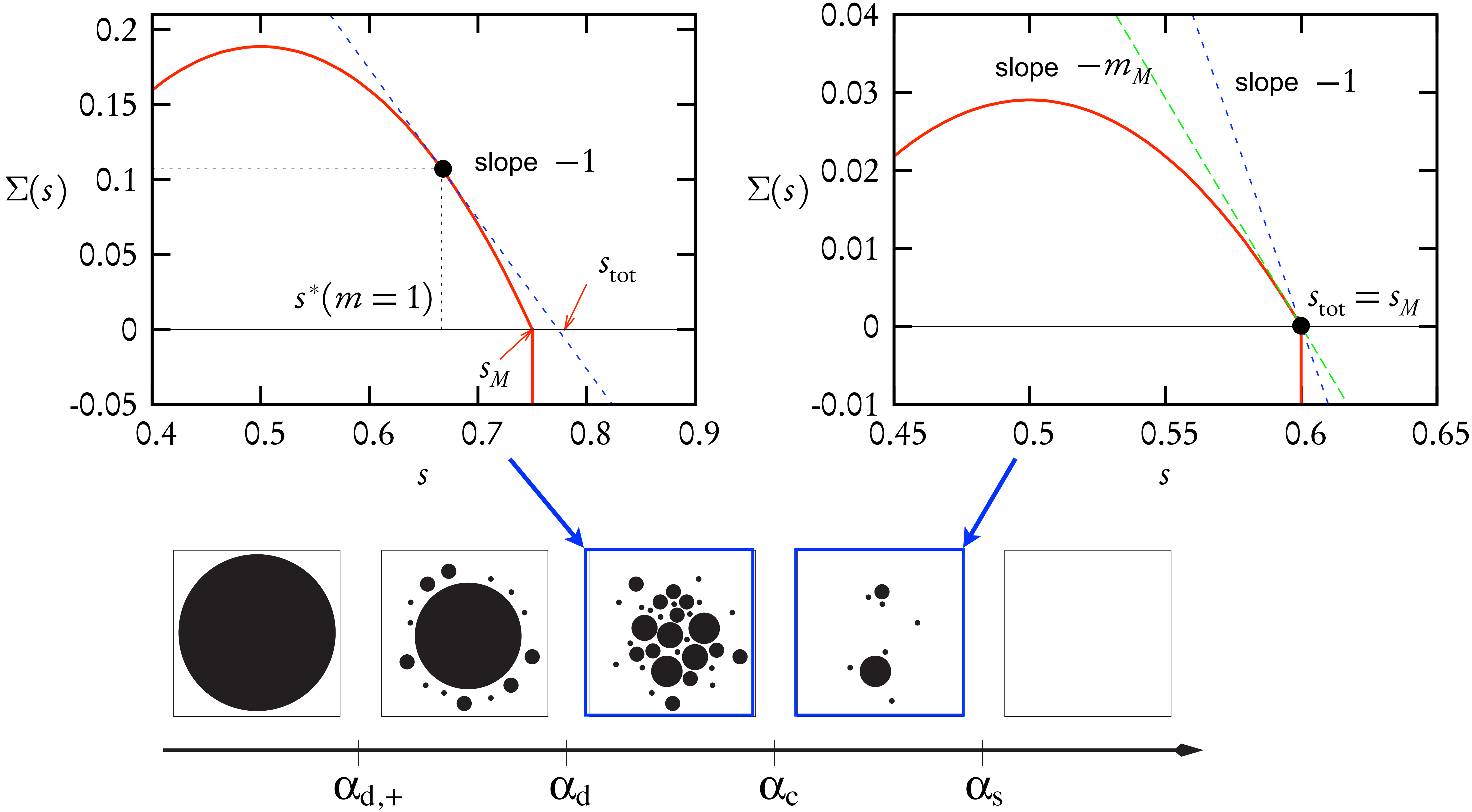
Dominating clusters have size e^{Ns^*}

$$s^* = \arg\max_{s:\Sigma(s)\geq 0}[\Sigma(s)+s]$$

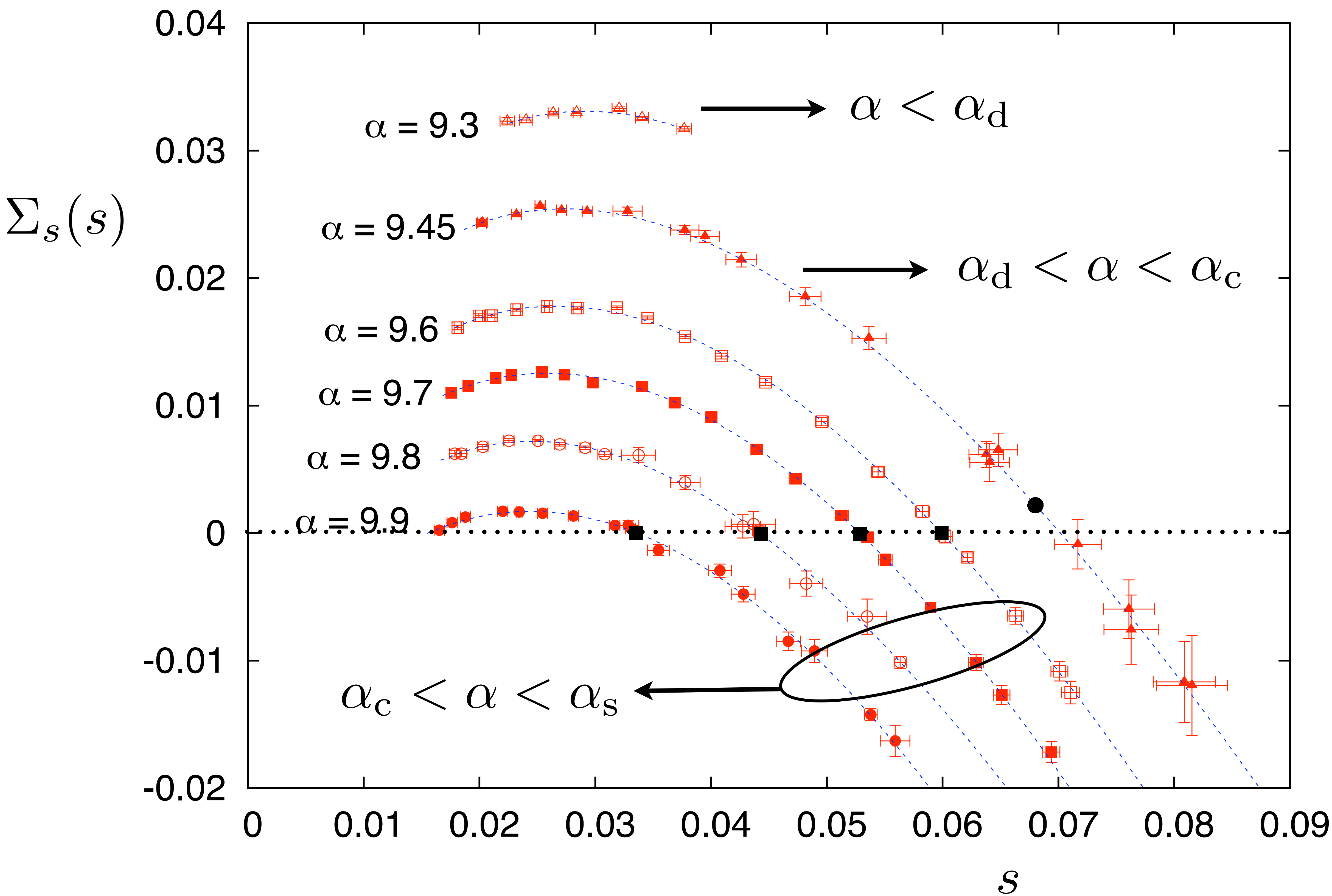
Counting solutions clusters



Counting solutions clusters

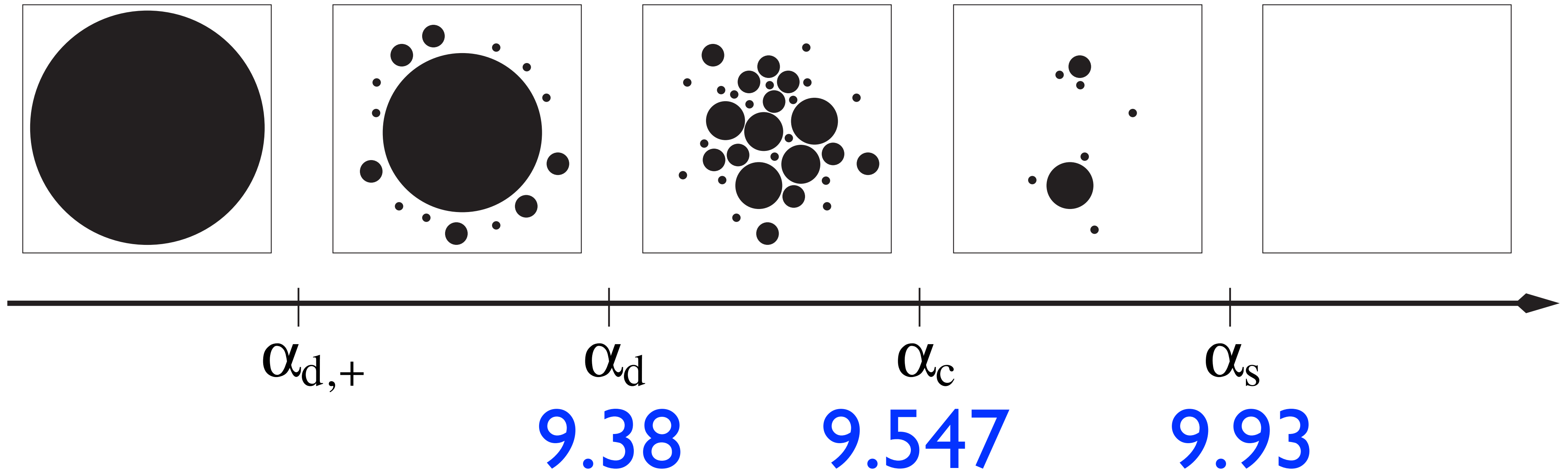


Random 4-SAT

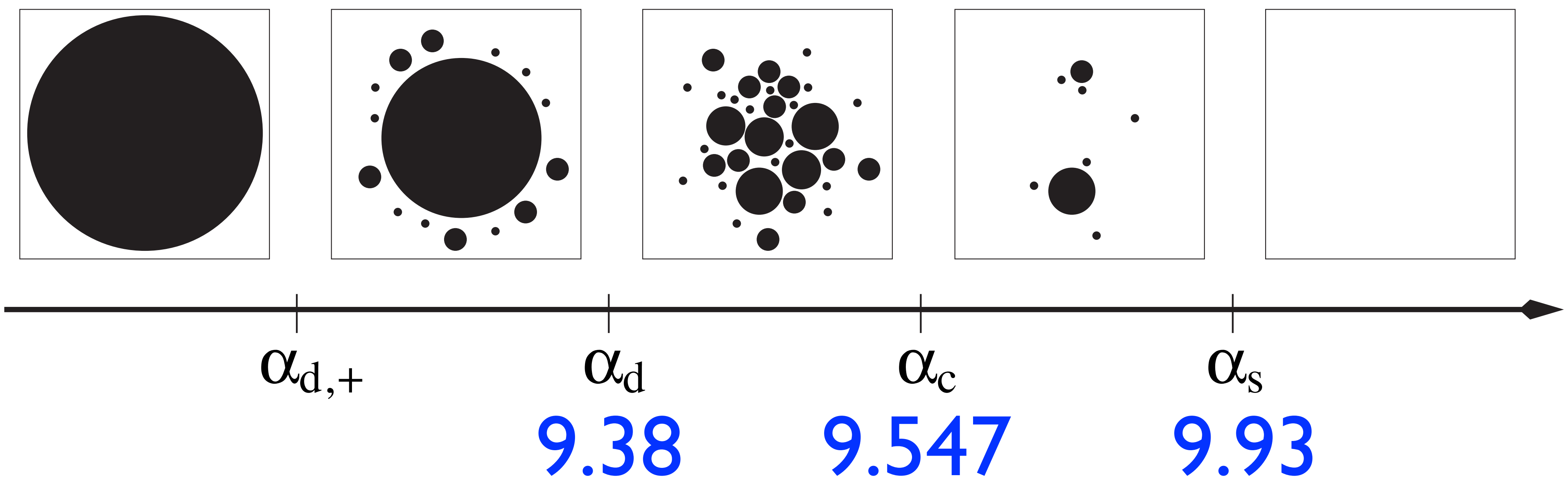


$\alpha = 9.3$
$\alpha = 9.45$
$\alpha = 9.6$
$\alpha = 9.7$
$\alpha = 9.8$
$\alpha = 9.9$
$\alpha < \alpha_{\rm d}$
$\alpha_{\rm d} < \alpha < \alpha_{\rm c}$
$\alpha_{\rm c} < \alpha < \alpha_{\rm s}$
$\Sigma_s(s)$
s
0.04
0.03
0.02
0.01
0
-0.01
-0.02
0
0.01
0.02
0.03
0.04
0.05
0.06
0.07
0.08
0.09

