Journal of Statistical Physics, Vol. 104, Nos. 5/6, September 2001 (© 2001)

Retrieval Properties of Bidirectional Associative Memories

D. Gandolfo[1, 2, 4] **and L. Laanait,**[1, 3, 5]

September 22, 2000; revised April 20, 2001

We study a two-layer neural network made of N and $M(N)$ neurons, producing a two-way association search for a family of $p(N)$ patterns, where each pattern is a pair of two independent sub-categories of information having respectively N and $M(N)$ components. In terms of the ratio $\gamma = \lim_{N\to\infty} M(N)/N$, we study the retrieval capability of this network and show that there exists, at least, three regimes of association for which we determine the evolution of the threshold $\alpha_c(\gamma)$ of the storage capacity $\alpha = \lim_{N\to\infty} p(N)/N$.

KEY WORDS: Bidirectionality; storage capacity; categorized patterns; hamming distance.

1. INTRODUCTION

The Hopfield model [H] of associative memory has been thoroughly investigated in the past [A, AGS, N]. However this seemingly seductive approach to the modelization of the brain, acting as a learning and retrieval system, exhibits an obvious drawback in that it does not allow for a categorization of the stored informations. Indeed, our souvenirs for example are clearly made of different pieces of information, we record people by their names, physiognomy, social behaviors and so on, in such a way that any one of these characteristics may lead to the recollection of an individual.

Several attempts to take into account compound information in the Hopfield model have been proposed, either by a change of the synaptic

[1] Centre de Physique Théorique, CNRS, Luminy, F-13288 Marseille Cedex 9, France.
[2] PhyMat, Département de Mathématiques, UTV, F-83957 La Garde, France.
[3] On leave from Département de Physique, Ecole Normale Supérieure, Rabat, Morocco.
[4] E-mail: gandolfo@cpt.univ-mrs.fr
[5] E-mail: laanait@yahoo.fr

0022-4715/01/0900-1389$19.50/0

couplings between neurons or by a modification of the structure of the underlying state space.

Introduction of correlations between stored patterns [GTA] was among the first works in this direction. The Hopfield model with correlated patterns has received much attention lately both through numerical and mean field type approaches [GTA, CT] and also by rigorous methods [N, GLMR2, GLMR3]. It is worth mentioning that this generalization of the Hopfield model originated in neurophysiological experiments performed on a primate [M, SM] where it was observed that a structurally uncorrelated temporal sequence of patterns is converted into spatially correlated attractors in the monkey brain.

Another early proposal (Parga, Virasoro, 1986) to take these features into account was the introduction [PV] of a hierarchical tree-based structure to store the patterns, without imposing their *a priori* orthogonality. The method relies on the spin glass mean field theory and requires only a slight modification of the Hebb's learning rule. It is closer, in the neural science point of view, to the way our brain processes information [Kl] but requires more investigations concerning its storage capability. In [S] (Sourlas, 1988) introduced a multi-layer neural network for hierarchical patterns in which greater storage capacities than the common one, $\alpha_c \sim 0.14$, was predicted (see [H]).

A different hierarchical storage method was considered in [B] for a network of N neurons organized in L different modules, each comprising N_k neurons ($\sum_1^L N_k = N$). Each module codes for a specific part of the pattern information. In the case of a two-modules model, the author exhibits interesting mean field behaviors concerning the stability of the compound information stored. For example, the stability of the patterns increases when the fraction of category coding neurons decreases. Also, when damage is done to the synaptic junctions, there exists a threshold above which the network recall the category of the patterns but not its details.

In this article we are interested in the study of a different approach from the two cases above. This approach was proposed by (Kosko, 1988) in [K] and called a bidirectional associative memory (BAM). The aim here is to learn and retrieve a pair of associated patterns by a two-layer non linear neural network producing a two-way association search.

p patterns $\{\Xi^\mu\}_{\mu \in \{1, \ldots, p\}}$ are to be stored in a two layers network $(\boldsymbol{\sigma}, \boldsymbol{\tau})$ of respectively N and M neurons, $\boldsymbol{\sigma} = \{\sigma_i\}_{i \in \{1, \ldots, N\}}$, $\boldsymbol{\tau} = \{\tau_j\}_{j \in \{1, \ldots, M\}}$ where no synaptic junctions exist between neurons on a same layer. Each pattern Ξ^μ is a pair (ξ^μ, η^μ), where ξ^μ is an N-vector $(\xi_1^\mu, \ldots, \xi_N^\mu)$ of i.i.d. $(\mathbb{Z}_2)$-r.v. to be stored in the $\boldsymbol{\sigma}$-layer and likewise, η^μ is an M-vector $(\eta_1^\mu, \ldots, \eta_M^\mu)$ of i.i.d. $(\mathbb{Z}_2)$-r.v. to be stored in the $\boldsymbol{\tau}$-layer. ξ^μ and η^μ coding for different pieces (categories) of the same information.

