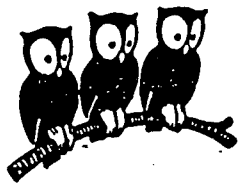
Physique statistique des réseaux de neurones et de l'optimisation combinatoire

Werner Krauth

ORSAY
n° d'ordre :

DEPARTEMENT DE PHYSIQUE
DE L'ECOLE NORMALE SUPERIEURE

UNIVERSITE DE PARIS-SUD
CENTRE D'ORSAY



THESE

présentée
Pour obtenir
Le Titre de DOCTEUR EN SCIENCE

par

M. Werner KRAUTH

SUJET : Physique statistique des réseaux de neurones et de l'optimisation combinatoire

Soutenue le 14 juin 1989 devant la Commission d'examen

M. Gérard TOULOUSE	Président
M. Bernard DERRIDA	Rapporteur
M. Nicolas SOURLAS	Rapporteur
M. Hendrik Jan HILHORST	
M. Marc MEZARD	Invité

DEPARTEMENT DE PHYSIQUE
DE L'ECOLE NORMALE SUPERIEURE

UNIVERSITE DE PARIS-SUD
CENTRE D'ORSAY

THESE

présentée
Pour obtenir
Le Titre de DOCTEUR EN SCIENCE

par
M. Werner KRAUTH

SUJET : Physique statistique des réseaux de neurones et de l'optimisation combinatoire

Soutenue le 14 juin 1989 devant la Commission d'examen

M. Gérard TOULOUSE	Président
M. Bernard DERRIDA	Rapporteur
M. Nicolas SOURLAS	Rapporteur
M. Hendrik Jan HILHORST	
M. Marc MEZARD	Invité

REMERCIEMENTS

Cette thèse a été effectuée au Département de Physique de l'Ecole Normale Supérieure, dont je voudrais remercier le directeur, Edouard Brézin, et l'ensemble de ses membres. J'ai fait partie du Laboratoire de Physique Statistique dès son début en 1988, et j'ai bénéficié d'échanges fructueux avec le directeur, Pierre Lallemand, et la plupart de ses membres.

Mon premier contact avec l'ENS date du mois de juin 1986, à l'occasion de la conférence de Heidelberg sur les verres de spin. Là, j'ai osé parler à Gérard Toulouse de mon projet de faire une thèse en France. Je lui suis profondément reconnaissant de m'avoir écouté, et ensuite d'avoir accepté d'être mon Directeur de thèse au Département de Physique de l'ENS. Tout au long de mon parcours j'ai beaucoup estimé sa confiance et ses conseils.

Si j'ai eu la rare chance de progresser pendant mes années de thèse à un rythme égal à celui de mes premières années de physique à l'Université, je le dois en premier lieu à Marc Mézard. C'est lui qui a eu la charge de guider mes recherches et de m'initier aux problèmes de la physique des systèmes désordonnés. J'ai toujours estimé son grand talent, sa gentillesse et sa patience.

J'ai également entretenu les relations fructueuses et amicales avec Jean-Pierre Nadal, avec lequel j'ai eu des discussions quotidiennes pendant une longue période.

Tous les travaux présentés dans cette thèse sont les fruits de collaborations avec M. Mézard, J.-P. Nadal et M. Opper. Un grand nombre de résultats, auxquels je me refère dans la thèse comme 'nos résultats', sont l'aboutissement de leurs idées. Qu'ils voient dans cette thèse un signe de ma reconnaissance pour leur apport - souvent décisif.

J'ai aussi eu la chance de bénéficier de nombreux échanges

scientifiques avec B. Derrida, E. Gardner, H. Gutfreund, D. d'Humières, N. Sourlas, J. Vannimenus, G. Weisbuch et J. Yedidia. Je remercie H. J. Hilhorst pour avoir accepté de faire partie de mon jury de thèse et B. Derrida et N. Sourlas pour m'avoir fait l'honneur d'en être les rapporteurs.

Je tiens aussi à remercier Cécile Combier, la responsable du centre de calcul du Département de Physique, pour des relations sans faille, et pour ses conseils toujours bénéfiques. Je salue ici la sagesse du Département de Physique de s'être doté de moyens considérables pour le calcul vectoriel.

Enfin, je voudrais remercier J. Vannimenus pour son aide dans la recherche d'un logement, S. Großmann pour ses bons offices auprès de l'ENS et les organisations de bourse, I. Gazan et M. Manceau pour le tirage de cette thèse, et les fondations C. Duisberg Stiftung et Studienstiftung des Deutschen Volkes, pour m'avoir permis de poursuivre mes études.

AVANT-PROPOS

Cette thèse s'inscrit dans le cadre d'une 'industrie de pointe' essayant de pousser plus loin les limites de la physique statistique.

En physique statistique, les modèles abondent. Ils sont parfois assez réalistes (comme le modèle de Heisenberg pour les aimants), parfois caricaturaux (comme le modèle d'Ising), souvent très loin de la réalité physique (caricatures de caricatures). Un exemple de la dernière classe est le modèle gaussien de Berlin et Kac [1]. Les auteurs en disent: '*It is irrelevant that the models may be far removed from physical reality if they can illuminate some of the complexities of the transition phenomena*'[1]. Une universalité sous-jacente (une sorte de continuité dans un espace hypothétique de modèles) lie les modèles entre eux ; elle rend pertinente l'étude des systèmes simples. Dans le cas du modèle de Berlin et Kac, d'ailleurs, la compréhension spectaculaire de cette universalité est venue d'une révolution bien postérieure, la théorie du groupe de renormalisation.

Ces dernières années un grand intérêt a été porté à des modèles collectifs dans d'autres sciences que la physique, notamment en économie, en neurobiologie, ou en informatique. Nous avons, pour cette thèse, travaillé sur un de ces problèmes - les réseaux de neurones - qui

peuvent être vus comme des modèles simples du comportement du cerveau, ou comme modèles d'une classe d'ordinateurs parallèles.

Le système particulier que nous appellerons 'modèle de Hopfield' consiste en un grand nombre de processeurs extrêmement simplifiés sans mémoire et avec une seule instruction. Chaque processeur interagit avec les autres par des couplages plus ou moins forts et se met à jour en fonction de *la somme* de tous les signaux reçus. Le modèle a surtout été étudié comme modèle d'une mémoire associative [2]. Par des modifications successives des couplages il est possible (dans un sens qui sera défini par la suite) de stocker un certain nombre de 'patterns' dans le réseau. Après cet *apprentissage*, le réseau se *rappelle* les 'patterns' et leur *associe* des états du réseau qui sont peu différents de l'un d'entre eux.

Ce modèle, qui est par nature collectif et parallèle, offre une vue d'un processus de calcul antithétique de celle de la célèbre machine de Turing [3] qui, elle, est par construction séquentielle. Cette vue est aussi beaucoup plus proche de ce que nous imaginons être le fonctionnement de notre cerveau.

S'il existe une correspondance exacte entre la machine de Turing et tout ordinateur sériel [4], les réseaux de neurones de nos jours sont davantage des modèles dans le sens de la physique statistique indiqué au début, c'est-à-dire allégoriques, potentiellement capables d'éclaircir les complexités du cerveau ou d'un futur ordinateur cellulaire, capables de nous faire concevoir de nouvelles idées générales. J'espère que le lecteur s'en apercevra à travers l'étude détaillée présentée dans ce travail.

Pour ce texte j'ai essentiellement reproduit les publications auxquelles j'ai été associé durant mon travail de thèse ; je les ai en outre augmentées d'introductions et de commentaires.

Le texte est organisé en deux chapitres de volumes inégaux, dont le premier regroupe les travaux sur les réseaux de neurones, le deuxième les quelques travaux sur des problèmes provenant de l'optimisation combinatoire.

CHAPITRE I

RESEAUX DE NEURONES

1. Introduction

Le système modèle (fig. 1) d'une mémoire associative que nous étudions est constitué d'un nombre N de processeurs que, suivant l'usage, nous appellerons 'neurones'. Chaque neurone i peut prendre deux valeurs σ_i=+1 (neurone 'actif') ou σ_i=-1 (neurone 'inactif'). Le neurone i peut communiquer son état à chacun des autres neurones j par un couplage 'synaptique' de poids J_{ji}.

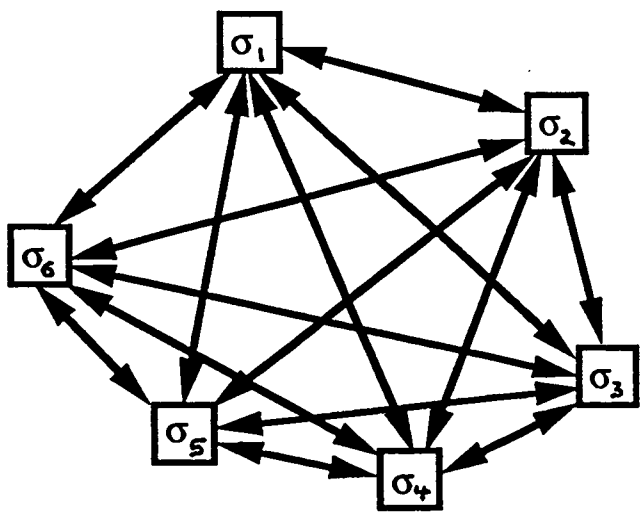


fig. 1 Système modèle d'un réseau de neurones, constitué de neurones i=1,...,N (ici N=6). Le neurone i peut communiquer son état $\sigma_i=\pm 1$ au neurone j par un couplage 'synaptique' J_{ji} ($J_{ij} \neq J_{ji}$).

Les deux régimes du système, l'*apprentissage* et le *rappel*, sont séparés l'un de l'autre (c'est-à-dire que les échelles de temps de leurs dynamiques sont supposées différentes). Dans le régime de rappel, la matrice des couplages est figée. L'état du réseau pendant la phase de rappel est donc décrit par le vecteur à N dimensions σ donnant les états de tous les neurones

$$\begin{aligned} \sigma &= (\sigma_1, \sigma_2, \ldots, \sigma_N) \\ &= (\pm 1, \pm 1, \ldots, \pm 1) \end{aligned} \qquad (1.1)$$

La configuration $\sigma = \sigma(t)$ du réseau évolue aux intervalles de temps discrets $t=0,1,\ldots$, selon la dynamique suivante:

$$\sigma_i(t+1) = \text{signe}(\Sigma_j J_{ij}\, \sigma_j(t)), \quad i=1,\ldots,N,\ t=1,2,\ldots \qquad (1.2)$$

Ceci est une dynamique (parallèle) de Monte Carlo à température zéro, qui se généralise facilement à des températures non-nulles [2].

Pendant l'apprentissage, un nombre p de 'patterns' $\xi^\mu, \mu=1,\ldots,p$ sont présentés au réseau. Les patterns sont des configurations possibles du système (correspondant à un état donné (actif ou inactif) de chaque neurone):

$$\begin{aligned} \xi^\mu &= (\xi^\mu_1, \xi^\mu_2, \ldots, \xi^\mu_N) \quad \mu=1,\ldots,p \\ &= (\pm 1, \pm 1, \ldots, \pm 1) \end{aligned} \qquad (1.3)$$

Ce sont ces patterns ξ^μ que le réseau doit être capable de reconnaître

dans la phase de rappel. Par définition le pattern ξ^μ est mémorisé par le réseau quand il est un point fixe de la dynamique éq. (1.2):

$$\text{mémoire:} \qquad \text{si } \sigma(t=0) = \xi^\mu \ , \ \sigma(t>0) = \xi^\mu \tag{1.4}$$

On dit que la mémoire est en plus *associative*, si un état initial σ proche de ξ^μ (dans un sens qui sera précisé par la suite), est ramené vers le pattern ξ^μ

$$\text{associativité:} \qquad \text{si } \sigma(t=0) \approx \xi^\mu \ , \ \sigma(t \to \ldots\infty) = \xi^\mu \tag{1.5}$$

c'est-à-dire si le pattern ξ^μ peut être récupéré à partir de données ($\sigma(t=0)$) incomplètes. Les paradigmes (1.4) et (1.5) sont essentiellement dus a Hopfield [2]. Evidemment, les définitions de la mémoire peuvent facilement être généralisées pour tenir compte d'un taux d'erreur dans le rappel.

Afin de quantifier les possibilités de ces réseaux et de comparer diverses règles d'apprentissage, un cas d'étude extrêmement important est celui des *patterns aléatoires*. Les patterns ξ^μ sont dits aléatoires, si leurs composantes sont tirées au hasard

$$\xi_i^\mu = \begin{cases} +1 \text{ avec probabilité } (1+m)/2 \\ -1 \text{ avec probabilité } (1-m)/2 \end{cases} \tag{1.6}$$

le paramètre m est le *biais* des patterns. L'introduction d'une distribution de probabilité pour les patterns rend possible l'étude des réseaux de

neurones dans la limite $N \to \infty$, la *limite thermodynamique.*

La *règle d'apprentissage* est choisie afin de mémoriser (c'est-à-dire stabiliser) les patterns ξ^μ, et afin de les entourer de *bassins d'attraction* aussi grands que possibles (c'est-à-dire de ramener le plus de configurations vers les patterns ξ^μ).

L'analyse que nous présentons dans ce chapitre s'appuie en grande partie sur la notion de la *stabilité*. Cette notion, qui sera introduite dans la section suivante (I.2), permet une discussion unifiée de la *mémorisation* et de l'*associativité* dans le réseau de Hopfield. Plus encore, la stabilité donne un critère à la fois clair et simple pour construire des algorithmes d'apprentissage avec de très bonnes propriétés associatives. Toujours dans la section I.2, un *algorithme de stabilité optimale* - le 'minover' - sera présenté. Nous discuterons les liens de ce nouvel algorithme avec l'algorithme du *perceptron* (de vingt ans son aîné).