Random 4-SAT

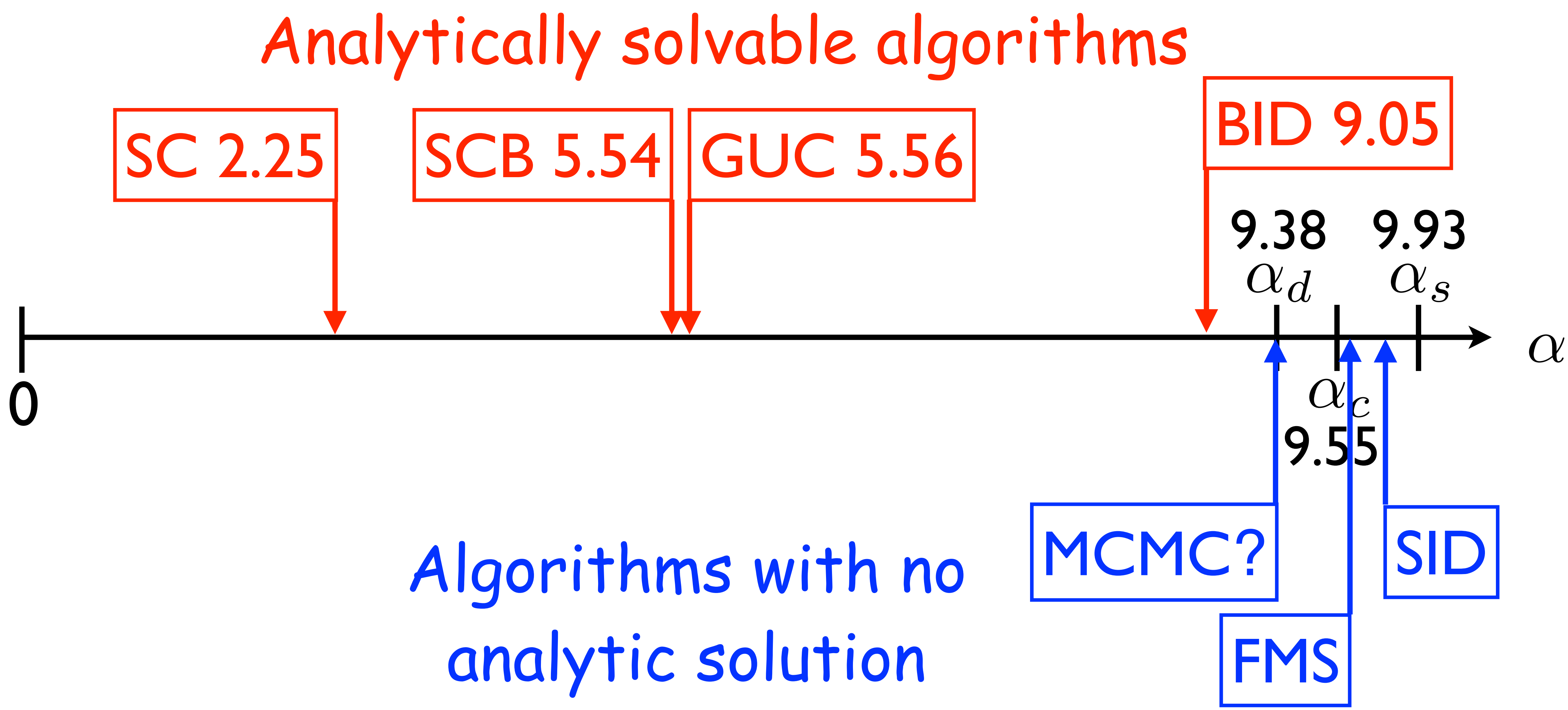


Random 4-SAT

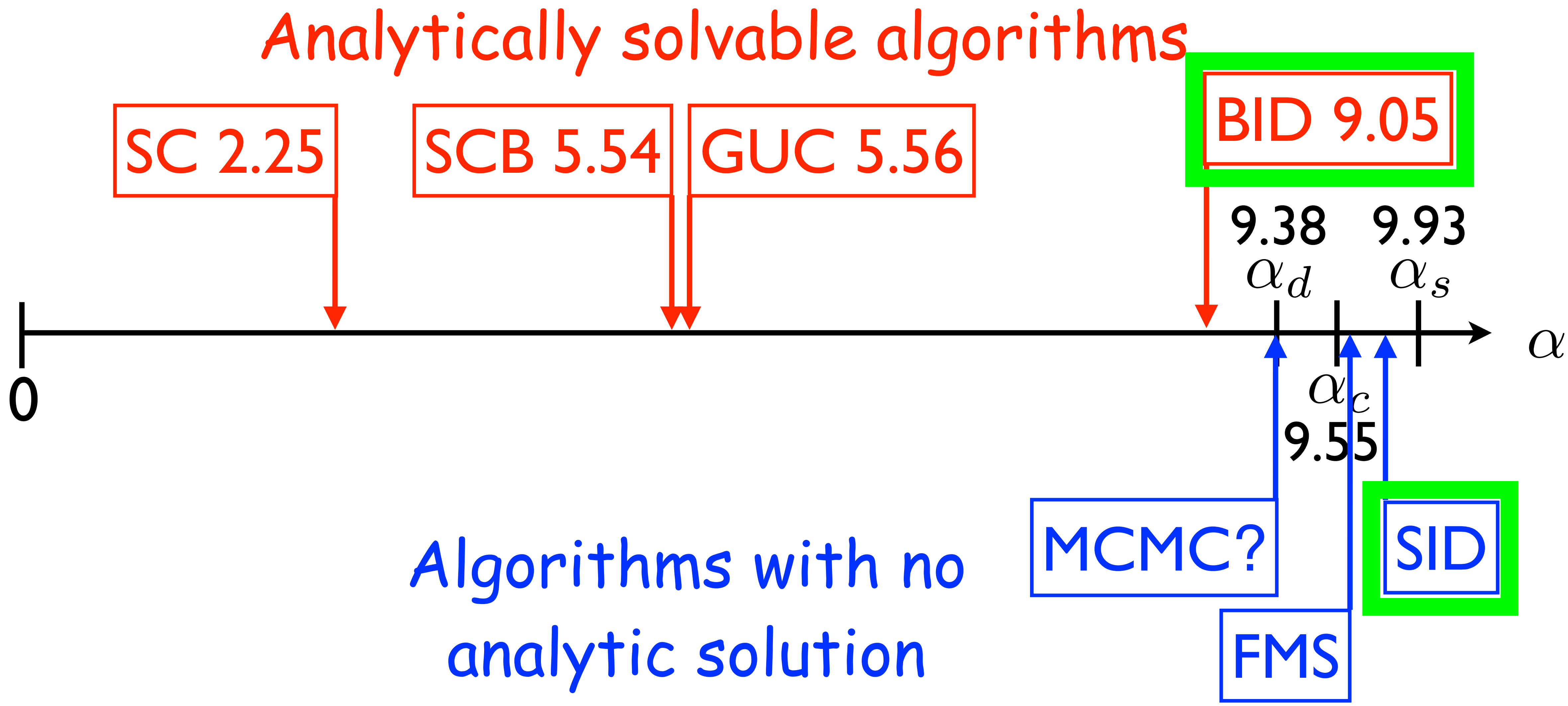


$5.56 \leq \alpha_a$ rigorous bounds $7.91 < \alpha_s < 10.23$

Algorithms performances
(random 4-SAT)



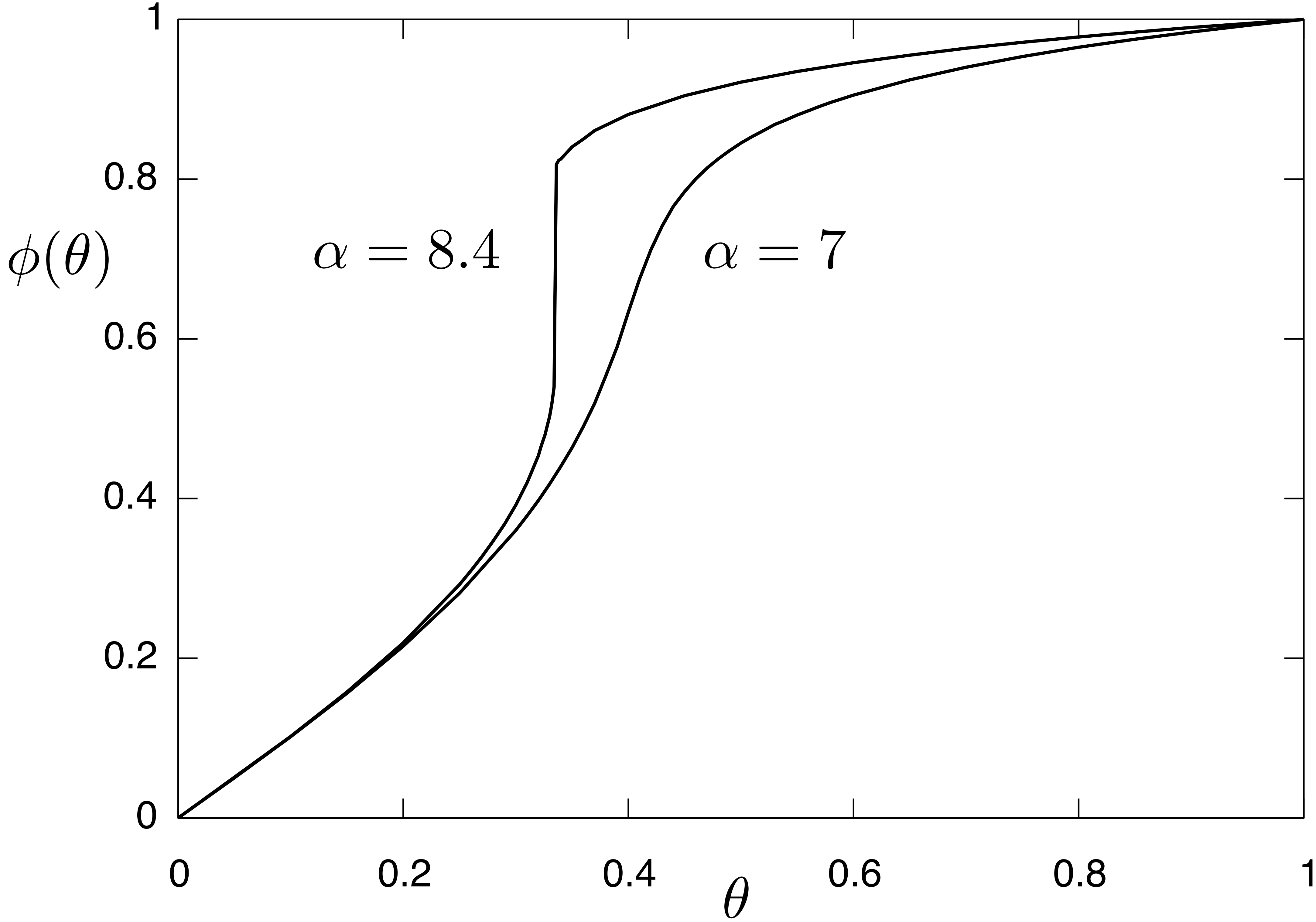
Algorithms performances
(random 4-SAT)



