Subdominant Dense Clusters Allow for Simple Learning and High Computational Performance in Neural Networks with Discrete Synapses

Carlo Baldassi,[1,2,*] Alessandro Ingrosso,[1,2] Carlo Lucibello,[1,2] Luca Saglietti,[1,2] and Riccardo Zecchina[1,2,3]

[1]*Politecnico di Torino, Corso Duca degli Abruzzi 24, I-10129 Torino, Italy*
[2]*Human Genetics Foundation-Torino, Via Nizza 52, I-10126 Torino, Italy*
[3]*Collegio Carlo Alberto, Via Real Collegio 30, I-10024 Moncalieri, Italy*

We show that discrete synaptic weights can be efficiently used for learning in large scale neural systems, and lead to unanticipated computational performance. We focus on the representative case of learning random patterns with binary synapses in single layer networks. The standard statistical analysis shows that this problem is exponentially dominated by isolated solutions that are extremely hard to find algorithmically. Here, we introduce a novel method that allows us to find analytical evidence for the existence of subdominant and extremely dense regions of solutions. Numerical experiments confirm these findings. We also show that the dense regions are surprisingly accessible by simple learning protocols, and that these synaptic configurations are robust to perturbations and generalize better than typical solutions. These outcomes extend to synapses with multiple states and to deeper neural architectures. The large deviation measure also suggests how to design novel algorithmic schemes for optimization based on *local entropy* maximization.

In the past decades, various methods borrowed from statistical physics have been quite successful in studying the basic properties of neural-like systems [1]. A well known general result is that, as is the case for other optimization problems, training neural networks is qualitatively different if the variables—the synaptic weights—are constrained to take discrete values. Yet, the standard equilibrium analysis, which suggests that the solutions to the constrained problem would be inaccessible, is at odds with some recent heuristic algorithmic advances [2–5], which demonstrate that simple effective protocols may be devised at least in some simple scenarios. Therefore, it is conceptually important to understand the underlying reason for this discrepancy, which may be relevant for larger classes of problems.

Furthermore, the modulation of synaptic efficacy is the elementary computational step for biological information storage and for large scale machine learning architectures and neuromorphic devices. While most models assume that synapses are continuous, it is extremely important to understand the practical implications of constraining the synaptic states. Biological considerations and recent experimental evidence [6,7] suggest that synaptic efficacies store a few bits each (between 1 and 5). Machine learning applications (especially hardware implementations) could benefit from using simpler synaptic models and update

protocols, a research direction that is currently hampered by the difficulty of devising effective learning protocols.

The learning problem in neural networks with constrained synapses, even in its simplest formulation—the perceptron with N binary synapses—is known to be intractable in the worst case [8]. In the typical case, its equilibrium description is dominated in the large N limit by an exponential number (in N) of local minima [9–12], which easily trap standard search strategies based on free energy minimization, e.g. Monte Carlo algorithms [13,14] (a situation typical of spin glass phases, which is common to many hard random optimization problems [15–17]); moreover, the optimal synaptic configurations are typically geometrically isolated (i.e., they have mutual Hamming distances of order N), and thus are even harder to find for local search strategies [14].

Here, however, we show that the standard analysis does not capture the properties that are relevant for effective learning strategies. We introduce a novel large deviation analysis that reveals the existence of a different class of solutions, clustered in dense regions. These solutions have radically different properties from the dominating ones, and they are accessible to simple learning protocols. Numerical experiments support the results of this analysis, and show that the same picture also holds for complex neural architectures trained on real-world benchmarks. The analysis generalizes to the case of multilevel synapses.

In a more general sense, these findings highlight the key role that subdominant states have in understanding the practically relevant properties of a prototypical complex system. We have no reason to believe that this scenario is specific to this particular family of problems; in fact, we also show that an optimization strategy inspired by our

analysis is also effective on the random K-satisfiability (K-SAT) problem.