La stabilité donne un critère, une fonction de coût, qui peut être attribuée à toute règle d'apprentissage possible. Ceci est l'origine d'une approche due à E. Gardner, qui a introduit des calculs analytiques sur *l'espace des règles d'apprentissage*. La section I.3 est une introduction à cette approche propre à la physique statistique. Les calculs *à la Gardner* permettent en particulier de déterminer analytiquement les performances de l'algorithme 'minover'. La section I.4 sera consacrée à l'étude (numérique et analytique) du *régime de rappel*. Nous tâcherons surtout d'étayer l'hypothèse initiale selon laquelle la stabilité doit être optimisée afin d'optimiser les propriétés associatives du réseau.

Dans la section I.5, enfin, nous reviendrons sur le problème de

l'apprentissage, mais cette fois-ci dans le cas où les efficacités synaptiques sont contraintes à prendre deux valeurs $J_{ij}=\pm 1$. Pour ce système, le rôle de la stabilité reste inchangé par rapport aux systèmes à couplages J_{ij} continus. Les analyses sur la phase de rappel dans la section I.4 restent donc valables. En revanche, le calcul de la stabilité optimale est profondément changé. D'une part, le problème de l'algorithme qui cherche un réseau de stabilité optimale (résolu par le 'minover' dans le cas continu) est à ce jour sans solution véritable : nous utilisons un algorithme d'énumération pour estimer les valeurs de la stabilité optimale dans la limite thermodynamique. D'autre part, le calcul à la Gardner sur l'espace des règles d'apprentissage discrètes se complique nettement par l'existence d'une phase verre de spin. Celle-ci nécessite l'inclusion d'effets dits de *brisure de la symétrie des répliques*. Néanmoins la double approche numérique et analytique fournit des résultats précis et cohérents.

2. Règles d'apprentissage

Pour les patterns aléatoires non-biaisés (m=0), Hopfield [2] a proposé un algorithme d'apprentissage selon la règle de Hebb [5]:

$$J_{ij} = \begin{cases} 1/p \sum_{\mu=1,..,p} \xi_i^\mu \xi_j^\mu & (i \neq j) \\ 0 & (\text{sinon}) \end{cases} \qquad (1.7)$$

Il se trouve qu'avec cette règle, pour le cas de patterns aléatoires, on arrive à mémoriser un nombre considérable de patterns p=0.14 N avec un très faible nombre d'erreurs.

Depuis la proposition de Hopfield, ce modèle, utilisant des patterns aléatoires, a été étudié en grand détail [6]. Cette étude est facilitée par le fait qu'avec le choix de la règle de Hebb éq. (1.7), la matrice (J_{ij}) des connections synaptiques est symétrique. En conséquence, la dynamique de Monte-Carlo éq. (1.2) peut s'écrire comme une dynamique de relaxation à la température T pour un système décrit par un Hamiltonien; le réseau de neurones devient alors un 'vrai' système physique.

Toujours à propos de ce problème d'apprentissage, un grand nombre d'algorithmes différents ont été proposés [7,8,9,10] afin d'améliorer à la fois le nombre de patterns que le modèle est capable de mémoriser et les propriétés associatives du réseau.

Finalement, la connexion profonde du problème d'apprentissage sur le réseau de Hopfield avec l'apprentissage bien connu du perceptron [11] a

été révélée. En fait, la condition de stabilité des patterns (voir éqs. (1.2) et (1.4)) s'écrit comme

$$\{\ \xi_i^\mu = \text{signe}(\Sigma_{j(\neq i)}\, J_{ij}\, \xi_j^\mu)\quad \mu=1,...,p\ \}\quad i=1,...,N \qquad (1.8)$$

(la notation $\Sigma_{j(\neq i)}$ indique une sommation uniquement sur l'indice j, excluant le terme j=i). L'équation (1.8) équivaut à

$$\{\ 0 < \kappa \leq \Sigma_{j(\neq i)}\, J_{ij}\, \xi_i^\mu\, \xi_j^\mu \quad \mu=1,...,p\ \}\quad i=1,...,N \qquad (1.9)$$

Dans la formule (1.9) les patterns ξ_i^μ sont des paramètres, tandis que les poids synaptiques J_{ij} sont les variables dynamiques de l'apprentissage. Pour chaque neurone i, la condition de stabilité de tous les patterns ne fait intervenir que les connexions J_{ij}, j=1,...,N, c'est-à-dire le vecteur de la ième ligne de la matrice (J_{ij}). Les conditions de stabilité éq.(1.9) sur la matrice (J_{ij}) se découplent donc en N systèmes autonomes (à l'intérieur des parenthèses { }) si les différentes lignes de la matrice (J_{ij}) sont indépendantes (voir l'article [I]).

Ainsi, le problème complet éq. (1.9) de l'apprentissage se décompose en N systèmes avec N-1 unités d'entrée et une seule unité de sortie, c'est-à-dire N systèmes du type perceptron :

$$0 < \Sigma_{j(\neq i)}\, J_{ij}\, \xi_i^\mu\, \xi_j^\mu \quad \mu=1,...,p \qquad (1.10)$$

L'éq. (1.10) peut s'écrire comme équation vectorielle entre le vecteur de ligne $\mathbf{J} = (J_{i1}, J_{i2}, ..., J_{iN})$ et les patterns 'auxiliaires' $\eta^\mu = \xi_i^\mu\, \xi^\mu$

$$0 < \mathbf{J} \cdot \eta^{\mu} \ \mu=1,...,p \quad \Leftrightarrow \quad 0 < \min_{\mu} \mathbf{J} \cdot \eta^{\mu} \qquad (1.11)$$

Il est important de réaliser que l'éq. (1.11) offre une formulation géométrique de la mémoire équivalente à la définition dynamique initiale (éq .(1.4)) : un vecteur **J** mémorise (au neurone i) tous les patterns ξ^{μ} si et seulement s'il a un recouvrement positif avec les patterns 'auxiliaires' η^{μ} (voir fig. 2).

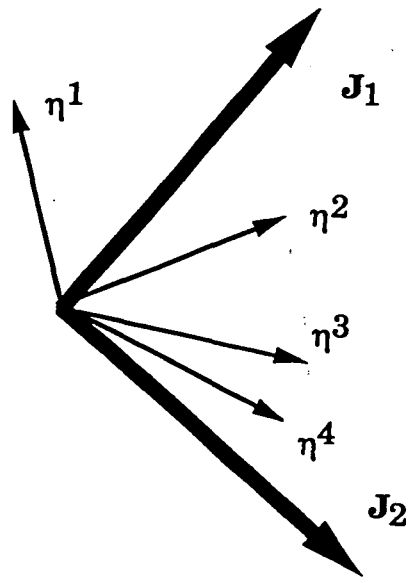


*fig. 2 L'image géométrique de l'apprentissage (voir éq. (1.11)). Un pattern ξ^{μ} est appris par le vecteur **J** si le recouvrement de **J** avec le pattern 'auxiliaire' η^{μ} est positif. En termes géométriques l'apprentissage consiste à trouver un vecteur **J** ayant un recouvrement positif avec tous les vecteurs η^{μ}. J_1 est un tel vecteur, J_2 ne l'est pas.*

Si elle existe, une telle solution **J** peut être trouvée par le célèbre *algorithme du perceptron* [11] : commençant par un vecteur $\mathbf{J} = \mathbf{J}(t=0) = (0,0,...,0)$ on cherche, à chaque instant du temps d'apprentissage t=0,1,..., un pattern η^{μ} qui n'est pas encore appris $\mathbf{J}(t) \cdot \eta^{\mu} < 0$ et on l'ajoute à la solution du pas de temps précédent: $\mathbf{J}(t+1)=\mathbf{J}(t)+\eta^{\mu}$. L'apprentissage est terminé (t=M) s'il n'existe plus de patterns avec recouvrement négatif, le

problème (1.11) étant alors résolu. Le *théorème de convergence du perceptron* [11] assure que le temps d'apprentissage avec l'algorithme du perceptron est fini si et seulement si le problème (1.11) a une solution **J**. (Il existe des méthodes basées sur la programmation linéaire ([12], I) qui permettent de décider en un temps fini (qui croît comme une fonction polynomiale de la taille N du système) si le problème de l'apprentissage n'a aucune solution).

En conséquence, l'algorithme du perceptron serait une bonne extension de la règle de Hebb dans le modèle de Hopfield [8], [13] : pour un problème donné, il est possible de déterminer une matrice (J_{ij}) (par N applications de l'algorithme du perceptron) qui, stabilise tous les patterns, à condition qu'il existe au moins une telle solution. Cet algorithme (comme ses variantes) converge en général vers des matrices non-symétriques ($J_{ij} \neq J_{ji}$). Un signe de sa robustesse est qu'il peut être modifié pour préserver la symétrie de la matrice (J_{ij}) [8] (à condition qu'il existe une matrice symétrique stabilisant tous les patterns).

Si les méthodes *classiques* permettent déjà de donner une réponse algorithmique au problème de la mémorisation, la transcription d'un autre résultat connu depuis les années 60 [14] permet d'établir un résultat d'une grande importance théorique : le nombre maximal de patterns aléatoires et sans biais qui peuvent être mémorisés sans erreurs dans le modèle de Hopfield se comporte dans la limite $N \to \infty$ comme $p \sim 2N$.

Ce résultat est important en ce qu'il fournit une valeur de la capacité de mémorisation de patterns aléatoires dans le modèle de Hopfield, indépendamment de toute règle d'apprentissage explicite. Si certaines

règles ont des performances restant très en deçà de cette limite, les règles élaborées à partir de l'algorithme du perceptron permettent d'atteindre la limite théorique de Cover (car elles convergent à condition qu'il existe une solution). Notons enfin, que la limite de mémorisation de Cover s'applique à la mémorisation des patterns *stricto sensu*, la limite de mémorisation avec un faible taux d'erreurs n'étant pas connue.

Jusqu'ici nous nous sommes préoccupés du problème de la mémorisation (éq. (1.4)) des patterns ξ^μ, laissant de côté la condition de l'associativité (éq. (1.5)), qui est d'une égale importance. A première vue, ceci semble être un problème beaucoup plus compliqué. La mémorisation (éq. (1.4)) des p patterns ξ^μ implique (sur chaque neurone) p conditions sur la dynamique à un pas de temps à partir de ces patterns (ceci est le contenu de l'équation (1.8)). L'associativité éq. (1.5), au contraire, correspond à un grand nombre de conditions sur la dynamique à plusieurs pas de temps à partir de toute configuration σ proche d'un des patterns.

Nous avons proposé dans ce contexte [15], [I] de considérer des vecteurs $\mathbf{J}$ de norme fixée (comme $|J_{ij}| \leq 1$, j=1,...,N; ou $\sum_j J_{ij}^2 \leq N$) vérifiant les conditions :

$$0 < \kappa\sqrt{N} \leq \mathbf{J}\ \eta^\mu\ \mu=1,...,p \quad \Leftrightarrow \quad 0 < \kappa\sqrt{N} \leq \min_\mu \mathbf{J}\cdot\eta^\mu \qquad (1.12)$$

avec une *stabilité* κ aussi grande que possible (pour des vecteurs $\mathbf{J}$ de norme bornée): pour la norme du maximum ($|J_{ij}| \leq 1$, j=1,...,N) on voit facilement [I] que chaque configuration initiale σ qui diffère du pattern ξ^μ en moins de $\kappa\sqrt{N}/2$ positions est ramenée vers ce pattern en un pas de

temps. Regardons sans restriction de généralité le cas où seuls les s premiers éléments de σ diffèrent de ξ^μ ($\sigma_i = -\xi_i^\mu$, i=1,...,m ; $\sigma_i = \xi_i^\mu$, i=m+1,...,N). La condition de convergence en un pas de temps est (voir éqs. (1.8), (1.9))

$$0 < \xi_i^\mu \Sigma_{j(\neq i)} J_{ij} \xi_i^\mu \sigma_j = \xi_i^\mu \{ -\Sigma_{j=1} J_{ij} \xi_j^\mu + \Sigma_{j=m+1} J_{ij} \xi_j^\mu \}$$
$$= \{ \underbrace{\xi_i^\mu \Sigma_{j=1} J_{ij} \xi_j^\mu}_{\geq \kappa\sqrt{N}} \underbrace{-2\ \xi_i^\mu \Sigma_{j=1} J_{ij} \xi_j^\mu}_{\leq 2\cdot m} \} \qquad (1.13)$$

(à cause de la condition $|J_{ij}| \leq 1$). Pour assurer la convergence on a donc la condition $m < \kappa\sqrt{N}/2$. A partir d'un tel argument on est conduit à rechercher des algorithmes d'apprentissage optimisant la stabilité, pour l'une ou l'autre définition de la norme du vecteur **J**.