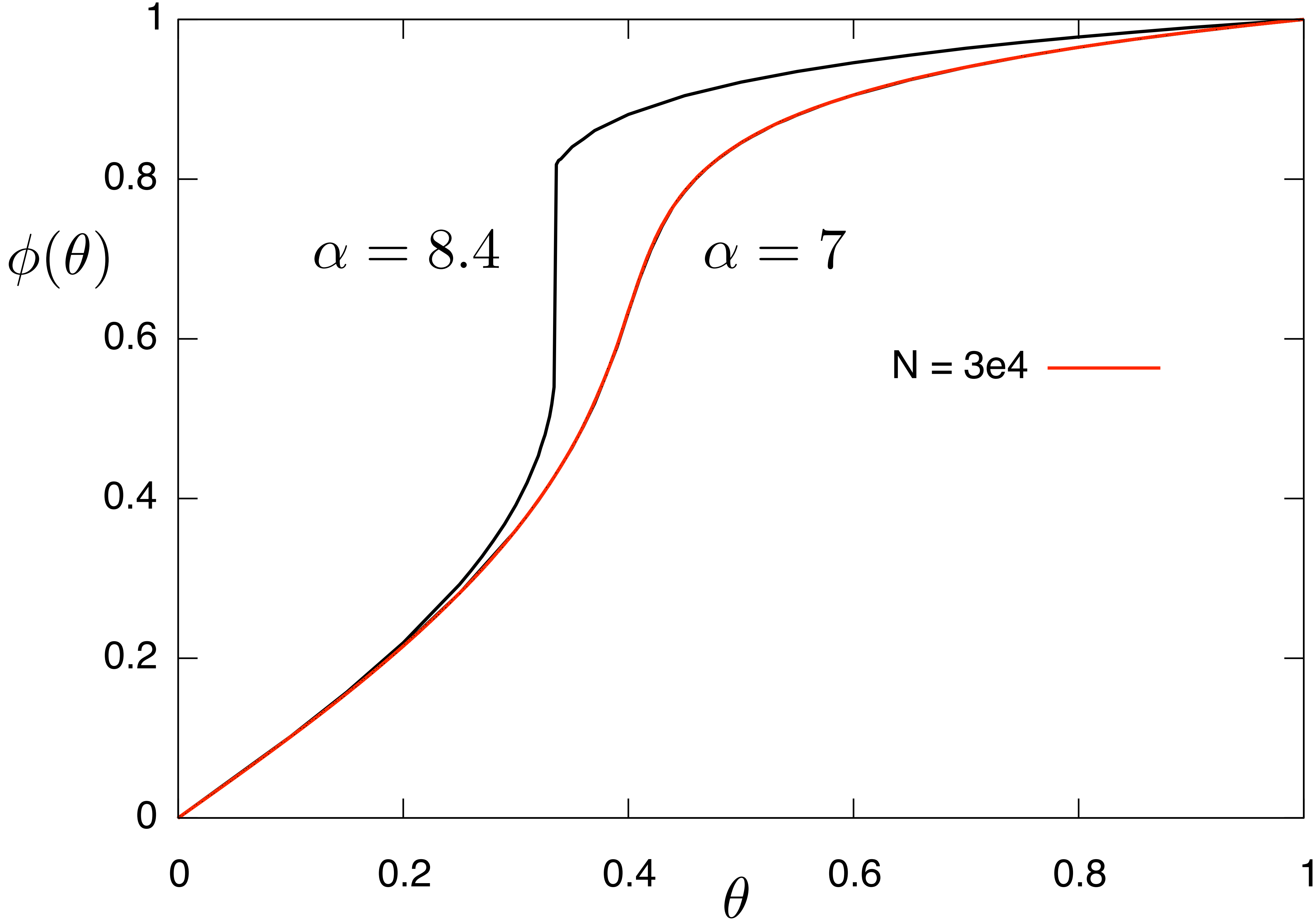
Beliefs/surveys inspired decimation (BID/SID)

- while (there are unassigned variables)
  - compute marginals (with BP or SP)
  - choose an unassigned variable (randomly / the most biased)
  - fix it (according to its marginal / to the most probable value)
  - simplify the formula by UCP