The model.—The single layer binary neural network (perceptron) maps vectors of N inputs $\xi \in \{-1, 1\}^N$ to binary outputs as $\tau(W, \xi) = \mathrm{sgn}(W \cdot \xi)$, where $W \in \{-1, 1\}^N$ is the vector of synaptic weights. Given αN input patterns ξ^μ with $\mu \in \{1, \ldots, \alpha N\}$ and their corresponding desired outputs $\sigma^\mu \in \{-1, 1\}^{\alpha N}$, and defining $\mathbb{X}_\xi(W) = \prod_{\mu=1}^{\alpha N} \Theta(\sigma^\mu \tau(W, \xi^\mu))$, where $\Theta(x)$ is the Heaviside step function, the learning problem is that of finding W such that $\tau(W, \xi^\mu) = \sigma^\mu$ for all μ, i.e., such that $\mathbb{X}_\xi(W) = 1$. The entries ξ_i^μ are independent and identically distributed (i.i.d.) unbiased random variables. There are two main scenarios of interest for the distribution of the desired outputs σ^μ: (1) the *classification* case, in which they are i.i.d. random variables, and (2) the *generalization* (or *teacher-student*) scenario, in which they are provided by a "teacher" device, i.e., another perceptron with synaptic weights W^T. In the classification scenario, the typical problem has a solution with probability 1 in the limit of large N up to $\alpha_c = 0.833$ [9], after which the probability of finding a solution drops to zero. α_c is called the *capacity*; we also use this term for the maximum value of α for which a solution can be found by a specific algorithm. In the teacher-student scenario, the problem has exponentially many solutions up to $\alpha_{TS} = 1.245$, after which there is a first-order transition and only one solution is possible: the teacher itself [1,10]. One additional quantity of interest in this scenario is the generalization error rate $p_e = (1/\pi) \arccos[(1/N) W \cdot W^T]$, which is the probability that $\tau(W, \xi^\star) = \tau(W^T, \xi^\star)$ when $\xi^\star$ is a previously unseen input.

The standard zero-temperature equilibrium analysis of this model is based on a probability measure defined by the partition function $Z_{eq} = \sum_{\{W\}} \mathbb{X}_\xi(W)$; the typical case is described by taking the quenched average $\langle \log(Z_{eq}) \rangle_\xi$ over the realizations of the patterns.

Effective learning algorithms.—Only a handful of heuristic algorithms are currently believed—based on numerical evidence—to be able to solve the classification problem and achieve a nonzero capacity in the limit of large N in a subexponential running time: reinforced Belief Propagation (BP) [2], reinforced Max-Sum [3], SBPI [4], and CP + R [5] (a brief description of each is provided in the Supplemental Material [18]). In the classification case, they achieve capacities between $\alpha \simeq 0.69$ and $\alpha \simeq 0.75$. They all share the property of being local and distributed, and have typical solving times that scale almost linearly with the size of the input. SBPI and CP + R additionally have extremely simple requirements (only employing finite discrete quantities and simple, local, and online update schemes), making them appealing for practical purposes and reasonably plausible candidates for biological implementations. A qualitatively similar scenario holds in the generalization case, where all these algorithms perform well except in a finite window $1 \lesssim \alpha \lesssim 1.5$ around α_{TS}.

These results are not captured by the standard spin glass theory; in particular, the effectiveness of the utterly simplified algorithms SBPI and CP + R is in striking contrast with a glassy energy landscape in which solutions are isolated.

Numerical experiments.—We investigated this issue numerically, and found evidence that, in fact, the solutions found by the algorithms are typically not isolated; rather, they belong (with high probability at large N) to large connected clusters of solutions. More precisely: (1) from a given solution $\tilde{W}$, a random walk process over neighboring configurations in the space of solutions can reach distances of order N from the starting point; (2) the number of solutions at a distance of order N from $\tilde{W}$ grows exponentially with N (this can be estimated from the analysis of the recurrence relations on the average growth factor of the number of solutions at varying distances, and using the random walk processes for sampling the local properties relevant to those relations).

Furthermore, we used the standard BP method on single problem instances to estimate the entropy of the solutions at varying distance (controlled via a Franz-Parisi potential [21]) from a reference solution $\tilde{W}$ obtained from a heuristic solver, and found that the results do not match the predictions of the equilibrium analysis [14], see Fig. 1.

Teacher-student case.—We also extended the equilibrium analysis [14] to the teacher-student scenario, and found that: (1) typical solutions are isolated for all values of α even when adding a nonzero stability constraint, as in the classification case; (2) the teacher device is isolated and indistinguishable from all other typical solutions except for the generalization error; and (3) the results of estimates obtained from BP are consistent with the analytical calculation when using the teacher as a reference point, but not when using a solution provided by a heuristic solver (see the inset in Fig. 1). Finally, the generalization error for solutions found algorithmically is lower than what would be expected for a typical solution (see Fig. 3).