L'algorithme 'minover' que nous proposons dans [I] est capable de déterminer la solution $\mathbf{J}_{opt}$ avec stabilité optimale κ_{opt} parmi tous les vecteur **J** de norme euclidienne fixée $\Sigma_j J_{ij}^2 = N$, c'est-à-dire de trouver un vecteur maximisant une stabilité renormalisée

$$\kappa' = \{\min_\mu \mathbf{J}\cdot\eta^\mu\} / |\mathbf{J}| \qquad (1.14)$$

où $|\mathbf{J}|$ est la norme *euclidienne* de **J**. Cet algorithme est d'une certaine importance, car il se trouve que la stabilité $\Delta^\mu' = \mathbf{J}\cdot\eta^\mu / |\mathbf{J}|$ correspondant à la norme euclidienne est le paramètre pertinent pour la dynamique du réseau proche d'un pattern ξ^μ (voir la section I.4)

L'algorithme 'minover' converge vers la solution optimale (c'est-à-dire un vecteur de stabilité minimale optimale) pour tout ensemble de

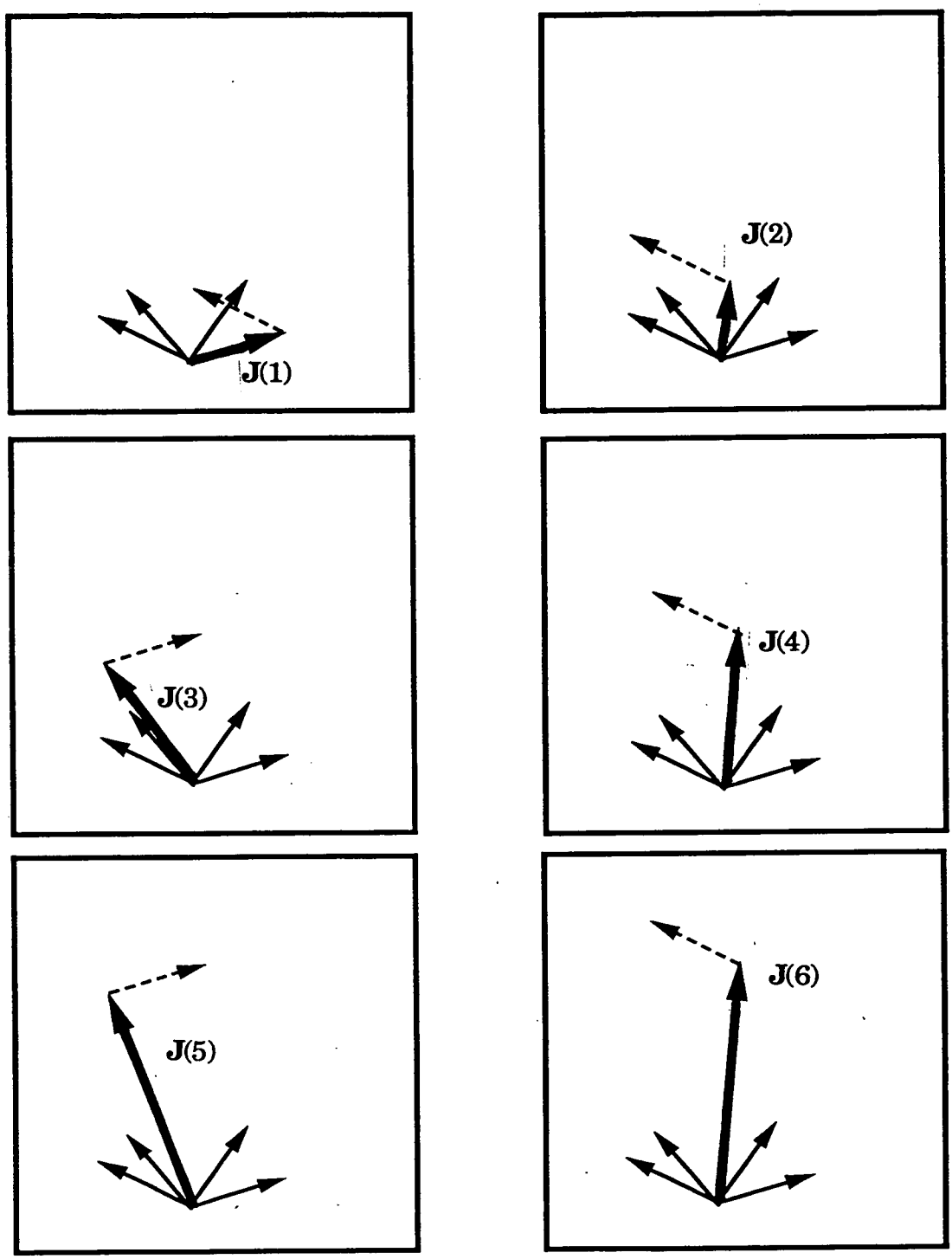


fig. 3 L'algorithme 'minover' en action. A chaque itération t, on ajoute au vecteur J(t) le pattern auxiliaire η^{μ} *ayant un recouvrement minimal avec J(t).*

patterns pour lequel une solution avec stabilité positive existe. Il est relié à l'algorithme du perceptron en ce qu'il est de nature itératif (voir fig. 3). A chaque instant du temps d'apprentissage le vecteur $\mathbf{J}(t)$ est mis à jour suivant $\mathbf{J}(t+1)=\mathbf{J}(t)+c\cdot\eta^{\mu(t)}$, où $\eta^{\mu(t)}$ est le pattern ayant un *recouvrement minimal* avec la solution $\mathbf{J}(t)$ (dans le perceptron $\eta^{\mu(t)}$ était un pattern quelconque ayant un recouvrement inférieur à zéro). Le critère d'arrêt est aussi différent de celui du perceptron (voir I).

Comme nous le verrons plus loin, l'algorithme 'minover' donne une réponse satisfaisante au problème d'une règle d'apprentissage créant des bassins d'attraction les plus grands possibles.

3. Apprentissage : Calculs analytiques d'après Gardner

L'algorithme 'minover' est une généralisation directe (algorithmique) du perceptron. Il est également possible [15,16,17], sur le plan théorique, de généraliser d'une manière substantielle les résultats 'classiques' de Cover [14] sur le perceptron, utilisant des méthodes propres à la physique statistique. La voie menant vers ces calculs aujourd'hui assez répandus a été ouverte par E. Gardner [15] . Nous traitons dans cette section-ci les idées principales de ce calcul et de ses généralisations, qui permettent de déterminer pour des patterns aléatoires (même biaisés) la valeur de la stabilité optimale $\kappa_{opt}(\alpha)$ [15], la distribution des champs $\Delta^\mu = \mathbf{J}\ \eta^\mu / |\mathbf{J}|$ [III,18], ainsi que le temps de convergence de l'algorithme [19]. Dans la prochaine section (I.4) nous traiterons du rôle exact de la stabilité dans la dynamique.

La solution $\mathbf{J}_{opt}$, qui se calcule avec l'algorithme 'minover' a, en effet, deux propriétés :

a) sa stabilité est la stabilité optimale κ_{opt}

b) c'est la seule solution de stabilité κ_{opt}

Nous nous sommes appuyés sur la première de ces propriétés pour construire le vecteur $\mathbf{J}_{opt}$, en proposant des changements successifs qui faisaient augmenter la stabilité.

Dans un travail tout à fait remarquable, E. Gardner montre que le deuxième critère (b)) peut être utilisé pour calculer analytiquement la stabilité pour le cas des patterns aléatoires. Plus précisément, le travail de Gardner consiste à considérer l'hypersphère en N dimensions

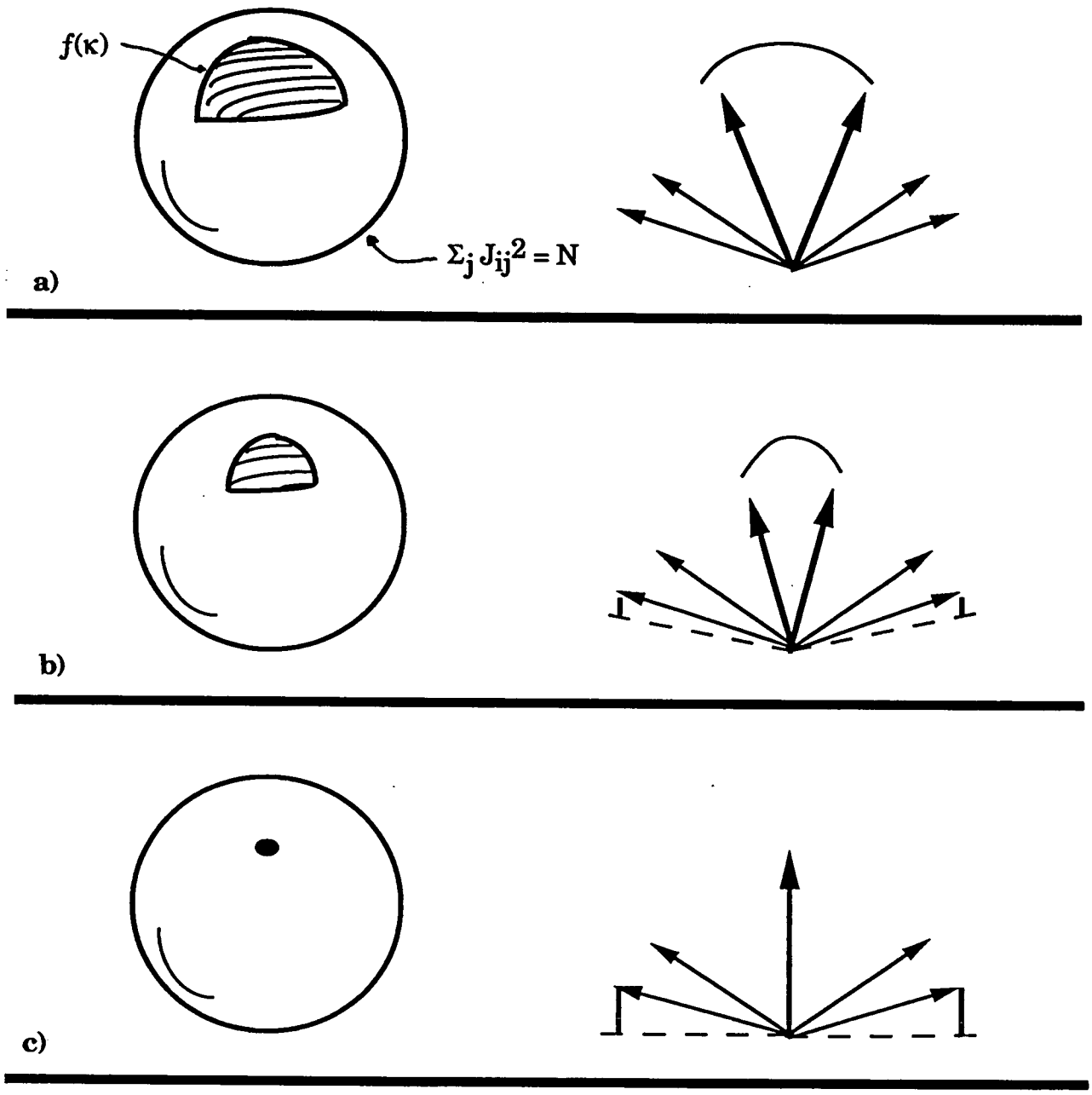


fig. 4 L'espace $f(\kappa)$ sur l'hypersphère $\Sigma_j J_{ij}^2 = N$ pour $\kappa = 0$ (a). $f(\kappa)$ est un domaine connexe ($\kappa > 0$), étendu sur la surface de la sphère ; le recouvrement de deux solutions J_α et J_β est petit. L'espace $f(\kappa)$ se rétrécit comme κ augmente (b). $f(\kappa_{opt})$ n'est qu'un point, le recouvrement entre deux solutions est 1 (solution unique) (c).

$\Sigma_j J_{ij}^2 = N$, c'est-à-dire l'espace de tous les vecteurs $\mathbf{J}$ avec $|\mathbf{J}|^2 = N$. Dans cet espace, le sous-espace $f(\kappa)$ de vecteurs $\mathbf{J}$ qui stabilisent tous les patterns avec une stabilité $\geq \kappa$ (voir l'éq. (1.12)) est connexe. Pour une valeur de κ en dessous de la stabilité optimale, le rapport $V(\kappa)$

$$V(\kappa) = \frac{\int \Pi_j \, dJ_{ij} \, \Pi_\mu \Theta(\Sigma_j J_{ij} \eta_j^\mu - \kappa \sqrt{N}) \, \delta \, (\Sigma_j J_{ij}^2 - N)}{\int \Pi_j \, dJ_{ij} \, \delta \, (\Sigma_j J_{ij}^2 - N)} \qquad (1.15)$$

entre le volume du sous-espace $f(\kappa)$ et le volume de l'hypersphère est fini (v. fig. 4a). En conséquence, le recouvrement typique de deux solutions $\mathbf{J}_\alpha$ et $\mathbf{J}_\beta$ vérifiant tous les deux cette condition (1.12)

$$q_{\alpha\beta} = \mathbf{J}_\alpha \mathbf{J}_\beta / (|\mathbf{J}_\alpha| \cdot |\mathbf{J}_\beta|) \qquad (1.16)$$

est plus petit que 1 (c'est-à-dire que $\mathbf{J}_\alpha$ et $\mathbf{J}_\beta$ ne sont pas identiques, voir fig. 4a). Le recouvrement typique entre deux solutions se rapproche de 1 quand la stabilité κ se rapproche de la stabilité optimale κ_{opt} (v. fig. 4b).

Pour un ensemble de patterns, comme des patterns aléatoires (éq. (1.6)), il est intéressant de connaître les *valeurs typiques* du volume de $f(\kappa)$. Un résultat fondamental de la physique statistique [20, 21] prescrit dans ce cas de moyenner sur des quantités extensives du système. Ici il faut donc calculer la *moyenne du logarithme* de $V(\kappa)$ sur un ensemble de patterns. ($\log V(\kappa)$ a la dimension d'une entropie; en physique statistique c'est l'entropie qui est une quantité thermodynamique extensive. Nous reviendrons sur cette question à la fin de cette section.)