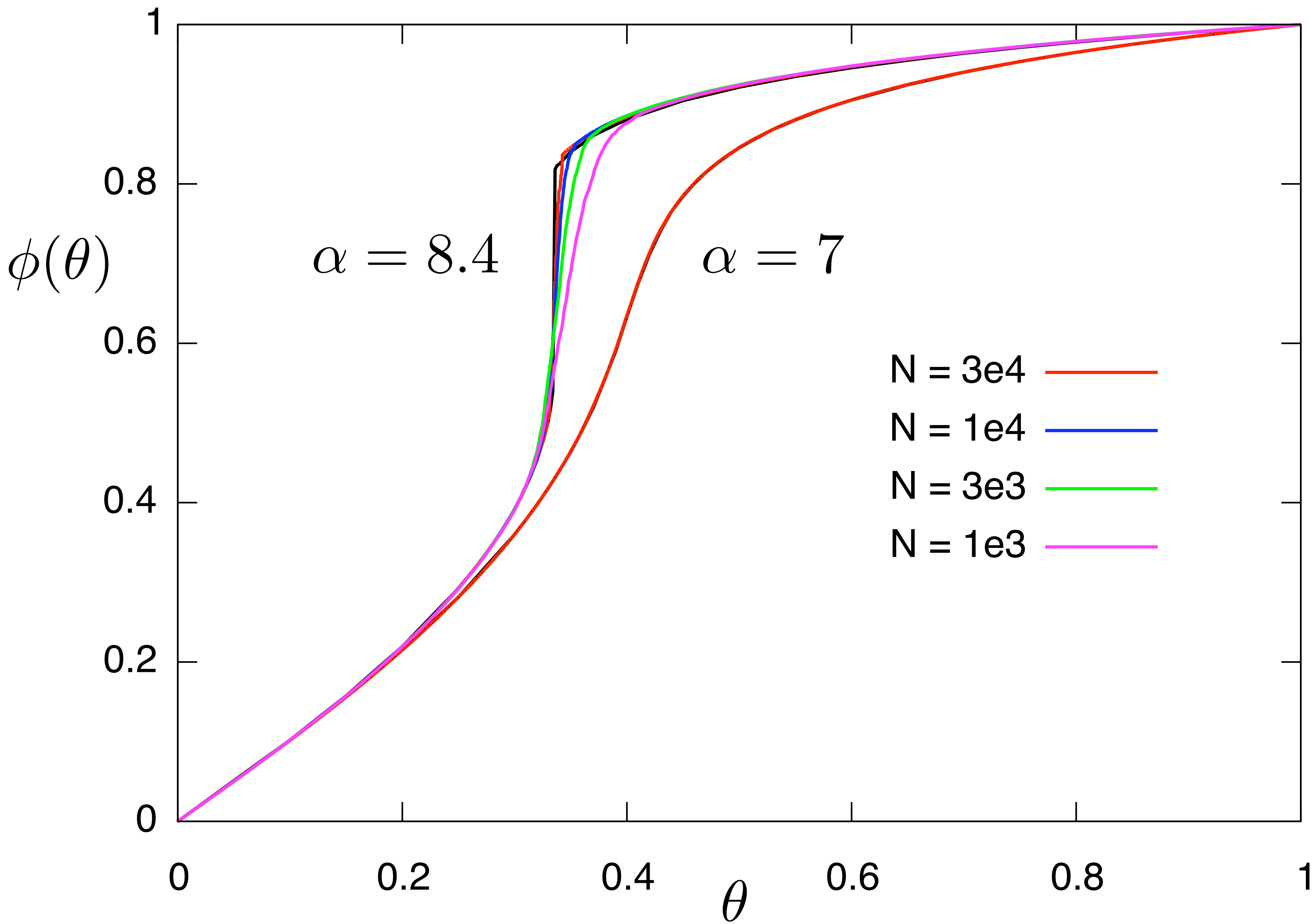
BID for random 4-SAT



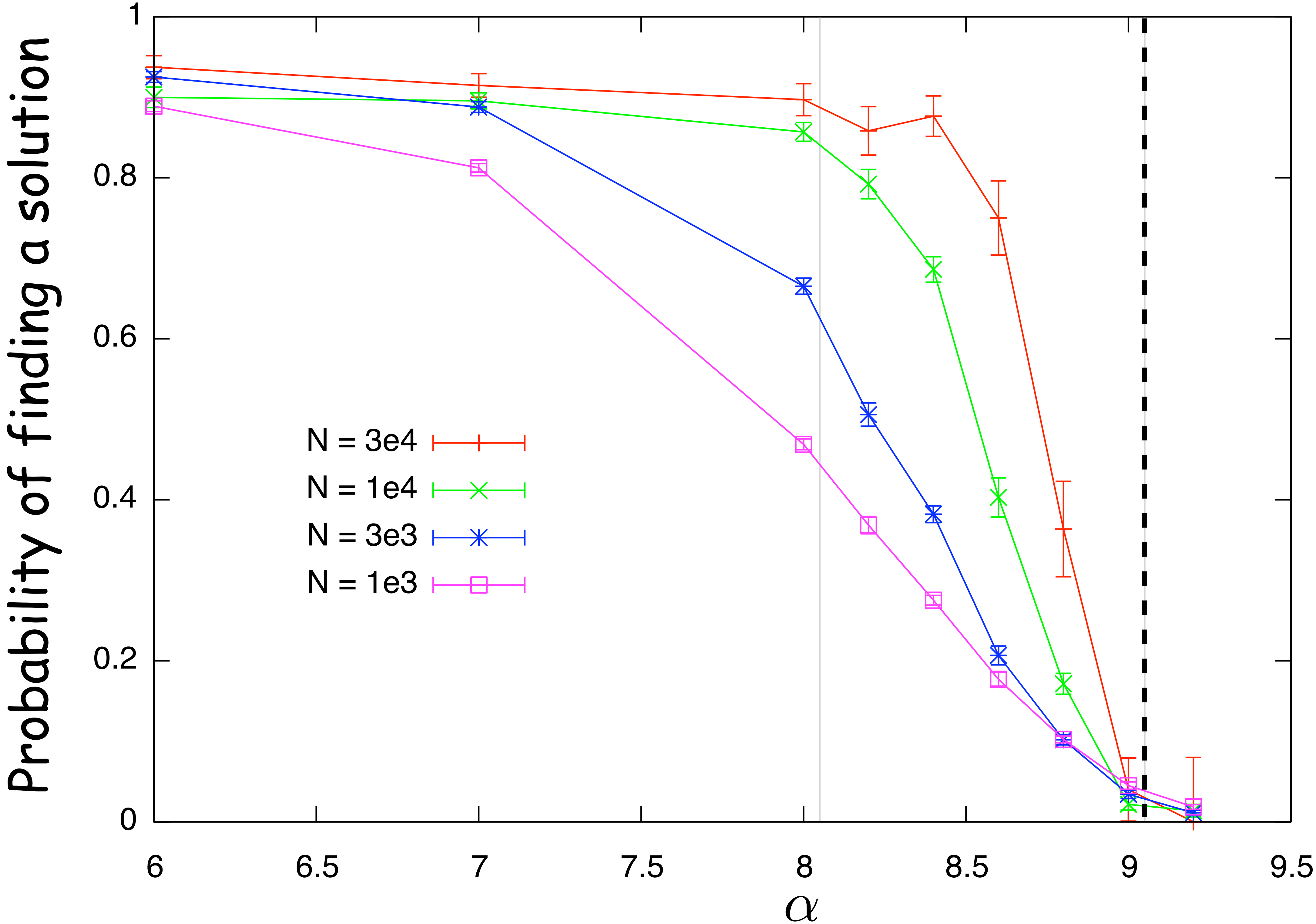
BID for random 4-SAT



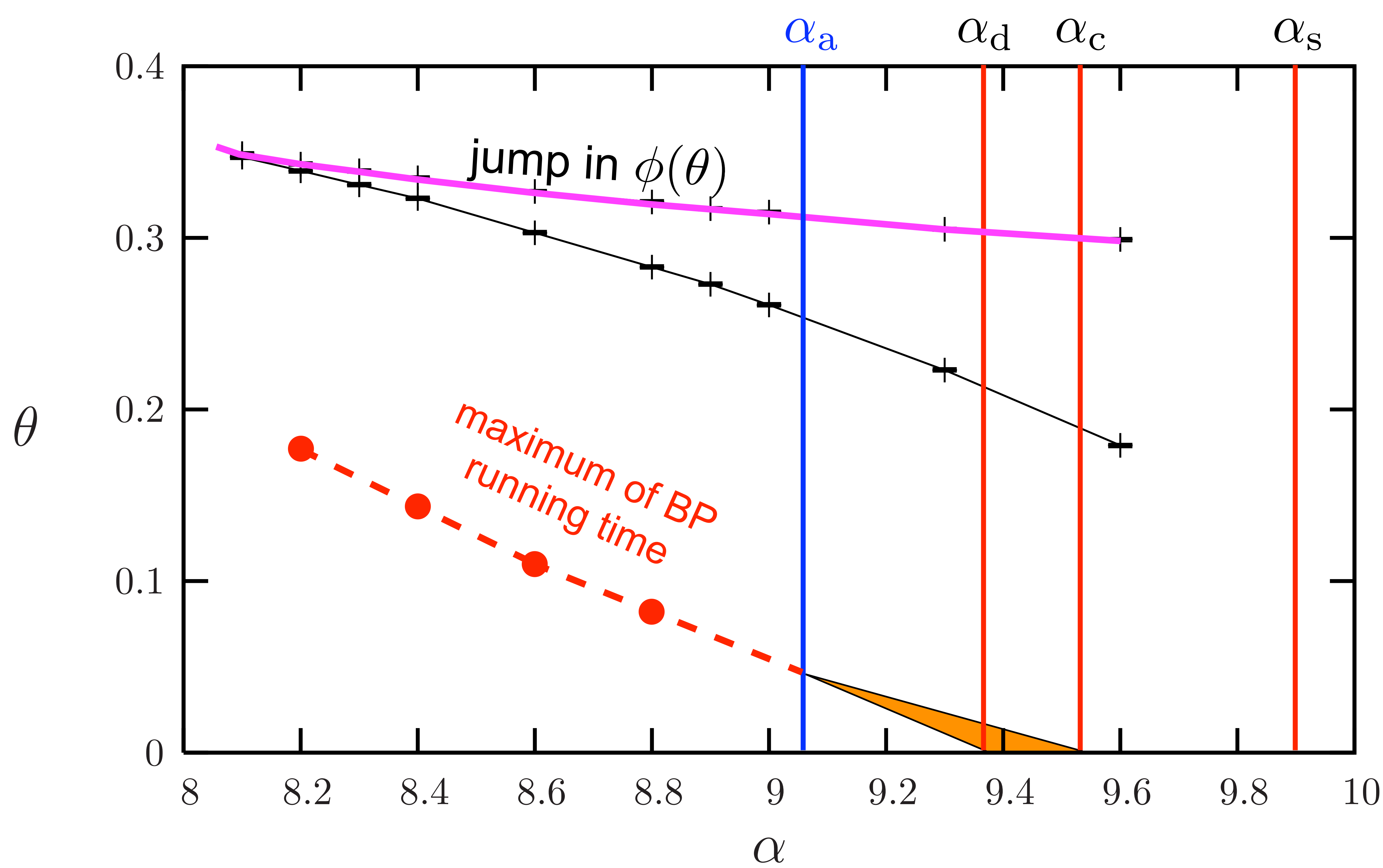
BID for random 4-SAT