Large deviation analysis.—These results indicate that calculations performed at thermodynamic equilibrium are effectively blind to the solutions found by the heuristic algorithms. Traditionally, in the context of replica theory, similar situations have been addressed by looking for subdominant states [22]. However, this is insufficient in the present case.

A different analytical tool is thus needed for obtaining a description of this regime, which—according to the numerical evidence—is characterized by regions with a high density of solutions. Clearly, the statistical weight of the individual solutions must be modified, by favoring the ones that are surrounded by a large number of other solutions. Therefore, we studied the following large-deviation free energy density function:

$$\mathcal{F}(d, y) = -\frac{1}{Ny} \log \left(\sum_{\{\tilde{W}\}} \mathbb{X}_\xi(\tilde{W}) \mathcal{N}(\tilde{W}, d)^y \right) \quad (1)$$
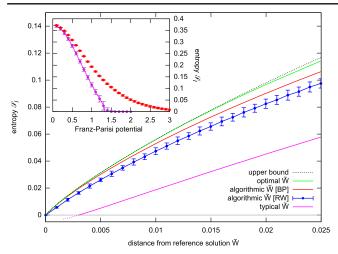
FIG. 1 (color online).   Numerical evidence of the existence of clusters of solutions. Entropy at a given distance from a reference solution $\tilde{W}$, in the classification case at $\alpha = 0.4$. From bottom to top: (magenta) theoretical prediction for a typical $\tilde{W}$; (blue) numerical estimate based on a random walks on connected solutions starting from one provided by SBPI, with $N = 1001$; (red) estimate from belief propagation using a solution from SBPI, with $N = 10001$; (green) theoretical curve for the optimal $\tilde{W}$ as computed from Eq. (1); and (dotted black) upper bound ($\alpha = 0$ case, all configurations are solutions). The random-walk points underestimate the number of solutions since they only consider single-flip-connected clusters; the BP curve is lower than the optimal curve because in the latter $\tilde{W}$ is optimized as a function of the distance, while in the former it is fixed. Inset: comparison between a typical solution and one found with SBPI, in the teacher-student case at $\alpha = 0.5$ with $N = 1001$. Larger potentials correspond to smaller distances. Top points (red): SBPI reference solution, with the entropy computed by BP; bottom curve (magenta): theoretical prediction for a typical solution; bottom points (purple): BP results using the teacher as reference.

where $\mathcal{N}(\tilde{W}, d) = \sum_{\{W\}} \mathbb{X}_\xi(W) \delta(W \cdot \tilde{W}, N(1 - 2d))$ counts the number of solutions W at normalized Hamming distance d from a reference solution $\tilde{W}$ (δ is the Kronecker delta symbol) and y has the role of an inverse temperature. This free energy describes a system in which each configuration $\tilde{W}$ is constrained to be a solution, and has a formal energy density $\mathcal{E}(\tilde{W}) = -(1/N) \log \mathcal{N}(\tilde{W}, d)$ that favors configurations surrounded by an exponential number of other solutions, with y controlling the amount of reweighting. The regions of highest local density are then described in the regime of large y and small d.

The relevant quantities are computed through the usual statistical physics tools; of particular importance is the entropy density of the surrounding solutions, the *local entropy*:

$$\mathcal{S}_I(d, y) = -\langle \mathcal{E}(\tilde{W}) \rangle_{\xi, \tilde{W}} = \frac{1}{N} \langle \log \mathcal{N}(\tilde{W}, d) \rangle_{\xi, \tilde{W}}, \quad (2)$$

which is simply given by $\mathcal{S}_I(d, y) = \partial_y(y\mathcal{F}(d, y))$. The signature for the existence of a dense and exponentially large cluster of solutions is that $\mathcal{S}_I(d, y) > 0$ in a

neighborhood of $d = 0$. Another important quantity is the *external entropy*, i.e., the entropy of the reference solutions $\mathcal{S}_E(d, y) = -y[\mathcal{F}(d, y) + \mathcal{S}_I(d, y)]$, which must also be non-negative.

The special case $y = 1$ is essentially equivalent to the computation of Ref. [11]; $\mathcal{S}_I(d, y)$ reduces to the computation à la Franz-Parisi of [14] in the limit $y \to 0$.