Pour le calcul de la moyenne $< >$ du logarithme de $V(\kappa)$ sur les patterns aléatoires η^μ, Gardner a utilisé la méthode des répliques [21], [22] :

$$< \log V(k) > = \lim_{n\to 0} [\{ < V^n > -1\} / n] \qquad (1.17)$$

avec

$$V^n = \frac{\int \Pi_{j,a} dJ_{ij}^a \Pi_\mu \Theta(\Sigma_j J_{ij}^a \eta_j^\mu - \kappa\sqrt{N})\, \delta(\Sigma_j J_{ij}^{a2} - N)}{\int \Pi_{j,a} dJ_{ij}^a \delta(\Sigma_j J_{ij}^{a2} - N)} \qquad (1.18)$$

où l'indice a varie de 1 à n ; dans la limite $N\to\infty$ on trouve

$$< V^n > = e^{Nn(\min_q \{ \alpha \int \mathcal{D}t \ln H(\frac{\sqrt{qt+\kappa}}{\sqrt{1-q}}) + 1/2\ln(1-q) + 1/2 \frac{q}{1-q}\}} \qquad (1.19)$$

avec $\mathcal{D}t = \frac{\exp(-t^2/2)\,dt}{\sqrt{2\pi}}$ $\qquad H(x) = \int_x^\infty \mathcal{D}z$

Dans l'éq. (1.19) l'hypothèse de la *symétrie des répliques* est prise, c'est-à-dire que tous les paramètres d'ordre $q_{\alpha\beta}$ sont supposés égaux $q_{\alpha\beta}=q$ [15] (voir plus loin). Le résultat du calcul donne dans la limite $q\to 1$ (solution unique) pour la capacité optimale $\alpha_c=p/N$

$$\alpha_c = \frac{1}{\int_{-\kappa}^{\infty} \mathcal{D}t\,(t+\kappa)^2} \qquad (1.20)$$

Le calcul de Gardner qui vient d'être présenté en quelques mots permet donc de déterminer, pour des patterns aléatoires, le volume typique des sous-ensembles $f(\kappa)$ comme fonction de κ, ou κ_{opt}, dans la limite $N \to \infty$. Le passage de l'équation (1.17) à l'équation (1.20) est loin d'être trivial, et s'appuie en grande partie sur les progrès des dix dernières années en physique statistique des systèmes désordonnés : la méthode des répliques [22] (ici le désordre est réalisé par les patterns ξ^μ, qui sont différents d'un système à l'autre, mais qui sont engendrés à partir de la même distribution de probabilité). Ce sont des développements assez récents qui ont permis de contrôler les hypothèses contenues dans la méthode des répliques, surtout l'ansatz de la *symétrie des répliques*. Dans le calcul de Gardner cette hypothèse est cohérente [15] et le résultat sans doute exact (dans la limite thermodynamique).

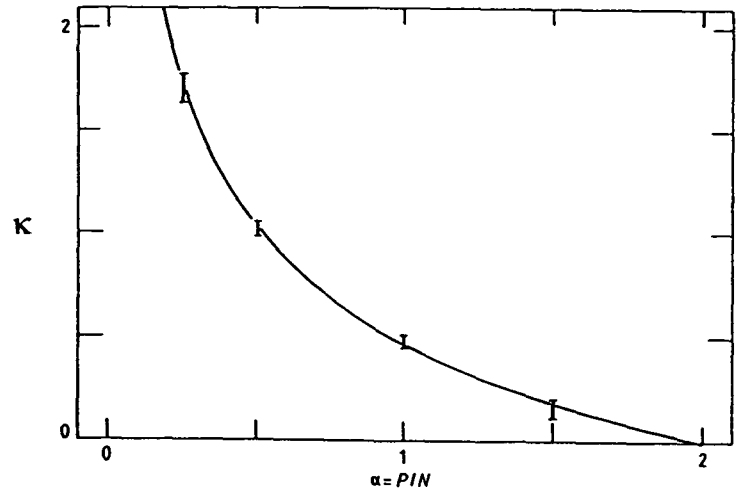


fig. 5 *Stabilité optimale pour des patterns aléatoires non-biaisés. Ici on compare la solution analytique de Gardner (éq. (1.20)) avec des simulations numériques utilisant l'algorithme 'minover' (voir [I]).*

Dans la fig. 5 nous montrons le résultat théorique (éq. (1.20)) et celui des simulations numériques. Les deux approches sont en bon accord. La stabilité minimale devient zéro à la valeur de la capacité $\alpha=p/N=2$ prédite par Cover [14]. La valeur de κ sera reliée à la taille (minimale) des bassins d'attraction dans la section I.4.

Si l'éq. (1.20) détermine la valeur de la stabilité optimale κ_{opt}, c'est-à-dire du plus petit des champs, $\kappa=\min_\mu(\Delta^\mu = \mathbf{J}\, \eta^\mu/|\mathbf{J}|)$, la distribution de probabilité des champs Δ^μ peut, elle aussi, être déterminé par un calcul de répliques [III, éq. (4.10)], [18] . Cette distribution est importante car elle gouverne les propriétés d'associativité (l'attractivité près du pattern ξ^μ étant déterminée par la stabilité Δ^μ). Un autre résultat concerne la durée de l'apprentissage, c'est-à-dire le nombre d'itération de l'algorithme à une certaine précision du résultat. La durée de l'apprentissage utilisant l'algorithme 'minover' a été calculée par Opper [19] pour un échantillon typique, améliorant la borne valable pour chaque échantillon [I, éq.(A1.5)] par un facteur numérique.

Même si la méthode des répliques permet de calculer un certain nombre de propriétés de la matrice de couplages ayant la meilleure stabilité, les quantités se rapportant à plusieurs lignes de la matrice (J_{ij}) *en même temps* restent hors de portée. Une des plus élémentaires de ces quantités est la symétrie de la matrice (J_{ij}) que nous définissons comme

$$\eta = (\Sigma_{i,j}\, J_{ij}J_{ji})/\,(\Sigma_{i,j}\, J_{ij}J_{ij}) \qquad (1.21)$$

Cette symétrie de la matrice optimale doit donc en général être déterminée numériquement, a partir de simulations utilisant l'algorithme 'minover'. Pour des valeurs faibles de la capacité α, on démontre facilement (III, éq. 18) que l'algorithme 'minover' donne la même solution qu'une autre règle d'apprentissage, où la matrice (Jij) est symétrique : la pseudo-inverse [9,10]. Dans la limite $\alpha \to 0$, la matrice (J_{ij}) de l'algorithme optimal est donc, elle aussi, symétrique. Pour des valeurs de α s'approchant de la valeur critique on trouve numériquement que la symétrie η reste élevée (voir la fig. 6 pour le résultat numérique).

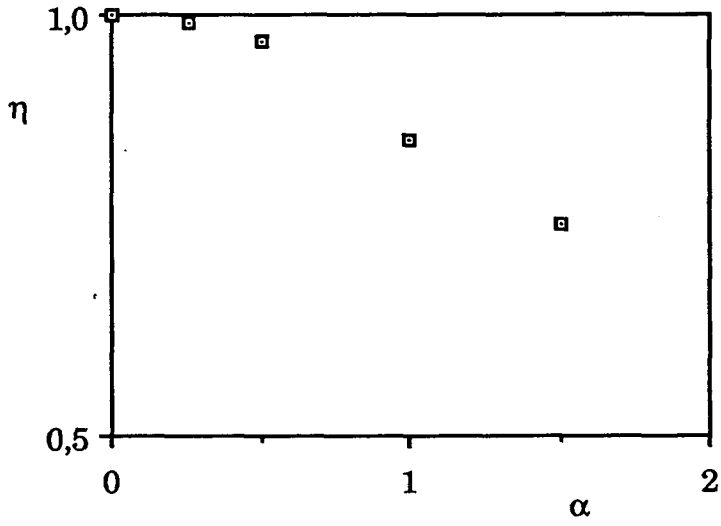


fig. 6 La valeur de la symétrie η de la matrice (J_{ij}) calculée avec l'algorithme 'minover' en fonction de la capacité α pour des patterns aléatoires.

Notons enfin que la valeur de la capacité optimale, elle aussi, est une quantité se rapportant à toutes les lignes de la matrice (J_{ij}) en même temps et , à cet égard, la seule telle quantité facilement accessible. En toute rigueur, il aurait fallu définir le volume V sur toutes les lignes :

$$V(\kappa) = \frac{\int \Pi_{i,j}\, dJ_{ij}\, \Pi_\mu \Theta(\Sigma_j \xi_i^\mu J_{ij} \xi_j^\mu - \kappa\sqrt{N})\, \Pi_i\, \delta(\Sigma_j J_{ij}^2 - N)}{\int \Pi_{i,j}\, dJ_{ij}\, \Pi_i\, \delta(\Sigma_j J_{ij}^2 - N)} \qquad (1.22)$$

donc on aurait

$$\log V(\kappa) = \Sigma_i \log(V_i) \tag{1.23}$$

avec

$$V_i(\kappa) = \frac{\int \Pi_j dJ_{ij} \Pi_\mu \Theta(\Sigma_j \xi_i^\mu J_{ij} \xi_j^\mu - \kappa\sqrt{N})\, \delta\,(\Sigma_j J_{ij}^2 - N)}{\int \Pi_j dJ_{ij}\, \delta\,(\Sigma_j J_{ij}^2 - N)} \tag{1.24}$$

La moyenne < > sur les patterns ξ^μ donne alors

$$<\log V(k)> = \Sigma_i < \log(V_i)> = N \cdot <\log(V_i)> \tag{1.25}$$

puisque le volume $<\log V_i>$ ne dépend plus de l'indice i. De ce fait (sous condition que le volume typique puisse s'obtenir en moyennant le logarithme du volume) la capacité α pour une ligne est égale à la capacité pour toute la matrice (J_{ij}), même si les vecteurs $\eta^\mu = \xi_i^\mu \xi^\mu$ ne sont pas indépendants d'une ligne à l'autre.

4. Dynamique du réseau dans le régime de rappel

La dynamique dans la phase de rappel (l'évolution au cours du temps d'un état σ) est au coeur des réseaux de neurones, pour lesquels il s'agit de modéliser un processus de calcul comme dynamique d'un système physique [2]. D'habitude, les systèmes considérés en physique statistique sont dotés d'une fonction d'énergie (car les couplages sont symétriques). Pour ces systèmes, l'étude des propriétés d'équilibre est primordiale. Leur dynamique est fortement contrainte par le fait qu'elle est une *dynamique de relaxation.* Par contre, les systèmes que nous considérons sont en général privés de propriétés d'équilibre (la matrice (Jij) n'étant pas symétrique), et ne sont définis que par leur*dynamique*.

On pourrait donc s'attendre que la dynamique des réseaux de neurones (asymétriques) soit très compliquée. Nous montrons, toutefois, que ceci n'est pas le cas, au moins près d'un pattern.

Naturellement, la dynamique du rappel est étroitement liée à celle de l'apprentissage. Nous avions indiqué que le lien le plus important était donné par la *stabilité*, qui à la fois 'ancre' les patterns ξ^μ dans le réseau et 'creuse' de grands bassins d'attraction. C'est cette hypothèse qui était à la base de la partie sur l'apprentissage, et qui sera au centre de notre intérêt dans la présente section.

Une démonstration très claire du rôle de la stabilité dans le réseau de Hopfield a été obtenue dans des simulations numériques de Forrest [23]. Dans ces simulations on regarde l'évolution au cours du temps d'états aléatoires σ ayant un certain recouvrement initial q_0 avec un pattern ξ^γ

$$q_0 = 1/N \sum \sigma_i(t=0)\xi_i^\gamma \qquad (1.26)$$

et on s'intéresse principalement à leur probabilité de converger vers ce pattern. Forrest a comparé, pour un réseau et un ensemble de patterns donné ξ^μ, $\mu=1,...p$, la règle de Hebb avec un algorithme du type perceptron [8] pour deux valeurs différentes de la stabilité.

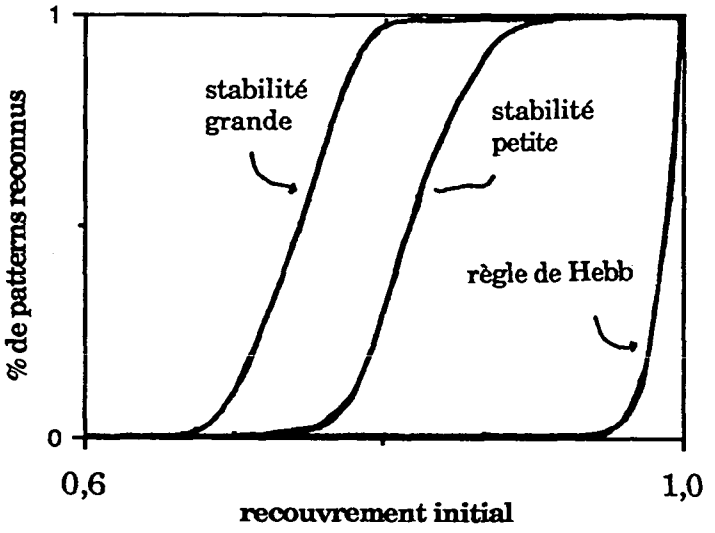


fig. 7 Bassins d'attraction pour trois règles d'apprentissage (schématiquement, après Forrest [23]).

La fig. 7 est basée sur les résultats numériques de Forrest, que nous retraçons schématiquement. Il se trouve que les bassins d'attraction pour la règle de Hebb sont très petits. Pour les règles à stabilité non-nulle, la taille du bassin d'attraction est d'autant plus grande que la stabilité est grande. L'analyse de taille finie de Forrest indique que dans la limite thermodynamique $N\to\infty$ la probabilité $P(q_0)$ de converger vers le pattern ξ^γ devient une fonction en escalier : pour tout recouvrement q_0 inférieur à une valeur critique q_c la probabilité $P(q_0)$ est zéro, tandis que pour q_0 dépassant q_c, $P(q_0)$ est égal à 1 : le pattern ξ^γ est reconnu avec probabilité 1. Il est donc possible de parler simplement du rayon des bassins

d'attraction, puisque ceux-ci sont essentiellement sphériques. La valeur de q_c ne tend pas vers 1 dans la limite $N \to \infty$. Ce dernier résultat est très surprenant : il contraste avec l'analyse de la section I.2 (voir éq. 1.13) où la convergence *après un pas de temps* n'était assurée que pour des configurations ayant un recouvrement initial de l'ordre de $1 - O(1/\sqrt{N})$. Bien entendu, la convergence considéré ici est une convergence qui n'est pas sûre, même si elle a lieu avec une probabilité 1.