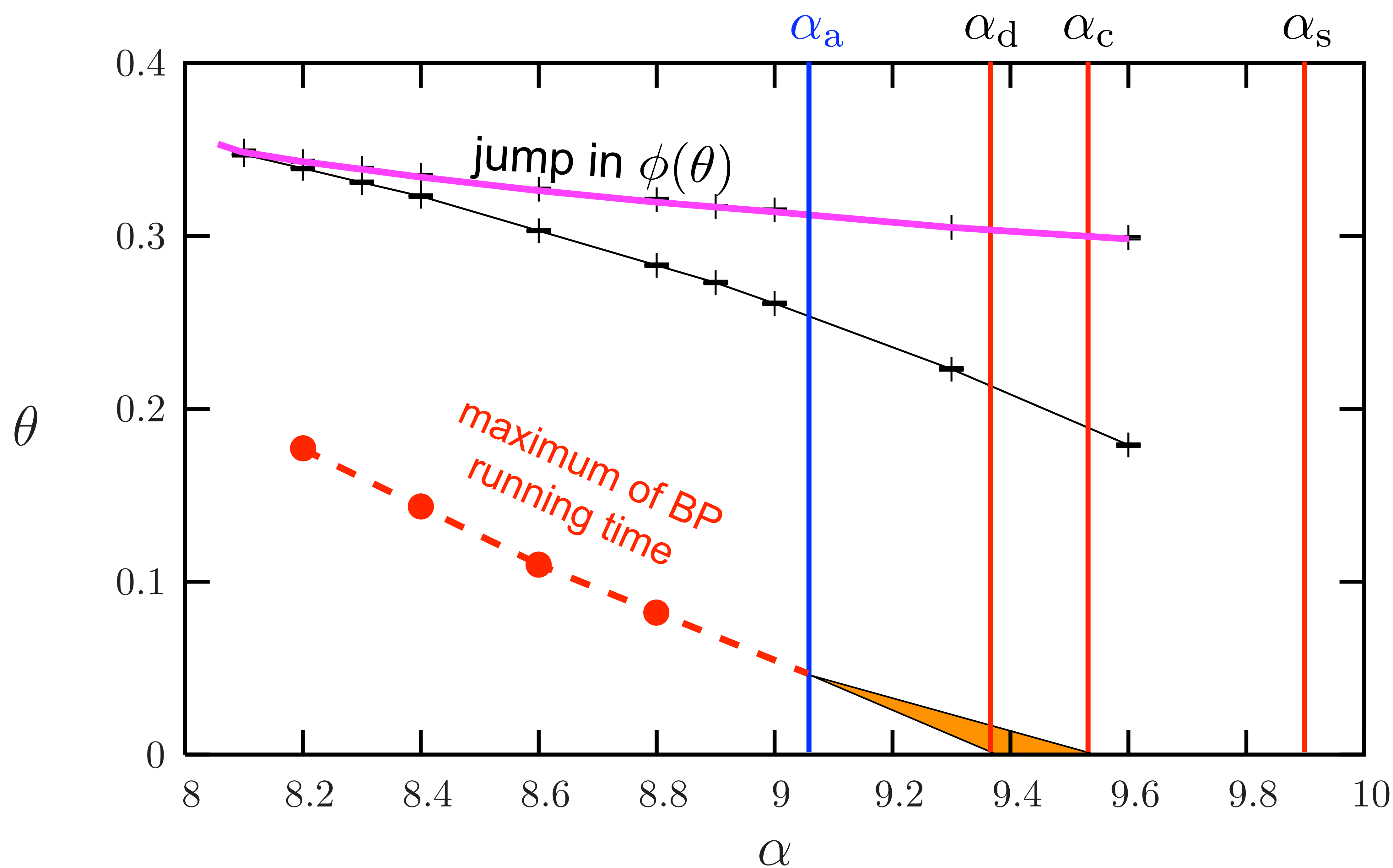
BID for random 4-SAT



BID for random 4-SAT



BID for random 4-SAT



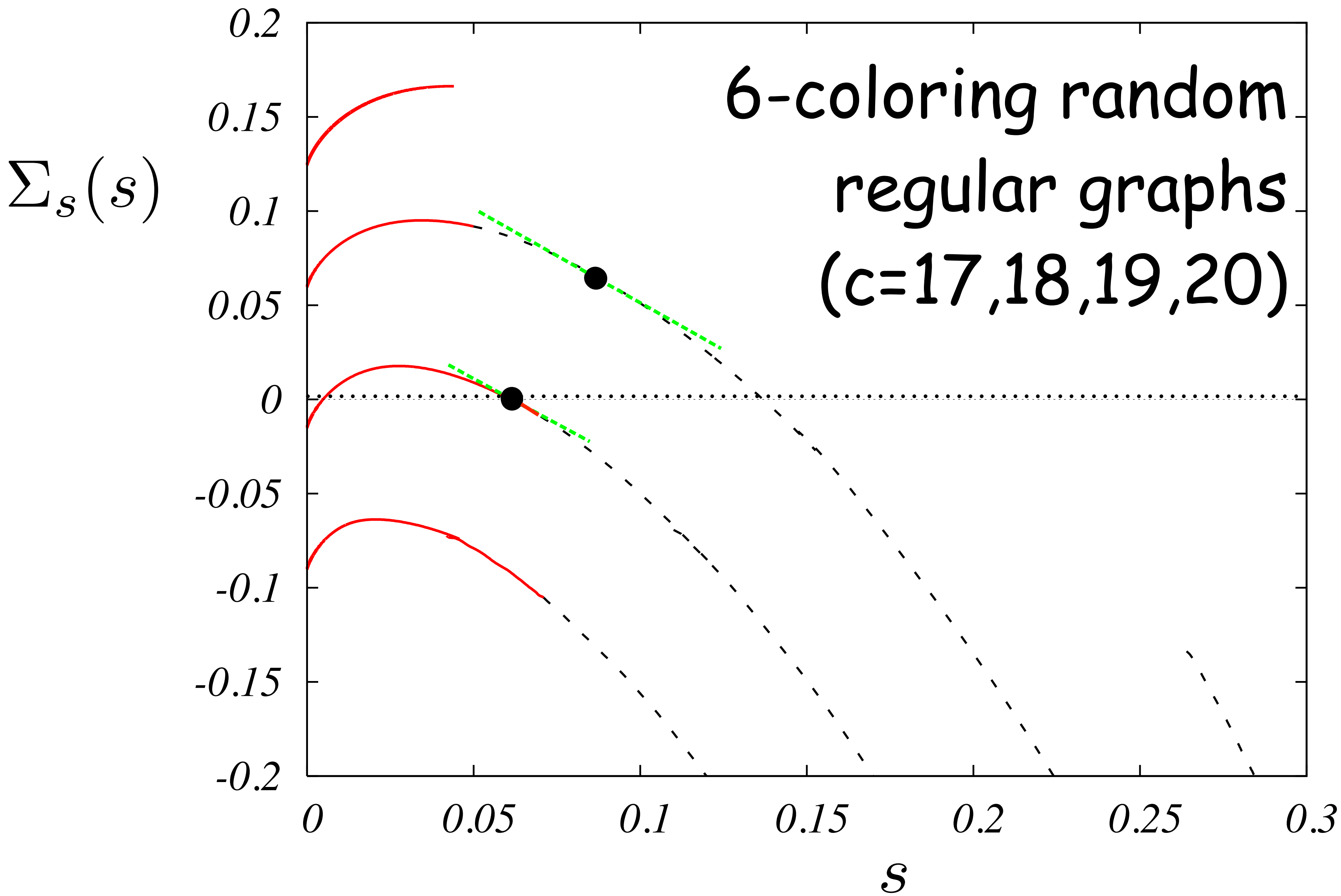
proved for large k in Coja-Oghlan, Pachon-Pinzon, arxiv:1102.3145