We computed Eq. (1) by the replica method in the replica-symmetric (RS) ansatz, resulting in an expression involving 13 order parameters to be determined by the saddle point method. The analytical expressions and the details of the computation are reported in the Supplemental Material [18]. It turns out that, for all values of α and d, there is a value of y beyond which $\mathcal{S}_E(d, y) < 0$, which is unphysical and signals a problem with the RS assumption. Therefore, we sought the value $y^\star = y^\star(\alpha, d)$ at which $\mathcal{S}_E(d, y^\star) = 0$, i.e., the highest value of y for which the RS analytical results are consistent. In the following, we thus drop the y dependency.

The solution to the system of equations stemming from the RS saddle point produces qualitatively very similar results for both the classification (with $\alpha < \alpha_c$) and the generalization (with $\alpha < \alpha_{TS}$) case. It displays a number of noteworthy properties (Fig. 2):

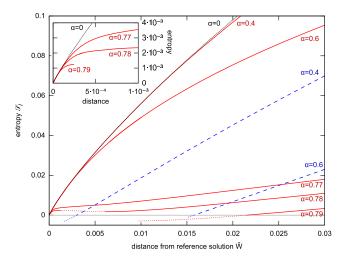(1) For all $\alpha < \alpha_c$, there is a neighborhood of $d = 0$ where $\mathcal{S}_I(d) > 0$, implying the existence of extensive

FIG. 2 (color online).   Large deviation analysis. Local entropy curves at varying distance d from the reference solution $\tilde{W}$ for various α (classification case). Black dotted curve, $\alpha = 0$ case (upper bound). Red solid curves, RS results from Eq. (1) (optimal $\tilde{W}$). Up to $\alpha = 0.77$, the curves are monotonic. At $\alpha = 0.78$, a region incorrectly described within the RS ansatz appears (dotted; geometric bounds are violated at the boundaries of the part of the curve with negative derivative). At $\alpha = 0.79$, the solution is discontinuous (a gap appears in the curve), and parts of the curve have negative entropy (dotted). Blue dashed curves, equilibrium analysis (typical $\tilde{W}$) [14] (dotted parts are unphysical): the curves are never positive in a neighborhood of $d = 0$. Inset: enlargement of the region around $d = 0$ (notice the solution for $\alpha = 0.79$, followed by a gap).

clusters of solutions. Furthermore, for all α, the curves for $S_I(d)$ are all approximately equal around $d = 0$; in particular, they all approximate the case for $\alpha = 0$ where all points are solutions. This implies that the clusters of solutions are extremely dense at their core. This is our chief result. The size of this dense region shrinks with α and vanishes at α_c.

(2) For large distances, as expected, $S_I(d)$ collapses with a second-order transition onto the equilibrium entropy; i.e., this regime is dominated by the typical solutions.

(3) Up to a certain α_U (where $\alpha_U \simeq 0.77$ in the classification case and $\alpha_U \simeq 1.1$ in the generalization case), the $S_I(d)$ curves are monotonic in d. Beyond α_U, there is a transition in which there appear regions of d (dotted in Fig. 2) that are not correctly described by the RS ansatz (since geometric bounds are violated; see the discussion in the Supplemental Material for details [18]), and must be described at a higher level of replica symmetry breaking (RSB). We speculate that this transition signals a change in the structure of the space of solutions: for $\alpha < \alpha_U$, the densest cores of solutions are immersed in a huge connected structure; for $\alpha > \alpha_U$, this structure fractures and the dense cores become isolated and hard to find.

(4) In the teacher-student scenario, the generalization properties of the optimal reference solutions $\tilde{W}$ are generally much better than those of typical solutions. This is clearly shown in Fig. 3, where we also show that the curve for small d is in striking agreement with that produced using solutions obtained from the SBPI algorithm. The
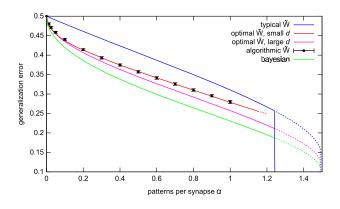