Le modèle de Hopfield apparaît donc comme un simple discriminateur de distances (recouvrements). Ceci est une force - le système reconnaît avec probabilité 1 un état σ suffisamment proche du pattern ξ^μ - et en même temps une faiblesse : la dynamique du modèle semble assez rigide, elle permet peu de structure.

Nous étudieons ensuite l'évolution au cours du temps de configurations initiales σ tirées au sort à partir d'un pattern ξ^μ de la manière suivante :

$$\sigma_i = \begin{cases} +\xi_i^\mu & \text{avec probabilité } (1+q)/2 \\ -\xi_i^\mu & \text{avec probabilité } (1-q)/2 \end{cases} \qquad (1.27)$$

(Dans la limite $N \to \infty$ le recouvrement q_0 (éq.(1.26)) est alors égal au biais q vers le pattern (éq. (1.27)) - à un terme de l'ordre de $1/\sqrt{N}$ près. Pour N fini, comme dans toute simulation numérique, il est préférable de travailler - comme l'a fait Forrest - dans le cadre de l'équation (1.26), qui

apporte moins de bruit statistique.)

Il se trouve que la dynamique à un pas de temps est soluble exactement dans le modèle (1.27) [III]. Prenons un état initial σ selon l'éq. (1.27). Pour une convergence vers le pattern nous cherchons à satisfaire à la condition

$$1= \text{signe } \Sigma_{j(\neq i)} J_{ij} \xi_i^{\mu} \sigma_j(t=0) \qquad i=1,...,N \tag{1.28}$$

pour ramener la configuration σ vers le pattern (voir l'éq. (1.8)). Ceci n'est en général pas possible (en un pas de temps), puisque l'information que porte le vecteur σ sur le pattern est trop faible. Nous pouvons cependant [III] déterminer la distribution de probabilité par rapport aux conditions initiales du champ

$$h^*= \Sigma_{j(\neq i)} J_{ij} \xi_i^{\mu} \sigma_j(t=0) \tag{1.29}$$

sans avoir à moyenner sur les matrices (J_{ij}) ou les patterns ξ^{μ}. De cette distribution nous allons ensuite calculer la probabilité que l'éq. (1.28) soit vérifiée, et en déduire le recouvrement à l'instant t=1.

Le champ h* est une somme de variables indépendantes. A condition que tous les éléments de la matrice (J_{ij}) soient du même ordre, le théorème de la limite centrale montre que h* est une variable gaussienne de moyenne

$$\begin{aligned} <h^*> &= < \Sigma_{j(\neq i)} J_{ij} \xi_i^{\mu} \sigma_j> = \Sigma_{j(\neq i)} \xi_i^{\mu} J_{ij} \xi_j^{\mu} <\sigma_j> \\ &= \Sigma_{j(\neq i)} \xi_i^{\mu} J_{ij} \ \xi_j^{\mu} q \end{aligned} \tag{1.30}$$

et d'écart quadratique moyen

$$<h^{*2}>-<h^{*}>^2 = \Sigma_{j,k}\, J_{ij}\, J_{ik}\{ <\sigma_j\, \sigma_k> - <\sigma_j><\sigma_k>\}$$
$$= \Sigma_j\, J_{ij}{}^2\, \{ 1-q^2 \} \qquad (1.31)$$

voir fig. 8 (les moyennes < > sont prises par rapport à la distribution éq.(20)). Puisque l'échelle des J_{ij}, donc celle des champs h*, est sans importance pour la dynamique éq. (1.2), le paramètre pertinent pour la dynamique est le *rapport* de la moyenne de la distribution de h* à sa largeur

$$\Delta^{\mu}_i = \Sigma_{j(\neq i)}\, \xi_i^{\mu}\, J_{ij}\, \xi_j^{\mu}\, q \,/\sqrt{\Sigma_j\, J_{ij}{}^2\, \{ 1-q^2 \}} \qquad (1.32)$$

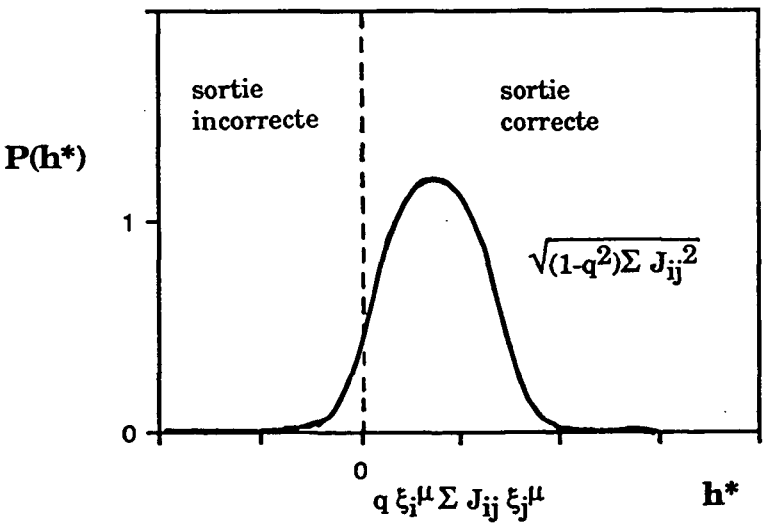


fig. 8 Distribution de probabilité du champ $h^* = \xi_i^{\mu} \Sigma_j J_{ij} \sigma_j(t=0)$. *Au premier pas de temps h* est une gaussienne dont le rapport moyenne/largeur dépend de la stabilité et de la norme euclidienne du vecteur* ***J****.*

Au pas de temps suivant l'état du neurone i sera

$$\sigma_i(t=1) = \begin{cases} + \xi_i^\mu \text{ avec probabilité } (1+q')/2 \\ - \xi_i^\mu \text{ avec probabilité } (1-q')/2 \end{cases} \qquad (1.33)$$

où q' est donné par

$$q' = \sqrt{\frac{2}{\pi}} \int_0^{\Delta\mu_i q/\sqrt{(1-q^2)}} dz\, e^{-z^2/2} \qquad (1.34)$$

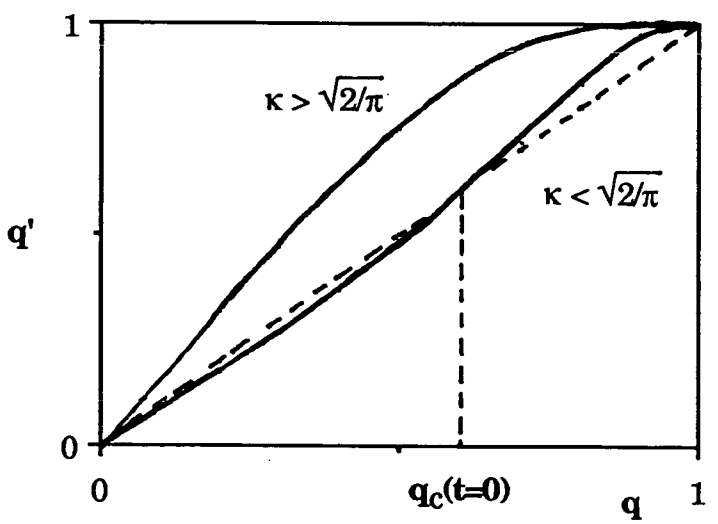


fig. 9 Recouvrement au pas de temps t=1 comme fonction du recouvrement q=q(t=0) pour deux valeurs de la stabilité κ. Une configuration s(t=0) ayant un recouvrement supérieur à $q_c(t=0)$ se rapproche du pattern au premier pas de temps avec probabilité 1 ($q_c(t=0)=0$ pour $\kappa > \sqrt{2/\pi}$).

L'éq. (1.32) est d'une importance fondamentale puisque nous y retrouvons la stabilité Δ^μ_i, avec une normalisation *euclidienne* du vecteur **J**. Le recouvrement q' à t=1 est une fonction croissante de κ et de q (voir fig. 9). Comme nous l'avons vu c'est avec l'algorithme 'minover' que nous pouvons optimiser la plus petite de ces stabilités Δ_i^μ. Cependant,

l'approche qui vient d'être présentée n'est valable qu'au premier pas de temps, et elle ne peut pas être réutilisée aux pas de temps suivants. La raison en est que les activités des neurones $\sigma_i(t>0)$ et $\sigma_j(t>0)$ sont en général *corrélées*, et non pas indépendantes ; donc la distribution du champ h* ne sera plus gaussienne. Nous ne pouvons donc pas itérer l'équation (1.34) et l'intersection de la fonction q'(q) avec la diagonale q'=q (le point $q_c(t=0\ )$ dans la fig. 9) ne nous donnera qu'une première approximation pour la valeur du recouvrement critique q_c. Ceci est dû aux corrélations des lignes de la matrice (Jij) et, comme nous l'avons vu, nous ne savons pas déterminer analytiquement ces corrélations dans le cadre d'un calcul à la Gardner (c'est-à-dire pour l'algorithme 'minover').

Dans l'impossibilité apparente de faire des progrès décisifs sur ce point du calcul du rayon des bassins d'attraction, plusieurs démarches restent toutefois possibles : D'une part il est possible de calculer explicitement l'évolution temporelle pour des règles d'apprentissage pour lesquelles une formule explicite des J_{ij} est connue. Un tel calcul a été fait pour la règle de Hebb, où la dynamique des deux premiers pas a été déterminée par Gardner et al [24].

D'autre part on a essayé de considérer des réseaux modifiés, pour lesquels (après moyenne sur les patterns) l'indépendance des états $\sigma_i(t)$ reste valable : ceci est le réseau dit 'fortement dilué' de Derrida et al [25]. Cette indépendance est aussi assurée pour un réseau où les corrélations des lignes de la matrice ont été détruites 'à la main' [II]. Enfin on a essayé de trouver des stratégies approximatives, voire phénoménologiques [18], qui permettent de donner des formules approchées pour la valeur du recouvrement critique. Dans la publication

II nous poursuivons une telle approche : nous considérons les lignes de la matrice (J_{ij}) comme indépendantes et y introduisons les corrélations à la main par le paramètre de symétrie η qui doit être déterminé indépendamment. Ceci est une méthode plutôt simple qui nous a permis toutefois de donner une bonne formule approchée de la taille des bassins d'attractions.

5. Le modèle de couplages binaires $J_{ij} = \pm 1$

Les développements précédents ont été généralisés en plusieurs directions : d'une part, des ensembles différents d'objets à mémoriser (patterns) ont été étudiés, comme des patterns biaisés [15], hiérarchiques [26], ou des patterns à plusieurs états. D'autre part on s'est intéressé aux cas où les couplages synaptiques (J_{ij}) sont contraints. Nous avons déjà fait allusion aux réseaux dilués (où la plupart des couplages sont mis à zéro), ou à un autre cas où seulement la normalisation est différente, comme par exemple $|J_{ij}| \leq 1$, j=1,...,N. Ce dernier problème [KrM] ne donne pas vraiment lieu à des effets très nouveaux : le calcul *à la Gardner* donne une fonction $\alpha_{crit}(\kappa)$ légèrement différente de l'éq.(1.20) ($\alpha_{crit}(0)$ vaut évidemment toujours 2) ; la symétrie des répliques est vérifiée et les résultats théoriques sont en bon accord avec les simulations [I], qui utilisent un algorithme de programmation linéaire.

Une extension bien différente est celle des couplages discrets, binaires $J_{ij}=\pm 1$. Ce cas est le prototype même des réseaux où la précision des J_{ij} est fixée, et ne croît pas avec la taille du système. Pour les patterns aléatoires, le formalisme de Gardner, ou les algorithmes du type perceptron ne s'appliquent que si on s'autorise de coder chaque couplage J_{ij} avec un nombre de bits m qui augmente avec la taille N du réseau comme $m \sim O(\log N)$. Ceci est une conséquence du fait que, dans ce cas, la complexité de l'algorithme pour un échantillon typique est une fonction polynomiale de N). Le modèle $J_{ij}=\pm 1$ pose donc le problème de la mémorisation *économique*.

Le problème des couplages binaires se pose à la fois en termes

d'algorithme d'apprentissage et de calcul théorique de la capacité.

Premièrement, les algorithmes du type perceptron ne semblent pas s'appliquer dans le cas des couplages binaires ; tous ces algorithmes sont de nature itératifs, car une solution initiale est améliorée au cours de l'apprentissage. Cette amélioration successive de la valeur des J_{ij} n'est pas possible si les couplages ne peuvent prendre que deux valeurs. Le problème d'apprentissage, à ce jour, n'a pas été résolu d'une manière satisfaisante. Un algorithme itératif permettrait d'établir facilement la capacité optimale du réseau ou la fonction $\alpha_{crit}(\kappa)$ de manière numérique.

$\kappa_{opt} = - \infty$

Do 1 $\tau=1,...,2^N$ ($J_{ij}(\tau)$: énumération complète)

$\Delta^\mu = \Sigma_j J_{ij}(\tau) \eta^\mu_j \quad \mu=1,...,p$

$\kappa_{opt} = \max (\kappa_{opt}, \min_\mu \Delta^\mu)$

1 continue

$\Rightarrow \kappa_{opt}$

fig. 10 *Algorithme élémentaire (énumération complète) pour le calcul de la stabilité optimale du modèle $J_{ij}=\pm 1$. Cet algorithme est d'une complexité $\alpha N^2 2^N$ (l'évaluation des Δ^μ en ligne 3 nécessite $p \cdot N$ opérations)*

Dans l'impossibilité d'utiliser un algorithme polynomial, nous avons eu recours à la plus élémentaire des méthodes numériques : l'énumération exhaustive (pour une autre méthode voir [28]). Le principe de la simulation [IV] est indiqué en fig. 10 : il s'agit d'extraire parmi des

2^N vecteurs $\mathbf{J} = (\pm1,\pm1,...,\pm1)$ un vecteur $\mathbf{J}_{opt}$ ayant la stabilité optimale, et de moyenner sur un grand nombre d'instances de vecteurs aléatoires η^μ. Nous avons pu réduire la complexité de cet algorithme exponentiel par un facteur N en utilisant une énumération complète suivant le 'code de Gray' [27], où chaque vecteur **J**, dans l'énumération de la fig. 10, diffère de son prédécesseur sur un bit seulement, ce qui réduit d'un facteur N le nombre d'opérations nécessaires au calcul des Δ^μ. Le code de Gray est expliqué figure 11.