Solution space structure vs. algorithmic behavior

- Most algorithms are <u>local</u>: take decisions looking at a bounded neighbourhood
- If strong correlations develop between <u>distant</u> variables, local algorithms are deemed to fail
- Is the condensation threshold $\alpha_{\rm c}$ the natural limit for local algorithms ?

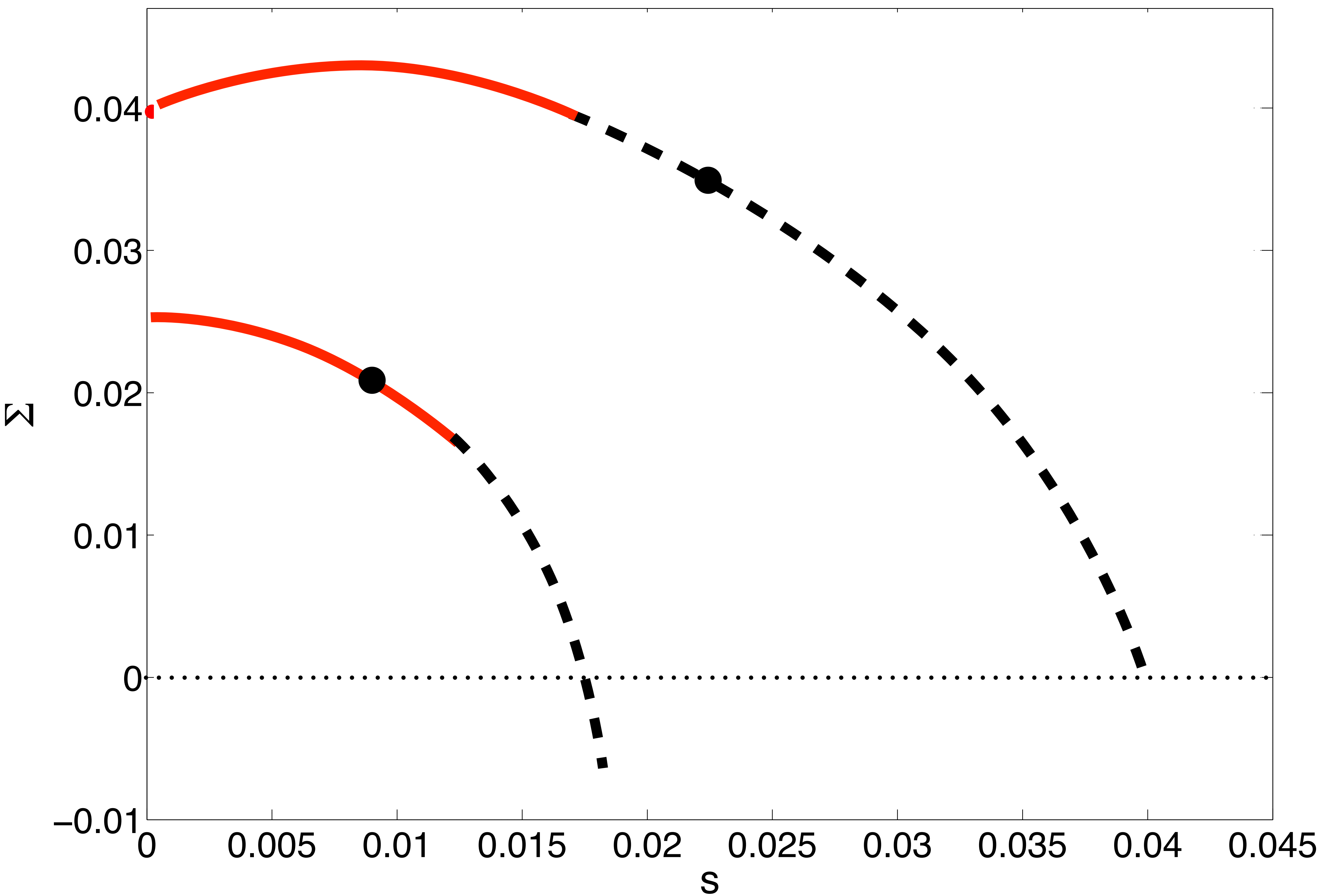
Frozen variables

Must take a specific value in a cluster
in order the formula to be SAT

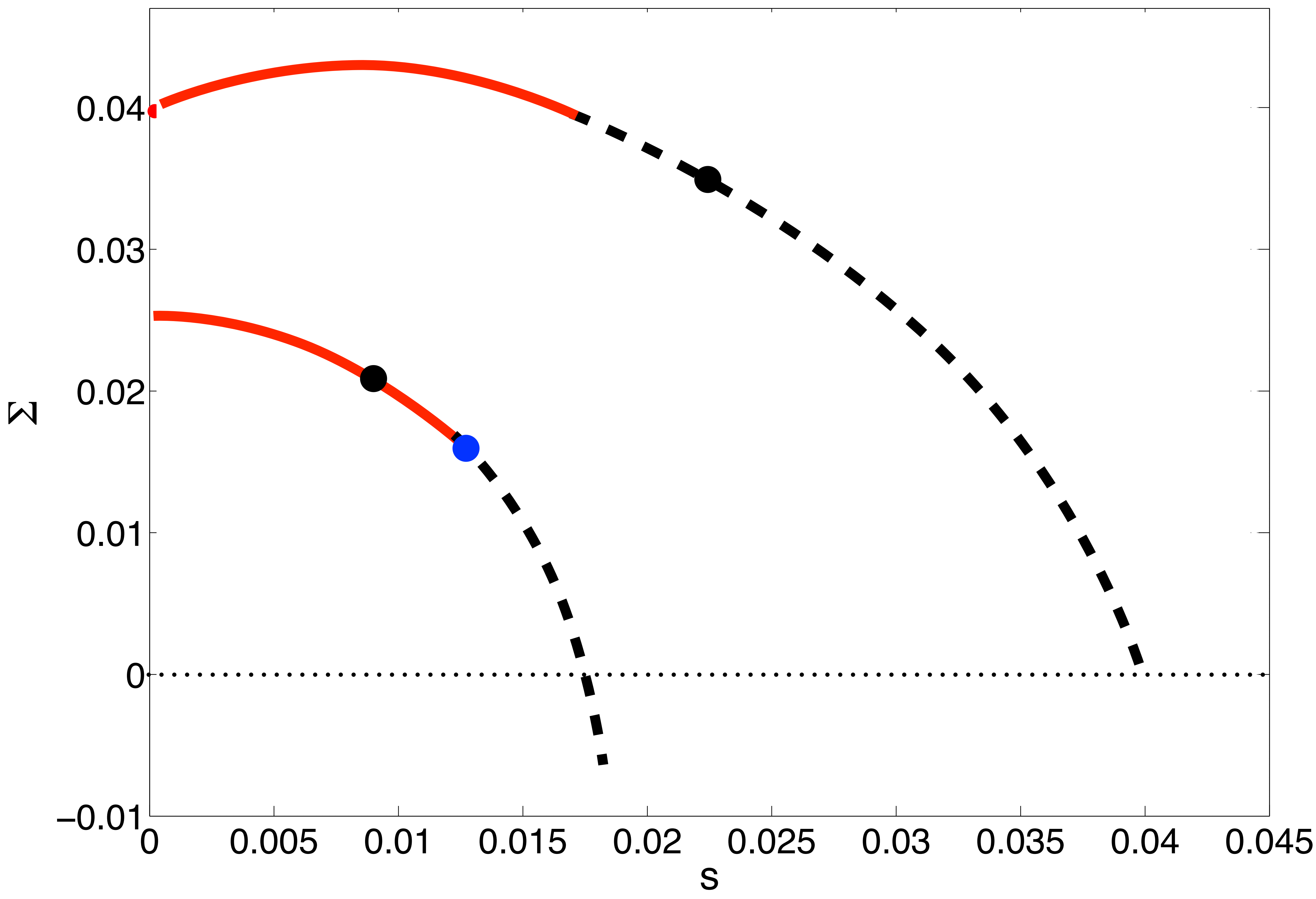


— dominating clusters have frozen variables

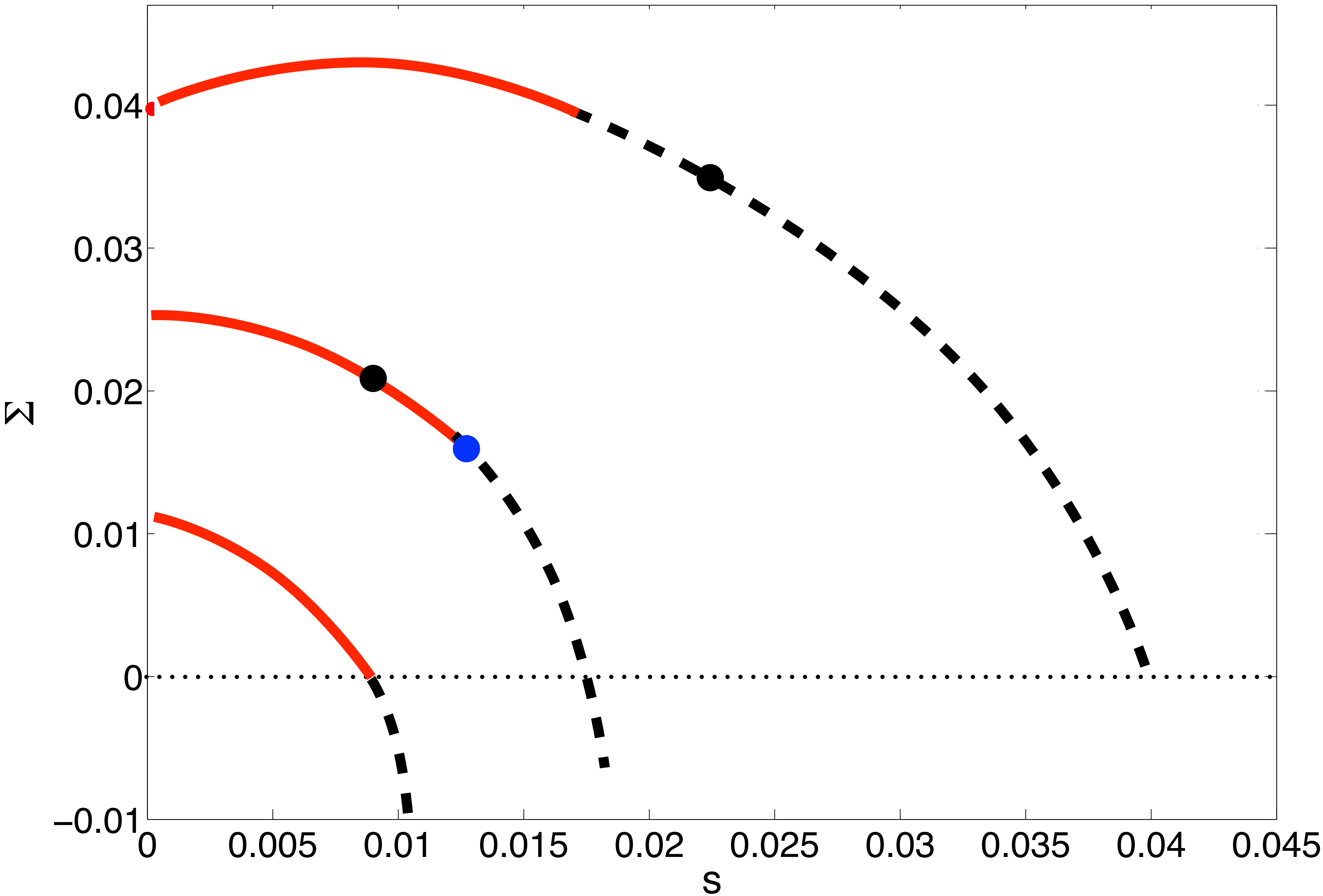
--- dominating clusters do not have frozen variables



0.04
0.03
0.02
0.01
0
−0.01
Σ
0
0.005
0.01
0.015
0.02
0.025
0.03
0.035
0.04
0.045
s



0.04
0.03
0.02
0.01
0
−0.01
Σ
0
0.005
0.01
0.015
0.02
0.025
0.03
0.035
0.04
0.045
s



0.04
0.03
0.02
0.01
0
−0.01
Σ
0
0.005
0.01
0.015
0.02
0.025
0.03
0.035
0.04
0.045
s

Summary of rigorous results
(random K-SAT)

Achlioptas, Coja-Oghlan, Ricci-Tersenghi, RSA '11

Theorem 2. *For every $k \geq 8$, there exists a value of $r < r_k$ and constants $\alpha_k < \beta_k < 1/2$ and $\epsilon_k > 0$ such that w.h.p. the set of satisfying assignments of $F_k(n, rn)$ consists of $2^{\epsilon_k n}$ nonempty cluster regions, such that*

1. *The diameter of each cluster region is at most $\alpha_k n$.*
2. *The distance between every pair of cluster-regions is at least $\beta_k n$.*

Theorem 3. *For any $0 < \delta < 1/3$, if $r = (1-\delta)2^k \ln 2$, then for all $k \geq k_0(\delta)$, Theorem 2 holds with*