FIG. 3 (color online). Generalization error (teacher-student scenario). From top to bottom: (blue) typical solution, (red) optimal $\tilde{W}$ from Eq. (1) at small d (we used $d = 0.025$ for numerical reasons and since the curve is not sensitive to the precise value of d in this regime; this solution disappears after $\alpha \simeq 1.2$), (black points) solutions from SBPI at $N = 10001$, 100 samples per point, (magenta) optimal $\tilde{W}$ from Eq. (1) at the value of d for which S_I is maximum (i.e., it equals the equilibrium entropy), and (green) Bayesian case: error from the average over all solutions. At $\alpha_{TS} = 1.245$ there is the first-order transition to perfect learning; between α_{TS} and $\alpha = 1.5$ there is a metastable regime; the dashed parts of the curves correspond to unphysical solutions of the RS equations with negative entropy.

generalization error decreases monotonically when increasing d, and it saturates to a plateau when $S_I(d)$ becomes equal to the entropy of the typical solutions [see point (2) above].

We expect this qualitative and quantitative picture, especially for $\alpha \lesssim \alpha_U$, to be quite robust. First, these results are convincingly supported by our numerical findings, where available. Furthermore, a slightly simplified model analyzed at a higher level of RSB and at $y \to \infty$ [see Eq. (3) below] yields almost indistinguishable results.

The analytical computations are straightforwardly generalized to the case of multilevel synapses and sparse patterns, and the results are qualitatively identical [23].

Multilayer network.—These theoretical results seem to extend to more complex architectures and nonrandom learning problems. We observed this by heuristically extending the CP + R algorithm to multilayer classifiers with L possible output labels, and training these networks on the MNIST database benchmark [24], which consists of 7×10^4 grayscale images of hand-written digits ($L = 10$). A description of the architecture and of the learning algorithm is provided in the Supplemental Material [18].

We observed that it is indeed very easy to achieve perfect learning on the whole training data set, and that very good generalization errors can be reached (e.g., 1.25% with order 10^7 synapses) despite the binary nature of the synapses and the fact that we did not specialize the architecture for this particular data set. Moreover, we did not observe any overfitting: the generalization error does not degrade by reaching zero training error, or by using larger networks.

As for the perceptron, we performed a random-walk process in the space of solutions, with similar results: the simplified algorithm reaches a solution that is part of a dense, large connected cluster, and the generalization properties of the starting solution are better than those of solutions found in later stages of the random walk (see Fig. 1B in the Supplemental Material [18]).

Optimization.—We also studied a variant of the free energy (1) without the constraint on $\tilde{W}$:

$$\mathcal{F}_U(d, y) = -\frac{1}{Ny} \log \left(\sum_{\{\tilde{W}\}} \mathcal{N}(\tilde{W}, d)^y \right). \qquad (3)$$

The analysis in this case requires at least an additional step of RSB, and will be presented in detail in a follow-up work [25]. Still, the results are very close to those reported for the constrained scenario; furthermore, the probability that the $\tilde{W}$ in this system are a solution tends exponentially to 1 with $d \to 0$, despite the removal of the explicit constraint. This suggests that we can algorithmically exploit $\mathcal{F}_U(d, y)$ to efficiently sample ground states of the system, and that such a strategy could be applied to different optimization problems as well.

As the most straightforward proof of concept in this direction we have developed a Monte Carlo Markov Chain

algorithm, using the local entropy $\mathcal{E}(\tilde{W})$ as an objective function. We call such a procedure the Entropy-driven Monte Carlo (EdMC) procedure [25]. $\tilde{W}$ is initialized at random; at each step $\mathcal{E}(\tilde{W})$ is computed efficiently by the BP algorithm, which is expected to give good results at small d in dense regions; random local updates of $\tilde{W}$ are accepted or rejected using a standard Metropolis rule at fixed temperature y^{-1}. In addition to the binary perceptron problem, we applied the algorithm to another (radically different) problem in which standard simulated annealing (SA) methods are known to fail, namely, the famous random K-SAT problem. As expected from our analysis, for the perceptron learning problem EdMC does not suffer from trapping in local minima even at $y = \infty$, up to at least $\alpha \simeq 0.65$, with a running time that scales almost linearly with the size of the problem (whereas SA solving time diverges exponentially). For the random K-SAT problem we explored different regimes, and in particular one (in 4-SAT) where SA method is known to fail: in all cases, EdMC succeeds in an almost linear number of steps, outperforming SA method. Preliminary quantitative data are given in the Supplemental Material [18].