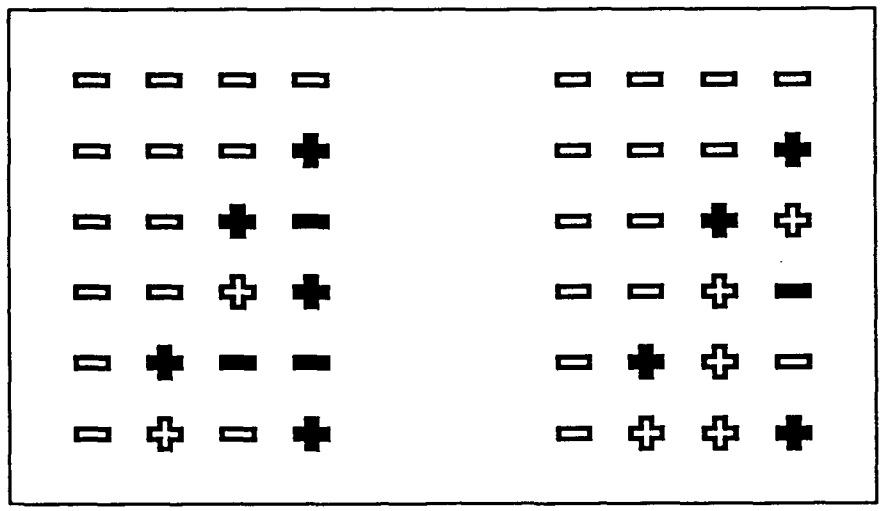


fig. 11 Début de l'énumération des vecteurs $\mathbf{J} = (\pm1,\pm1,\pm1,\pm1)$ *suivant l'énumération des nombres binaires (à gauche) et suivant le code de Gray (à droite). Dans le code de Gray chaque vecteur diffère de son prédécesseur en un bit seulement (foncé). L'utilisation du code de Gray permet de réduire la complexité de l'algorithme d'énumération (fig. 10) par un facteur N.*

Les simulations numériques (qui utilisent une distribution continue pour les patterns η^μ afin de réduire les effets de taille finie) suggèrent une valeur de la capacité critique égale à $\alpha_{crit}(\kappa=0) = 0.83$. (voir IV, fig. 2

pour l'allure de la fonction $\alpha_{crit}(\kappa)$). Ceci implique, qu'avec une précision sur les J_{ij} fortement réduite il est possible de mémoriser un nombre de patterns qui est du même ordre que pour les matrices (J_{ij}) réelles.

Il est intéressant de constater qu'avec la puissance de calcul d'un ordinateur moderne, l'utilisation d'un algorithme exponentiel ne s'interdit pas *eo ipso* ; la taille maximale considérée dans nos simulations est N=30, et les prédictions sont tout à fait précises.

Même si un algorithme d'apprentissage fait défaut, il est possible de déterminer les limites théoriques, *la capacité*, de ce modèle par un calcul *à la Gardner* . On verra surgir ici le problème intéressant du point de vue théorique, l'existence d'une phase verre de spin, qui se manifeste par la brisure de la symétrie des répliques *via* une transition du premier ordre.

Le calcul de Gardner pour les couplages $J_{ij}=\pm1$ se *formule* de la même façon que pour les J_{ij} continus. De plus, les couplages $J_{ij}=\pm1$ forment un sous-ensemble des couplages considérés dans le modèle 'sphérique' $\Sigma_j J_{ij}^2 = N$. Dans l'ensemble que nous appelions plus haut $f(\kappa)$, (l'ensemble des vecteurs $\mathbf{J}$ tel que $0 < \kappa\sqrt{N} \leq \min_\mu \mathbf{J}\cdot\eta^\mu$) il existe alors un sous-ensemble $f'(\kappa)$ de sommets de l'hypercube $\mathbf{J}=(\pm1,\pm1,...,\pm1)$. Le volume de $f'(\kappa)$ c $f(\kappa)$ (le nombre de solutions) est donné par une formule analogue à l'éq. (1.15), avec la transformation

$$\int dJ_{ij} \quad \rightarrow \quad \sum_{J_{ij}=\pm1} \tag{1.35}$$

La différence intéressante avec le problème considéré plus haut est que les sommets de l'hypercube se font pour ainsi dire assez rares : l'ensemble $f(\kappa)$ ne contient typiquement plus de sommets à des valeurs de

κ où l'extension spatiale de $f(\kappa)$ est toujours assez grande (le recouvrement typique q entre deux solutions dans $f(\kappa)$ est toujours plus petit que 1) [V] (voir fig. 12).

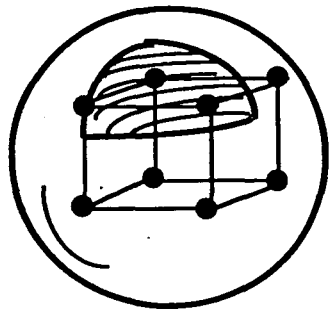


fig. 12 L'ensemble $f(\kappa)$ ne contient typiquement plus de sommets de l'hypercube $J = (\pm 1, \pm 1, \ldots, \pm 1)$ même si l'extension spatiale de $f(\kappa)$ est grande (cette image reste valable dans la limite $N \to \infty$.)

La théorie symétrique dans les répliques a été obtenue par Gardner et Derrida [17]. Les prédictions de ce calcul sont esquissées dans la fig. 13

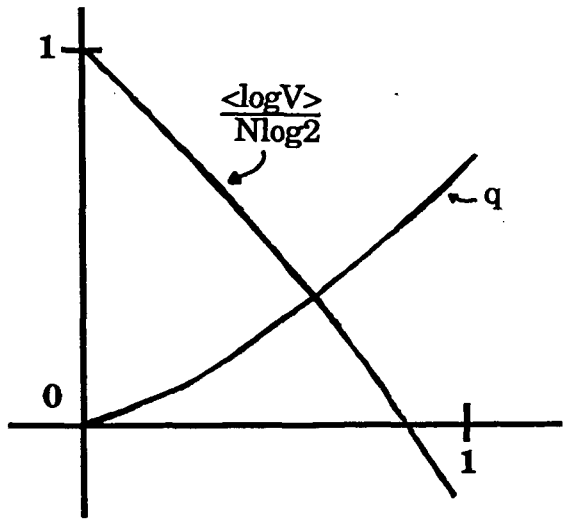


fig. 13 Prédictions du calcul symétrique dans les répliques pour le modèle $J=\pm 1$. La figure montre le logarithme du volume (divisé par Nlog2) des vecteurs J stabilisant tous les patterns η^μ comme fonction de α, et le recouvrement typique q entre les solutions.

pour la stabilité $\kappa = 0$ et nous verrons qu'il sont raisonnables pour une certaine gamme des valeurs de α. Dans la limite $\alpha \to 0$ (petit nombre de patterns) presque tous les sommets de l'hypercube sont des solutions à l'équation $\xi_i^\mu \Sigma_j J_{ij} \xi_j^\mu \geq 0$ car le nombre de telles contraintes tend vers zéro. En conséquence, le volume de l'ensemble $f'(0)$ est de l'ordre de 2^N (le nombre total de sommets $\mathbf{J} = (\pm 1, \pm 1, \ldots, \pm 1)$). Le logarithme du volume typique (divisé par le logarithme de 2^N) est donc égal à 1, comme prédit par la théorie symétrique. Egalement, le recouvrement typique q de deux solutions tend vers zéro (voir fig. 13).

En augmentant α, le volume typique de $f'(0)$ devient plus petit ; à une valeur de $\alpha = 0.833$, l'entropie, prédite par la théorie symétrique dans les répliques devient négative. Au delà de cette valeur, la théorie symétrique dans les répliques est clairement fausse, (car l'entropie dans un système discret ne peut devenir négative), même si la solution est localement stable. Le recouvrement q de deux solutions devient 1 à une valeur de $\alpha = 4/\pi = 1.27$. C'est donc cette valeur qui est la capacité critique dans le cadre de l'approximation symétrique dans les répliques.

La théorie utilisant la brisure de la symétrie des répliques *à la Parisi* à une étape [29] de ce problème (voir [22]), que nous exposons dans la publication V donne l'interprétation suivante : la théorie symétrique dans les répliques reste valable pour toute valeur de α et de κ où le volume de $f'(\kappa)$ est non-nul (pour k=0 donc jusqu'au point $\alpha = 0.833$). Au point où le nombre de sommets de l'hypercube dans $f(\kappa)$ devient plus petit que 1 (où l'entropie, c'est-à-dire le logarithme du nombre de sommets dans $f(\kappa)$, devient négative) une transition de phase du premier ordre se produit, et le système se gèle. La valeur théorique pour α_{crit} (figure 2, V), utilisant la brisure à une étape, est donc identique à la valeur de α où l'entropie de

la solution symétrique passe par zéro. Ce résultat est en très bon accord avec les simulation numériques (figure 2, IV), et une synopsis des données numériques et analytiques (figure 2 IV, figure 2 V) est donnée figure 14.

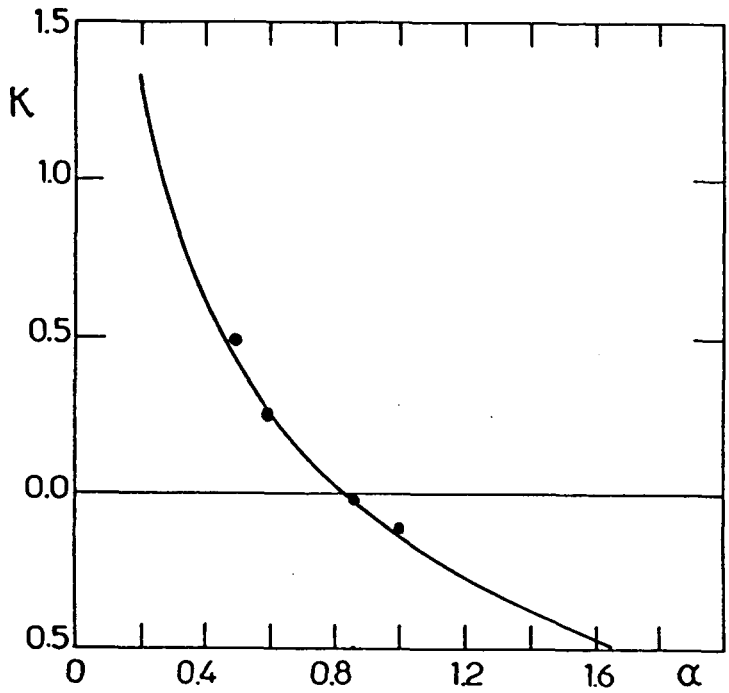


fig. 14 Stabilité optimale pour le réseau J=±1. Nous comparons les résultats analytiques avec ceux des simulations numériques (voir IV, V)

Les deux approches - théorie et simulations numériques - donnent donc une image cohérente. Il est toutefois utile de garder à l'esprit que les deux sont d'une telle complexité qu'elles suggèrent cette solution plutôt qu'elles ne la prouvent.

6. Conclusions

Pour conclure, nous rappelons que la théorie présentée dans ce chapitre a très largement été développée en analogie avec la théorie du champ moyen des verres de spin [22]. Les deux domaines sont en large partie complémentaires (voir fig. 15). Pour les verres de spin le désordre est réalisé par la distribution aléatoire des couplages J_{ij} (dans le modèle de Edwards-Anderson [21]) tandis que dans les réseaux de neurones ce sont les spins (patterns) qui portent le désordre. Il est amusant de constater qu'on retrouve à peu près les mêmes modèles. Comme chaque analogie fructueuse, celle entre les verres de spin et les réseaux de neurones a ses limites. La différence la plus fondamentale est que pour les verres de spin *l'état microscopique du système* n'a aucun intérêt, car les valeurs exactes des couplages ne sont pas accessibles expérimentalement. Pour les réseaux des neurones (comme pour le problème d'optimisation qui sera traité dans le deuxième chapitre) la situation est différente. C'est l'état microscopique du système qui nous intéresse, puisqu'il code des informations sur les patterns, les données extérieures du système.

Finalement, on a aussi pu se rendre compte - sur le plan méthodologique - de la dialectique entre le calcul théorique et les simulations numériques. Cette dialectique existait déjà entre le travail de Hopfield [2], tout à fait exploratoire et intuitif - et celui de Amit, Gutfreund et Sompolinsky [6] qui apporte au sujet rigueur et méthode. Même au niveau de cette thèse, c'est la complémentarité entre les deux procédés qui a dans une large mesure poussé nos recherches et qui nous

a permis de faire des progrès. Toutes les deux - l'analyse et la simulation - sont pleines de ressources et des méthodes inexplorées. Elles peuvent toutes les deux être originales et innovatrices, et se situent le plus souvent loin des sentiers battus.