In [K] a Hebb learning rule is used which updates successively the configurations in each layer according to the value of the neurons states of the other layer. The simulation of a small network ($N = 15$, $M = 10$, $p = 4$) has lead to the conclusion that, activating the dynamic of the network with an input configuration $(\boldsymbol{\sigma}, \boldsymbol{\tau})$ where, either $\boldsymbol{\sigma}$ or $\boldsymbol{\tau}$ is "close enough" from one of the categories of a stored pattern, say Ξ^{μ_0}, then the network quickly evolves to a stable fixed point corresponding to the perfect retrieval of Ξ^{μ_0}. This phenomenon is called "two-patterns reverberation" by the author. Moreover, with the same data as above, Kosko observed that if the network dynamics is activated with an initial configuration $(\boldsymbol{\sigma}, \boldsymbol{\tau}) = (\xi^{\mu_i}, \eta^{\mu_j})$, $i \neq j$ then the retrieved state corresponds to the pattern $(\xi^{\mu_i}, \eta^{\mu_i})$. This is explained in [K] by claiming that the fixed point represents a system energy local minimum and by the fact that N is larger than M. Nevertheless, the relative magnitude of M and N was not clearly elucidated there.

In [KPS] the phase diagram of a stochastic version of this model is investigated through mean field theory and the replica method, following ideas from [AGS]. These authors observe that the storage capacity increases with respect to the ratio M/N, $N \geqslant M$ and get a threshold storage capacity less than $\alpha_c \sim 0.14$ (see [H], [AGS]). However, referring to the mean field computations in [S], one would expect higher threshold storage capacities.

We have studied rigorously the BAM model in the limit: $\lim_{N \to \infty} M(N)/N = \gamma \in (0, 1]$ for the following three regimes of association (hereafter denoted $\mathscr{R}_I$, $\mathscr{R}_{II}$, $\mathscr{R}_{III}$):

- in $\mathscr{R}_I$ (resp. $\mathscr{R}_{III}$), a retrieved configuration $(\boldsymbol{\sigma}, \boldsymbol{\tau})$ is such that $\boldsymbol{\sigma}$ (resp. $\boldsymbol{\tau}$) is almost perfectly aligned with one of the patterns, say ξ^μ (resp. η^μ) and $\boldsymbol{\tau}$ (resp. $\boldsymbol{\sigma}$) only partially aligned with the corresponding η^μ (resp. ξ^μ), $\mu = 1, \ldots, p$,
- in $\mathscr{R}_{II}$, a retrieved configuration $(\boldsymbol{\sigma}, \boldsymbol{\tau})$ is such that both $\boldsymbol{\sigma}$ and $\boldsymbol{\tau}$ are only partially aligned with, respectively, η^μ and ξ^μ, $\mu = 1, \ldots, p$.

Among the results, we get that there exists a finite threshold capacity $\alpha_c(\gamma)$ satisfying

- in $\mathscr{R}_I$, for $\gamma > \gamma_0 \sim 0.3$, $\alpha_c(\gamma)$ decreases from $\alpha_c(\gamma_0) \sim 0.076$ to $\alpha_c(\gamma = 1) \sim 0.056$,
- in $\mathscr{R}_{II}$, $\alpha_c(\gamma)$ increases and then decreases as a function of γ in the interval $(0, 1]$,
- in $\mathscr{R}_{III}$, $\alpha_c(\gamma)$ is an increasing function of γ in the interval $(0, 1]$.

$\alpha_c(\gamma = 1)$ is quite close to the Newman value 0.056 for the standard Hopfield model [N].