$$\alpha_k = \frac{1}{k}, \quad \beta_k = \frac{1}{2} - \frac{5}{6}\sqrt{\delta}, \quad \epsilon_k = \frac{\delta}{2} - 3k^{-2}.$$

Theorem 8. *For every $k \geq 9$, there exists $c_k < r_k$ such that for all $r \geq c_k$, w.h.p. every cluster of $F_k(n, rn)$ has at least $(1 - 2/k) \cdot n$ frozen variables. As k grows,*

$$\frac{c_k}{2^k \ln 2} \to \frac{4}{5}.$$

What about <u>non-random</u> CSP?

- The locally tree-like topology is not strictly necessary
- Long range correlations and phase transitions are common to any high dimensional model
- The freezing of (random) subsets of variables in (random) directions can be the general driving mechanism for the onset of NP-hardness

What about <u>non-random</u> CSP?

- Can we identify strongly correlated subset of variables in a general model?
- Algorithmic problems related to short loops
- Loops corrections to mean-field approximations: Cluster Variational Methods (CVM), Generalized Belief Propagation (GBP), ...

What about the UNSAT phase?

- Clustering structure also in the UNSAT phase
- Succinct UNSAT certificates by uncovering frozen (or strongly correlated) variables
- Message-passing algorithms for determining the probability of being in the UNSAT certificate

Thanks !

References and more info can be found on web pages of

- me --> http://chimera.roma1.infn.it/FEDERICO
- Dimitris Achlioptas (UC Santa Cruz)
- Amin Coja-Oghlan (Univ. Warwick)
- Andrea Montanari (Stanford)
- Riccardo Zecchina (Politecnico Torino)