VERRES DE SPIN		RESEAUX DE NEURONES
$Z_J = < \exp(\beta \Sigma J_{ij} S_i S_j) >_S$	•	$V\xi\mu = <\Pi_\mu \Theta(\xi_i^\mu \Sigma Jij\ \xi_j^\mu - k\sqrt{N})>_J$
$F_J = \log\{< \quad >_S\}$	•	$F\xi\mu = \log\{< \quad >_J\}$
$F = < F_J >_J = < \log\{< \quad >_S\} >_J$	•	$F = < F\xi\mu >_{\xi\mu} = < \log\{< \quad >_J\} >_{\xi\mu}$

DESORDRE

$p(J_{ij}) = 1/\sqrt{2\pi} \exp(-1/2\ J_{ij}^2)$	•	$p(\xi_i^\mu) = \pm 1$ prob. 1/2

MODELE SPHERIQUE

$\Sigma_j S_j^2 = N$	•	$\Sigma_j J_{ij}^2 = N$
Kosterlitz et al.		Gardner

MODELE DISCRET

$S_j = \pm 1$	•	$J_{ij} = \pm 1$
(Sherrington-Kirkpatrick)		(modèle $J_{ij} = \pm 1$)

PARAMETRE D'ORDRE

$q_{\alpha\beta} = \Sigma_j S_j^\alpha S_j^\beta$	•	$q_{\alpha\beta} = \Sigma_j J_{ij}^\alpha J_{ij}^\beta$

fig. 15 *Analogie entre la théorie du champ moyen des verres de spin (à gauche) et les réseaux de neurones (à droite).*

CHAPITRE II

Optimisation combinatoire : L'exemple du voyageur de commerce

1. Introduction

Comme on a pu le voir dans le premier chapitre, la théorie des modèles des systèmes désordonnés n'est jamais sans ambiguïté : dans la théorie du champ moyen des verres de spin (le modèle de Sherrington-Kirkpatrick [30]), par exemple, un des résultats les plus célèbres est la valeur de l'énergie de l'état de base (voir fig. 15 pour la définition de l'hamiltonien et de la distribution de probabilité des liens). Cette valeur est obtenue dans le cadre du schéma de brisure de la symétrie des répliques et donne une énergie par spin égale à $E/N \approx -0.7633$ [29]. A cause du caractère un peu énigmatique de la théorie il a toujours été important d'étayer le résultat théorique par des simulations numériques. Or, comme on a également pu le constater à propos du modèle $J_{ij}=\pm 1$, le problème numérique, quoique parfaitement bien défini comme problème de l'optimisation combinatoire, souffre du fait qu'il est le plus souvent impraticable.

En effet, si sur le plan théorique on est confronté aux subtilités de la brisure de la symétrie des répliques, sur le plan numérique on rencontre le problème - aussi épineux - de la complexité algorithmique. Il a été démontré [31] que le calcul de l'état de base du modèle SK appartient à la classe des problèmes NP-complets. Ceci est une classe contenant un grand nombre de problèmes en optimisation combinatoire pour lesquels il n'existe vraisemblablement pas d'algorithme polynomial.

D'un point de vue physique nous comprenons assez bien cette difficulté pour les systèmes avec brisure : au dessous d'une température critique T_c, le système se trouve dans une phase verre de spin où l'état de Gibbs est décomposé en un nombre exponentiel d'états d'équilibre, qui ne sont reliés l'un à l'autre par aucune symétrie. Il semble donc très difficile de décider dans quel état se trouve la configuration avec l'énergie la plus basse.

D'une manière intuitive, la réalité du modèle SK semble donc exprimer à la fois une vérité physique (la phase verre de spin) et numérique (appartenance à une classe de problèmes d'optimisation de grande complexité).

Il est toutefois important de souligner qu'une correspondance exacte entre les propriétés statistiques et celles de la complexité algorithmique est loin d'être établie. Dans le cas général, elle est même peu probable. Pour une vraie correspondance, le point de vue de la physique est trop différent de celui de l'optimisation : en physique statistique, sont pris en compte seulement les échantillons avec un poids statistique considérable (échantillons 'typiques') tandis que la complexité algorithmique comporte une garantie pour tous les cas possibles. Toujours est-il que l'étude de l'optimisation pour un échantillon typique est d'un intérêt égal.

Un pas important vers la jonction des deux domaines a été franchi par Kirkpatrick, Gelatt et Vecchi [32] et par Mézard et Parisi [33,34,35]. Kirkpatrick et al. [32] avaient constaté que les méthodes numériques appliquées à la simulation des verres de spin - le recuit simulé - s'appliquaient aussi à des problèmes plus traditionnels de l'optimisation combinatoire, en particulier au problème standard : le voyageur de commerce (le *TSP*, de l'anglais 'Traveling Salesman Problem'). Comme l'indique d'ailleurs son nom, le recuit simulé consiste à faire des simulations de Monte Carlo d'un système physique à température finie et à baisser progressivement la température pour retrouver la solution du problème d'optimisation de départ à température nulle.

Sur le plan théorique, Mézard et Parisi [33,34] ont réussi à transférer la théorie des répliques du modèle SK à certains modèles en optimisation combinatoire, et surtout à proposer une solution d'un certain problème du TSP 'aléatoire'.

Notre publication VI concerne ce modèle du TSP. Elle se situe au croisement des approches théoriques et numériques. La publication contient d'une part le dernier pas vers une solution (symétrique dans les répliques) de ce modèle. Dans la section suivante nous allons décrire brièvement l'approche théorique de Mézard et Parisi afin de mettre la partie théorique de la publication dans son contexte. D'autre part, l'article contient des simulations numériques extensives. Nous mettrons en perspective le travail numérique dans la section II.3.

2. Théorie du voyageur de commerce (aléatoire)

Dans le problème du voyageur de commerce on se donne un ensemble de N points (villes) i=1,...,N et les distances l_{ij} entre points i et j. On considère des chemins fermés (des *tournées*) passant une fois par tous les points, et on cherche à trouver l'itinéraire dont la longueur totale est la plus petite possible. Formellement, chaque tournée Σ est définie par une permutation *cyclique* π de (1,2,...,N), et sa longueur totale (son coût, l'énergie) est

$$L_\Sigma = \Sigma_i \, l_{i\pi(i)} \qquad (2.1)$$

La tournée minimisant le coût est la *solution optimale* du TSP.

Le TSP est un problème NP-complet, et l'algorithme le plus rapide connu [36] pour résoudre ce problème est d'une complexité $2^N N^2$. La difficulté de trouver la solution optimale du TSP croît donc très rapidement avec la taille du problème, mais elle dépend aussi de la nature de la matrice de distances (l_{ij}) : des cas spéciaux sont connus où la résolution du problème devient facile, voire triviale [37].

Un cas non-trivial qui a beaucoup été étudié est le TSP avec des distances aléatoires, le TSP *aléatoire*, où les éléments de la matrice (l_{ij}) ($l_{ij} = l_{ji}$) sont des variables aléatoires uniformément réparties sur l'intervalle [0,1]. Le niveau de difficulté algorithmique pour résoudre une instance typique du TSP aléatoire n'est pas connu.

Mézard et Parisi ont étudié le problème du TSP aléatoire du point de vue de la physique statistique : le TSP est considéré comme un système physique dont l'espace de configurations est donné par l'ensemble des

tournées Σ, et où chaque configuration porte un poids de Boltzmann égal à

$$p(\Sigma) \sim \exp(-\beta L_\Sigma) \qquad (2.2)$$

où $\beta = 1/T$ est l'inverse de la température. La formule (2.2) conduit à une fonction de partition

$$Z = \Sigma_{conf} \exp(-\beta L) \qquad (2.3)$$

et à une énergie libre $F = -(1/\beta) \log Z$. Dans la limite $N \to \infty$ on s'attend à ce que l'énergie libre à température zéro, moyennée sur la distribution de probabilité des liens, donne la valeur du coût de la solution optimale pour le TSP, dans un échantillon typique. Si la limite de température zéro est l'intérêt principal de ce système (puisqu'elle redonne le problème du départ), les propriétés à température finie sont aussi importantes : elles peuvent potentiellement fournir des informations sur les configurations de basse énergie (des configurations quasi optimales).

Pour pouvoir calculer les variables thermodynamiques du TSP aléatoire, Mézard et Parisi ont utilisé une représentation compacte des tournées qui s'appuie sur l'équivalence (tout à fait abstraite) de certains modèles de polymères avec des systèmes de spins de Heisenberg dans la limite où la dimension des spins tend vers zéro [38]. A chaque point i est alors associé un spin s_i de dimension m, avec $s_i^2 = m$, dans la limite $m \to 0$. Avec cette représentation, la fonction de partition s'écrit comme

$$Z = \lim_{\gamma \to \infty} 1/\gamma^N \int \Pi_i d\mu(s_i) \exp\{\gamma \Sigma_{i<j} \exp(-\beta l_{ij}) s_i s_j\} \qquad (2.4)$$

où la mesure $d\mu(s_i)$ indique une intégration sur la surface de la sphère à m dimensions ($m \to 0$). Les intégrales sur cette mesure ont les propriétés suivantes [38] :

$$\int d\mu(s) = 1 ; \qquad \int d\mu(s)\, s = 0 ; \qquad \int d\mu(s)\, s^2 = 1 ;$$
$$\int d\mu(s)\, s^n = 0, \quad n>2 \qquad (2.5)$$

En développant l'exponentielle dans la formule (2.4) en puissances de γ on engendre pour l'ordre n tous les graphes avec un nombre n de liens. A l'aide de l'équation (2.5) on voit facilement que seuls les graphes formant une boucle subsistent après l'intégration sur les spins. Enfin, dans la limite $\gamma \to \infty$ seules les boucles fermées de longueur N (les tournées) ont un poids non négligeable. On voit facilement que ce poids est le poids de Boltzmann de l'éq. (2.3) : l'éq. (2.4) est donc bien équivalente à l'éq. (2.3).

Avec la représentation éq. (2.4), le TSP aléatoire acquiert une certaine ressemblance avec la théorie du champ moyen des verres de spin, et le calcul de l'énergie libre, moyennée sur la distribution des liens, peut se faire en principe par la méthode des répliques. Ce calcul, pour le TSP, est très compliqué [34]. Alors, Mézard et Parisi n'ont pu obtenir qu'une solution approchée pour les propriétés de cette solution à température zéro.

Pour ce problème du voyageur de commerce, Mézard et Parisi [35] ont également appliqué une autre formulation de la théorie du champ moyen des verres de spin, *la méthode de la cavité* [22]. Celle-ci est équivalente à la théorie des répliques, mais elle part des équations valables pour chaque

échantillon. Il apparaît que, pour des raisons techniques (ou parce qu'elle utlise directement des grandeurs physiques, comme la distribution des aimantations), cette approche donne des équations plus faciles à résoudre dans la limite $T \to 0$.

La méthode de la cavité, pour le TSP, considère un système de N spins $s_1,...,s_N$, décrits par l'éq. (2.4). Dans ce système à N spins, le spin i est supposé avoir acquis une magnétisation spontanée dans une direction $\langle s_i^a \rangle = \delta^{a1} m_i^c$. On ajoute un nouveau spin s_0, et on suppose que, pour déterminer l'action des anciens spins s_i sur le nouveau spin s_0, les corrélations des spins $s_1,...,s_N$ peuvent être négligées (cette hypothèse est équivalente à l'Ansatz de la symétrie des répliques). Ceci permet de décrire le système à N+1 spins par une fonction de partition 'effective'

$$Z_{N+1} = \int \Pi_{i=0}^{N} \int d\mu(s_i) \exp(h_i s_i) \exp\{(\Sigma s_0 s_j \gamma \exp(-\beta l_{0j})\} \qquad (2.6)$$

A l'aide de l'équation (2.6) on peut maintenant déterminer la magnétisation $\langle s_0 \rangle$ (qui, elle aussi, sera spontanée dans une direction) comme fonction des magnétisations dans le système à N spins

$$\langle S0 \rangle = \frac{\partial \log Z_{N+1}}{\partial h_0}\Big|_{h_0=0} \qquad m_i^c = \frac{\partial \log Z_{N+1}}{\partial h_i}\Big|_{h_0=0} \qquad (2.7)$$

Utilisant l'éq. (2.5), ceci conduit à la relation

$$m_0 = \gamma \frac{\Sigma \exp(-\beta l_{0j})\, m_j^c}{\Sigma_{j<k} \exp(-\beta l_{0j}) \exp(-\beta l_{0k})\, m_j^c\, m_k^c} \qquad (2.8)$$

L'équation (2.8) décrit la magnétisation au point i=0 comme fonction des magnétisations m_j^c j=1,...,N du système avec N spins auquel le spin i=0 n'appartient pas (la magnétisation du spin j changera de valeur en présence du spin i=0). Pourtant, en intégrant sur la distribution des liens l_{0i} (dénoté par < >), la distribution de m_0 est égale à celle des m_j^c, dans la limite $N\to\infty$. Ceci vient du fait qu'on suppose l'existence d'une limite thermodynamique pour la distribution des aimantations. L'éq. (2.8) implique donc une *relation de cohérence* pour cette distribution de probabilité

$$P(m) = <\delta(m-m_0)> = <\delta(m-m_i^c)> \qquad (2.9)$$

C'est cette relation de cohérence que nous avons résolue directement dans la limite $T\to 0$. La fonction P(m) que nous avons ainsi obtenue porte toute l'information sur le système. Elle permet de calculer, par exemple, la longueur du chemin optimal L=2.0415... , aussi bien que la distribution des longueurs des liens dans la solution optimale (voir la section II.3).

Avec cette méthode de la cavité, les prédictions précises de la théorie symétrique dans les répliques sont établies dans la limite $T\to 0$. Notons que la méthode de la cavité nous conduit à une formulation 'physique' de l'hypothèse sous-jacente : l'existence d'un seul état thermodynamique, c'est-à-dire que les tournées quasi optimales sont presque identiques [MPV].

Passons maintenant aux vérifications numériques des prédictions analytiques.