2. DEFINITIONS AND RESULTS

The patterns (ξ^μ, η^μ); $\mu \in \{1, \ldots, p(N)\}$ to be stored in the network are chosen in the configuration space of the activity states of the $N+M(N)$ neurons $(\boldsymbol{\sigma}, \boldsymbol{\tau}) \in \{\pm 1\}^N \otimes \{\pm 1\}^M$, organized as a two layers neural network viewed as a complete bipartite graph.

The learning of the pattern configurations runs through a non-linear mechanism involving weighted synaptic junctions between neurons $i \in \Lambda_N \equiv \{1, \ldots, N\}$ on layer $\boldsymbol{\sigma}$ and $j \in \Lambda_M \equiv \{1, \ldots, M\}$ on layer $\boldsymbol{\tau}$, whose expressions are given by

$$J_{ij} = \sum_{\mu=1}^{p} \xi_i^\mu \eta_j^\mu \tag{2.1}$$

The synaptic matrix (2.1) allows to define an energy function for this system [K]

$$H_{N,M}(\boldsymbol{\sigma}, \boldsymbol{\tau}) = -\frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{M} J_{ij} \sigma_i \tau_j \tag{2.2}$$

We introduce the local overlaps of the configuration $(\boldsymbol{\sigma}, \boldsymbol{\tau})$ with the pattern (ξ^μ, η^μ) as

$$\begin{cases} \mathfrak{S}_N^\mu(\boldsymbol{\sigma}) = \dfrac{1}{N} \displaystyle\sum_{i=1}^{N} \sigma_i \xi_i^\mu \\ \mathfrak{T}_M^\mu(\boldsymbol{\tau}) = \dfrac{1}{M} \displaystyle\sum_{j=1}^{M} \tau_j \eta_j^\mu \end{cases}, \qquad \mu = 1, \ldots, p \tag{2.3}$$

so that the energy function takes the form

$$H_{N,M}(\boldsymbol{\sigma}, \boldsymbol{\tau}) = -M(N) \sum_{\mu=1}^{p} \mathfrak{S}_N^\mu(\boldsymbol{\sigma})\, \mathfrak{T}_M^\mu(\boldsymbol{\tau}) \tag{2.4}$$

Obviously, the configuration $(\boldsymbol{\sigma}^0, \boldsymbol{\tau}^0)$ given by $\sigma_i^0 = \xi_i^\mu$, $i \in \Lambda_N$ and $\tau_j^0 = \eta_j^\mu$, $j \in \Lambda_M$ minimizes the energy (2.4).

The information stored in the network is carried by synaptic junctions between neurons on the two layers and the stable configurations around the local minima $(\boldsymbol{\sigma}^0, \boldsymbol{\tau}^0)$ of the energy function, are called *attractors*. We define the storage capacity α as

$$\alpha = \lim_{N \to \infty} p(N)/N \tag{2.5}$$

The results derived in [K] and further investigated in [KPS] suggest that the retrieved configurations are fixed points located near each stored pattern $\Xi^\mu = (\xi^\mu, \eta^\mu)$ and, furthermore, it was noticed in [K] that this retrieval is quite robust against damages (i.e. errors) imposed on the components of one or both pattern categories.

In order to state our result, some definitions are necessary :

Definition 2.1. Given two configurations $\boldsymbol{\sigma}$, $\boldsymbol{\sigma}'$ in $\mathbb{Z}_2^N$, we call $d(\boldsymbol{\sigma}, \boldsymbol{\sigma}')$ their Hamming distance. The sphere of radius $\delta_1 \in (0, \frac{1}{2})$ centered on $\boldsymbol{\sigma}$ is defined by:

$$S(\boldsymbol{\sigma}, \delta_1) = \{\boldsymbol{\sigma}' \in \mathbb{Z}_2^N : d(\boldsymbol{\sigma}, \boldsymbol{\sigma}') = \lfloor \delta_1 N \rfloor\}$$

where $\lfloor \delta_1 N \rfloor$ is the greatest integer less that or equal to $\delta_1 N$ (denoted $\delta_1 N$ hereafter).