---

*carlo.baldassi@polito.it

[1] A. Engel and C. Van den Broeck, *Statistical Mechanics of Learning* (Cambridge University Press, Cambridge, England, 2001).

[2] A. Braunstein and R. Zecchina, Learning by message-passing in neural networks with material synapses, Phys. Rev. Lett. **96**, 030201 (2006).

[3] C. Baldassi and A. Braunstein, A max-sum algorithm for training discrete neural networks, J. Stat. Mech.: Theory Exp. (**2015**) P08008.

[4] C. Baldassi, A. Braunstein, N. Brunel, and R. Zecchina, Efficient supervised learning in networks with binary synapses, Proc. Natl. Acad. Sci. U.S.A. **104**, 11079 (2007).

[5] C. Baldassi, Generalization learning in a perceptron with binary synapses, J. Stat. Phys. **136**, 902 (2009).

[6] D. H. O'Connor, G. M. Wittenberg, and S. S.-H. Wang, Graded bidirectional synaptic plasticity is composed of switch-like unitary events, Proc. Natl. Acad. Sci. U.S.A. **102**, 9679 (2005).

[7] T. M Bartol, C. Bromer, J. P. Kinney, M. A. Chirillo, J. N. Bourne, K. M. Harris, and T. J. Sejnowski, Hippocampal spine head sizes are highly precise, bioRxiv, 2015.

[8] E. Amaldi, On the complexity of training perceptrons, Kohonen et al, pages 55–60, 1991.

[9] W. Krauth and M. Mézard, Storage capacity of memory networks with binary couplings, J. Phys. II (France) **50**, 3057 (1989).

[10] H. Sompolinsky, N. Tishby, and H. Sebastian Seung, Learning from examples in large neural networks, Phys. Rev. Lett. **65**, 1683 (1990).

[11] H. Huang, K. Y. Michael Wong, and Y. Kabashima, Entropy landscape of solutions in the binary perceptron problem, J. Phys. A **46**, 375002 (2013).

[12] T. Obuchi and Y. Kabashima, Weight space structure and analysis using a finite replica number in the ising percep-tron, J. Stat. Mech.: Theory Exp. (**2009**) P12014.

[13] H. Horner, Dynamics of learning for the binary perceptron problem, Z. Phys. B Condens. Matter **86**, 291 (1992).

[14] H. Huang and Y. Kabashima, Origin of the computational hardness for learning with binary synapses, Phys. Rev. E **90**, 052813 (2014).

[15] O. C. Martin, R. Monasson, and R. Zecchina, Statistical mechanics methods and phase transitions in optimization problems, Theor. Comput. Sci. **265**, 3 (2001).

[16] M. Mézard and A. Montanari, *Information, Physics, and Computation* (Oxford University Press, New York, 2009).

[17] C. Moore and S. Mertens, *The Nature of Computation* (Oxford University Press, New York, 2011).

[18] See Supplemental Material at http://link.aps.org/supplemental/10.1103/PhysRevLett.115.128101, which includes Refs. [19] and [20], for a brief description of the heuristic algorithms, details of the analytical calculations, details on the multilayer network, and preliminary data on Entropy-driven Monte Carlo results.

[19] G. Hinton, S. Osindero, and Y.-W. Teh, A fast learning algorithm for deep belief nets, Neural Comput. **18**, 1527 (2006).

[20] Y. Bengio, A. Courville, and P. Vincent, Representation learning: A review and new perspectives, Pattern Analysis and Machine Intelligence, IEEE Transactions on, **35**, 1798 (2013).

[21] S. Franz and G. Parisi, Recipes for metastable states in spin glasses, J. Phys. I (France) **5**, 1401 (1995).

[22] L. Dall'Asta, A. Ramezanpour, and R. Zecchina, Entropy landscape and non-Gibbs solutions in constraint satisfaction problems, Phys. Rev. E **77**, 031118 (2008).

[23] C. Baldassi, F. Gerace, A. Ingrosso, L. Saglietti, and R. Zecchina (to be published).

[24] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, Gradient-based learning applied to document recognition, Proc. IEEE **86**, 2278 (1998).

[25] C. Baldassi, A. Ingrosso, C. Lucibello, L. Saglietti, and R. Zecchina (to be published).