Definition 2.2. Let $\xi^\mu \in \mathbb{Z}_2^N$, we call $\boldsymbol{\sigma}^{J_1}$ the configuration whose components disagree with those of ξ^μ on the set $J_1 \subset \Lambda_N$

$$(\boldsymbol{\sigma}^{J_1})_i = \begin{cases} -\xi_i^\mu : i \in J_1 \\ \xi_i^\mu : i \in J_1^c = \Lambda_N \setminus J_1 \end{cases} \tag{2.6}$$

where $|J_1| = \delta_1 N$, $\delta_1 \in (0, \frac{1}{2})$. Likewise for the configuration $\tau^{J_2} \in S(\eta^\nu, \delta_2)$, $\eta^\nu \in \mathbb{Z}_2^M$, $J_2 \subset \Lambda_M : |J_2| = \delta_2 M$, $\delta_2 \in (0, \frac{1}{2})$.

Definition 2.3. The configuration $(\boldsymbol{\sigma}, \tau)$ is said to be stable if there exists $\epsilon > 0$, $\delta_1 > 0$ and $\delta_2 > 0$ such that:

$$H_{N,M}(\boldsymbol{\sigma}, \tau) \leqslant \min_{\substack{\boldsymbol{\sigma}' \in S(\boldsymbol{\sigma}, \delta_1 \\ \tau' \in S(\tau, \delta_2)}} H_{N,M}(\boldsymbol{\sigma}', \tau') - \epsilon\, M(N) \tag{2.7}$$

Recall that $\gamma = \lim_{N \to \infty} M(N)/N \in (0, 1]$. Our main result is contained in the following

Theorem 2.1. For $\gamma > \gamma_0 > 0$ there exists strictly positive thresholds $\alpha_c(\gamma)$ so that for any $\alpha < \alpha_c(\gamma)$, one can find $\delta_1 \in (0, \frac{1}{2})$, $\delta_2 \in (0, \frac{1}{2})$ and some $\epsilon > 0$ (each depending on α) so that a strictly positive function $\Gamma(\gamma)$ can be found such that:

$$\mathbb{P}_{\xi,\eta}\left[\bigcap_{1 \leqslant \mu \leqslant p} \{H_{N,M}(\xi^\mu, \eta^\mu) - \min_{\substack{\boldsymbol{\sigma} \in S(\xi^\mu, \delta_1) \\ \tau \in S(\eta^\mu, \delta_2)}} H_{N,M}(\boldsymbol{\sigma}, \tau)\} < -\epsilon\, M(N)\right]$$

$$\underset{\text{N large}}{\geqslant} 1 - e^{-\Gamma(\gamma)N} \tag{2.8}$$

Remark 2.1. For $\gamma > \gamma_0 > 0$, the bound we obtain for the probability $\mathbb{P}_{\xi,\eta}$ implies, by applying the Borel–Cantelli Lemma that, if $(\xi^\mu, \eta^\mu)_{\mu \in \{1, \ldots, p\}}$ is a sequence of patterns such that $\alpha < \alpha_c(\gamma)$ then with probability 1, the event

$$\{H_{N,M}(\xi^\mu, \eta^\mu) \leqslant \min_{\substack{\sigma \in S(\xi^\mu, \delta_1) \\ \tau \in S(\eta^\mu, \delta_2)}} H_{N,M}(\sigma, \tau) - \epsilon\, M(N)\} \text{ occurs i. o.}$$

Comments on the theorem. Let $\delta_{1c}(\alpha_c(\gamma))$, $\delta_{2c}(\alpha_c(\gamma))$ the upper bounds on δ_1 and δ_2 as stated in the theorem. Introducing constants c_0 and c_1 such that $0 < c_0 < c_1 < 1$ then the regimes $\mathscr{R}_I$, $\mathscr{R}_{II}$, $\mathscr{R}_{III}$ defined in Section 1 are such that

- $\mathscr{R}_I$ is characterized by $0 < \delta_{1c} < c_0\, \delta_{2c}$. In this domain, the threshold capacity $\alpha_c(\gamma)$ decreases from $\alpha_c(\gamma_0 \sim 0.3) \sim 0.076$ to $\alpha_c(\gamma = 1) \sim 0.056$ (see Fig. 1).
- $\mathscr{R}_{II}$ is such that $c_0\, \delta_{2c} < \delta_{1c} < c_1\, \delta_{2c}$. There, the threshold capacity $\alpha_c(\gamma)$ increases and then decreases in the interval (0, 1].
- $\mathscr{R}_{III}$ corresponds to $0 < c_1\, \delta_{2c} < \delta_{1c}$. Here, the threshold capacity $\alpha_c(\gamma)$ increases in the interval (0, 1] (see Fig. 1).

Proof. We only outline the main ideas of the proof of Theorem 2.1 as it follows closely the techniques initiated in [N]. See [FMP, BG] for

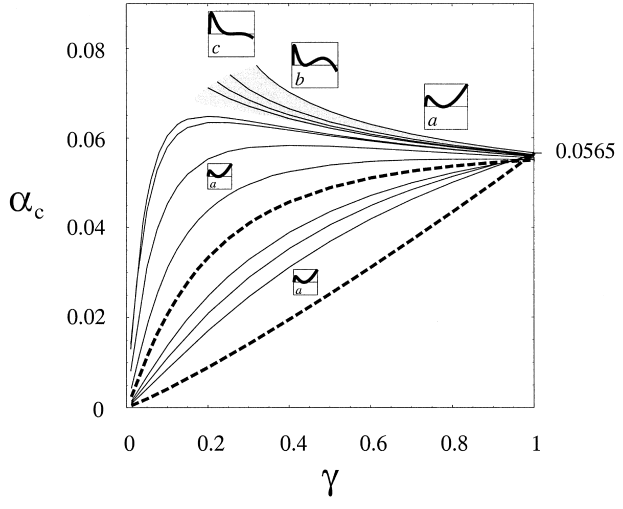

Fig. 1. Threshold capacity $\alpha_c(\gamma)$. The different regimes of behaviour of the BAM are outlined. $\mathscr{R}_I$: upper shaded part, $\mathscr{R}_{II}$: intermediate domain, $\mathscr{R}_{III}$: lower part bounded by dotted lines. In embedded windows: minima of $G(t_{\min}, \delta_c(\alpha_c(\gamma)))$.

other straightforward applications to the Potts–Hopfield and diluted Hopfield models, [L] for an extension to correlated patterns and also [Lou, T] for further improvements of Newman's result. Let us mention that very recently, using Fourier transform techniques, a new rigourous bound ($\alpha_c \simeq 0.113$) on the critical storage capacity of the Hopfield model has been found [FST].

In order to estimate the probability appearing in Eq. (2.8), on uses large deviation theory. The proof amounts at evaluating the difference in energy between a pattern (ξ^μ, η^μ) and the configurations $(\boldsymbol{\sigma}, \boldsymbol{\tau})$ located on spheres $S(\xi^\mu, \delta_1)$, $S(\eta^\mu, \delta_2)$ centered on this pattern.

This leads (using exponential Chebyshev–Markov inequality) to expectations with respect to Gaussian random variables, through the use of the law of large numbers and central limit theorem. One arrives eventually at the following bound for formula (2.8)

$$\mathbb{P}_{\xi,\eta}\left[\bigcap_{1\leqslant\mu\leqslant p}\{H_{N,M}(\xi^\mu,\eta^\mu)-\min_{\substack{\sigma\in S(\xi^\mu,\delta_1)\\ \tau\in S(\eta^\mu,\delta_2)}} H_{N,M}(\boldsymbol{\sigma},\boldsymbol{\tau})\}<-\epsilon\, M(N)\right]$$
$$\geqslant 1-\inf_{t\geqslant 0}\{e^{-NG(\alpha,\gamma,\delta_1,\delta_2)}\} \tag{2.9}$$

where

$$G(\alpha,t,\gamma,\delta_1,\delta_2)=-t\frac{\sqrt{\gamma}}{2}[\epsilon-(1-(1-2\delta_1)(1-2\delta_2))]$$
$$+\delta_1\log\delta_1+(1-\delta_1)\log(1-\delta_1)$$
$$+\gamma(\delta_2\log\delta_2+(1-\delta_2)\log(1-\delta_2))$$
$$+\frac{\alpha}{2}[\log(1-(1-\delta_1)\,\delta_2\,t^2)+\log(1-(1-\delta_2)\delta_1\,t^2)] \tag{2.10}$$

Then, for each positive $\gamma\in(0,1]$, one has to find a threshold capacity $\alpha_c(\gamma)$ and upper bounds $\delta_1(\alpha_c,\gamma)$ and $\delta_2(\alpha_c,\gamma)$ for δ_1 and δ_2, $\alpha<\alpha_c(\gamma)$, such that the function $G(\alpha,t,\gamma,\delta_1,\delta_2)$ reaches its minimum (in t) : $\Gamma(\gamma)$, which is nothing but the strictly positive function appearing in Theorem 2.1. This is completed numerically, as in [N]. The threshold capacities $\alpha_c(\gamma)$ are plotted in Fig. 1.

CONCLUSION

This study of the stability of the attractors in Bidirectional Associative Memories reveals that around any of the local energy minima $\Xi^\mu=(\xi^\mu,\eta^\mu)$,

$\mu = 1, \ldots, p$ (made of the two categories of information ξ^μ and η^μ), the BAM network retrieves, among others, the particular configurations $(\boldsymbol{\sigma}, \boldsymbol{\tau})$ where $\boldsymbol{\sigma}$ (resp. $\boldsymbol{\tau}$) agrees with the category ξ^μ (resp. η^μ) and the other $\boldsymbol{\tau}$ (resp. $\boldsymbol{\sigma}$) disagrees with the category η^μ (resp. ξ^μ) on a set of components whose cardinality depends on the threshold capacity.

It means that the bidirectionality allows to retrieve perfectly one of the categories for γ greater than a certain threshold γ_0. We recall that in the case of the standard (unidirectional) Hopfield model, the retrieval is perfect only when the storage capacity vanishes as N goes to infinity (see [A]).

ACKNOWLEDGMENTS

We are grateful to A. Messager and J. Ruiz for helpful discussions on related works. One of us (L.L.) is indebted to CPT–CNRS, Marseilles for kind hospitality. The support of CNR–Morocco under Grant: Physique nbr 037 is gratefully acknowledged.

REFERENCES

[A] D. J. Amit, Cambridge University Press, Cambridge, 1989.
[AGS] D. J. Amit, H. Gutfreund, and H. Sompolinsky, *Phys. Rev. A* **32**:1007 (1985).
[AT] D. J. Amit and M. Tsodyks, *Networks* **5**:1 (1994).
[B] N. Brunel, *J. Phys. I, France* **3**:1693 (1993).
[BG] A. Bovier and V. Gayrard, *J. Stat. Phys.* **72**:79 (1993).
[CT] L. F. Cugliandolo and M. Tsodyks, *J. Phys. A: Math. Gen.* **27**:741 (1993).
[FMP] P. A. Ferrari, S. Martinez, and P. Picco, *J. Stat. Phys.* **66**:1643 (1992).
[FST] J. Feng, M. Shcherbina, and B. Tirozzi, *Comm. Math. Phys.* **216**:139 (2001).
[GLMR1] D. Gandolfo, L. Laanait, A. Messager, and J. Ruiz, *Physica A* **264**:305 (1999).
[GLMR2] D. Gandolfo, L. Laanait, A. Messager, and J. Ruiz, in *Mathematical Results in Statistical Mechanics* (World Scientific, Marseille, 1999).
[GLMR3] D. Gandolfo, L. Laanait, A. Messager, and J. Ruiz, submitted to *J. Stat. Phys.*
[GTA] M. Griniasty, M. Tsodyks, and D. J. Amit, *Neural Comput.* **5**:1 (1993).
[H] J. J. Hopfield, *Proc. Nat. Acad. Sci. USA* **79**:2554 (1982).
[K] B. Kosko, *IEEE Transactions on Systems, Man and Cybernetics* **18**:49 (1988).
[Kl] R. L. Klatzky, Human Memory, W. H. Freeman, San Francisco, 1975.
[KPS] B. Kurchan, L. Peliti, and M. Saber, *J. Phys. I, France* **4**:1627 (1994).
[L] M. Löwe, *Ann. Appl. Prob.* **8**(4):1216 (1998).
[Lou] D. Loukianova, *C. R. Acad. Sci. Paris, Série I Math.* **318**(2):157 (1994).
[M] Y. Miyashita, *Nature* **335**:817 (1988).
[N] C. M. Newman, *Neural Networks* **1**:223 (1988).
[PV] N. Parga and M. A. Virasoro, *J. Phys. I, France* **47**:1857 (1986).
[S] N. Sourlas, *Europhys. Lett.* **7**:749 (1988).
[SM] K. Sakai and Y. Miyashita, *Nature* **354**:152 (1991).
[T] M. Talagrand, *Prob. Theor. Relat. Fields* **110**(2):177 (1998).