3. Algorithmique

Un grand nombre d'approches numériques pour le TSP ont en commun qu'elles sont par nature itératives et qu'elles gardent à chaque instant de l'itération une solution 'faisable' - une tournée $\Sigma(t)$. Un algorithme de ce type, couramment utilisé dans le TSP, a été proposé par Lin [39] (voir fig. 16) : à chaque instant de temps deux liens utilisés dans le chemin $\Sigma(t)$ sont remplacés (de manière aléatoire) par deux autres liens pour construire une tournée Σ'. La configuration au pas suivant est alors

$$\begin{aligned} &\text{si } L(\Sigma') < L(\Sigma(t)) \quad : \quad \Sigma(t+1) = \Sigma' \\ &\text{sinon} \quad : \quad \Sigma(t+1) = \Sigma(t) \end{aligned} \tag{2.10}$$

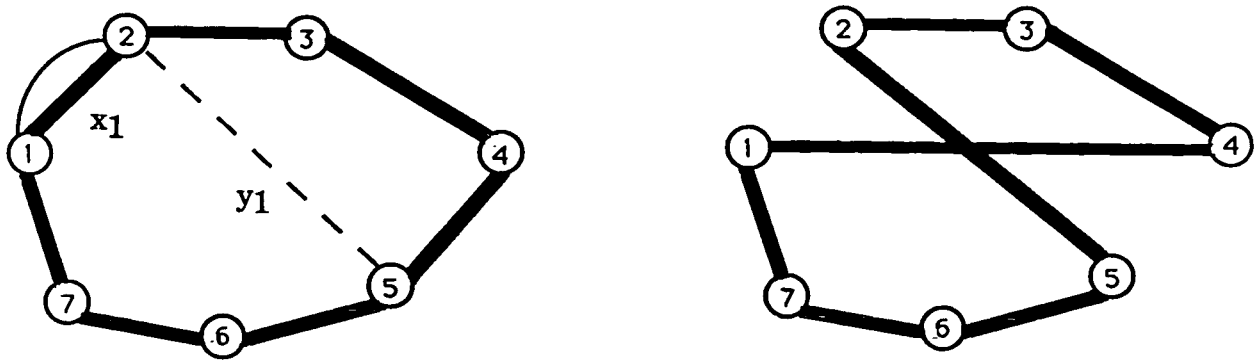


***fig. 16** Mouvement élémentaire 'de Lin' pour le problème du voyageur de commerce. Notons que le mouvement est défini par deux liens voisins : x_1 (qui doit être coupé) et y_1 (qui est ajouté à la tournée). Le lien entre les points 4 et 5 doit alors être coupé (le lien entre 1 et 4 ajouté) afin d'aboutir à une tournée à la suite du changement.*

L'algorithme de Lin (éq. (2.10)) crée donc une suite de configurations $\Sigma(0)$, $\Sigma(1)$,...., dont le coût décroît avec le temps $L(\Sigma(0)) \geq L(\Sigma(1)) \geq \ldots$.

L'itération éq. (2.10) peut être visualisée comme une dynamique dans un espace de configurations qui est schématisé dans la fig. 17 : dans cet espace deux configurations sont voisines (Σ_1,Σ_2 dans fig. 17) si elles sont accessibles l'une à partir de l'autre en un pas de temps.

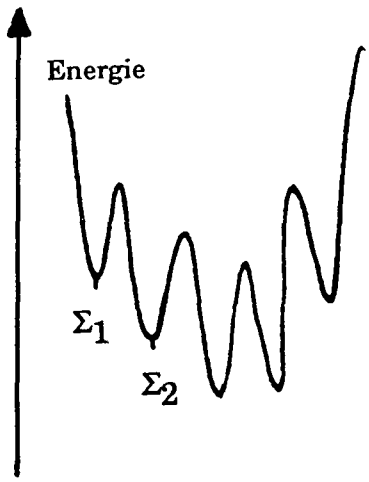


fig. 17 Paysage d'énergie d'un problème d'optimisation. Σ_1 et Σ_2 sont deux configurations métastables, séparées par une barrière d'énergie.

En général, le paysage d'énergie est très irrégulier : il existe un grand nombre de minima locaux - de configurations Σ_i telles que chaque configuration accessible à partir de Σ_i a un coût plus élevé. C'est vers un de ces états métastables que l'algorithme de Lin va en général converger. Les états métastables à coût élevé peuvent complètement écranter les configurations de basse énergie : pour le TSP aléatoire, par exemple, il n'est pas possible d'atteindre en temps polynomial des états d'énergie O(1), même en relançant l'algorithme de Lin avec des configurations initiales différentes un grand nombre de fois [40].

L'idée du recuit simulé consiste alors à remplacer la dynamique (2.10) par une dynamique de Monte Carlo à température finie

$$\text{si} \quad L(\Sigma') < L(\Sigma(t)) : \quad \Sigma(t+1) = \Sigma' \tag{2.11}$$

$$\text{sinon} : \quad \Sigma(t+1) = \Sigma' \quad \text{avec prob.} \quad \exp\{-\beta[L(\Sigma')-L(\Sigma(t))]\}$$

$$\Sigma(t+1) = \Sigma(t) \text{ avec prob. } 1-\exp\{-\beta[L(\Sigma')-L(\Sigma(t))]\}$$

c'est-à-dire à autoriser des changements défavorables avec une certaine probabilité, qui décroît avec la température. La simulation à une température non-nulle permet donc au système de franchir une barrière d'énergie et de passer vers des minima plus bas (voir la fig. 17). Au cours du temps, la température est abaissée (les changements défavorables de moins en moins souvent acceptés).

Il est évident que la proposition de Kirkpatrick et al. définit un cadre théorique plutôt qu'un algorithme proprement dit : la définition d'un 'plan de recuit' (annealing schedule) par exemple, reste un problème délicat. L'abaissement de la température ne doit pas être trop lent - afin de ne pas perdre du temps de calcul (en acceptant un nombre trop élevé de changements défavorables), en même temps un abaissement trop rapide entraînerait le système dans des états métastables à énergie très élevée.

Si on suit la logique du recuit simulé, une autre solution consisterait à adopter un choix plus large de mouvements élémentaires - comme échange de plus de deux liens à la fois, une méthode qu'on a l'habitude d'appeler l'algorithme de Lin3, Lin4, etc. [41]. Ici on rencontre le problème de la 'catastrophe combinatoire' : le choix de mouvements possibles $\Sigma_i \rightarrow \Sigma_i'$ devient énorme et l'exploration du paysage d'énergie (bien que plus régulier) impraticable [40].

Pour le TSP il existe pourtant un algorithme dû à Lin et Kernighan [42] dans lequel les mouvements élémentaires (toujours basés sur les échanges de liens) sont très compliqués, mais où la 'catastrophe combinatoire' est

évitée.

Dans l'algorithme de Lin-Kernighan on cherche - comme cela a été le cas pour les méthodes Lin3, Lin4,...- à faire des remaniements assez complexes : c'est à dire à trouver un ensemble de liens $X = \{x_1,x_2,...x_p\}$ à enlever du chemin Σ_i et un ensemble de liens $Y = \{y_1,y_2,...,y_p\}$ $(X \cap Y = \emptyset)$ à y ajouter de telle manière que Σ_i' est un chemin fermé.

La condition pour un changement favorable s'écrit comme suit :

$$\Delta L_p = \Sigma_{i=1,...,p} (L(x_i)-L(y_i)) > 0 \qquad (2.12)$$

(L(a) dénote la longueur du lien a)). L'idée pour réduire la complexité de l'algorithme vient du lemme suivant : Si l'équation (2.12) est vérifiée, il existe une permutation P de (1,2,...,p) telle que la suite des sommes partielles ΔL_k, k=1,...,p est positive :

$$\Delta L_k = \Sigma_{i=1,...,k} (L(x_{P(i)}) - L(y_{P(i)})) \geq 0; \quad k=1,...,p \qquad (2.13)$$

Dans l'algorithme de Lin et Kernighan ce fait est utilisé dans le sens inverse : si à une profondeur k l'éq. (2.13) est vérifiée (le bilan partiel est positif) la recherche est continuée ; la construction d'un remaniement est arrêtée dès que la condition (2.13) n'est plus satisfaite (voir fig. 18 pour une description plus complète de l'algorithme).

Cet algorithme nous a permis d'établir une borne supérieure pour la longueur de la tournée optimale :

$$L_{TSP} < 2.21 \quad (N \to \infty) \qquad (2.14)$$

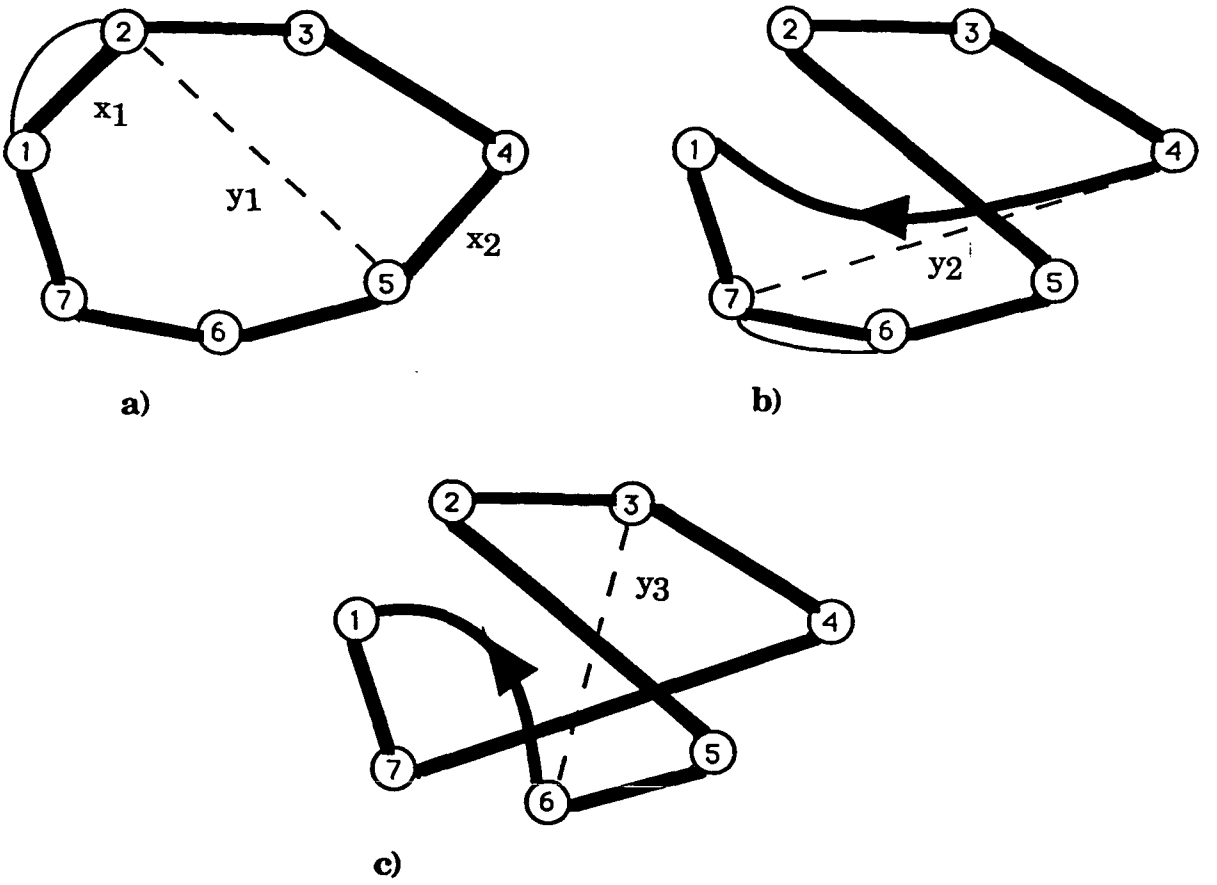


fig. 18 L'algorithme de Lin-Kernighan. ***a)*** *La configuration proposée au premier pas est* Σ_1*=(1234567). Le point 2 est choisi au hasard comme point de départ.* y_1 *est choisi comme le plus petit lien au départ de 2 ne faisant pas partie de* Σ_1 *et permettant fermer le chemin ultérieurement.* ***b)*** *La configuration proposée au deuxième pas est* Σ_2*=(1765234). Si* $(x_1\text{-}y_1)<0$ *: arrêt ; la nouvelle configuration est la plus courte de* Σ_1*,* Σ_2*. Si, par contre,* $(x_1\text{-}y_1)>0$ *: choisir* y_2 *: le lien le plus court au départ de 4 ne faisant pas partie de* Σ_1*(!) permettant fermer le chemin.* ***c)*** *La configuration proposée au troisième pas est* Σ_2*=(1743256). Si* $(x_1\text{-}y_1)+(x_2\text{-}y_2)<0$ *: arrêt ; la nouvelle configuration est la plus courte de* Σ_1*,* Σ_2*,* Σ_3*. Si, par contre,* $(x_1\text{-}y_1)+(x_2\text{-}y_2)>0$ *...(voir b).*

Il est important de constater que la longueur des chemins accessibles

par l'algorithme de Lin-Kernighan est du même ordre que la valeur théorique ; cet algorithme est donc plus performant que les méthodes utilisant l'algorithme de Lin, où la longueur optimale divergeait comme logN dans la limite thermodynamique $N \to \infty$ [40].

Si les algorithmes décrits ci-dessus gardent à chaque instant de temps une *configuration faisable* - une tournée, il existe d'autres méthodes dites *de relaxation* - qui travaillent toujours sur des classes de graphes plus larges que les tournées, tout en les incluant. Nous allons décrire ici une de ces méthodes, que nous avons utilisée pour déterminer une borne inférieure à la longueur optimale du TSP, la méthode dûe à Held et Karp [43].

La classe de graphes considérée dans cet algorithme sont des *1-arbres*. Un 1-arbre est un graphe sur les points 1,...,N qui, restreint à des points 2,...,N est un arbre complet (un graphe sans boucle connectant tous les points) et qui, de plus, lie le point 1 à ses deux plus proches voisins (voir fig. 19). Un 1-arbre de longueur minimale s'obtient facilement à partir d'un algorithme pour déterminer un arbre de longueur minimale sur les points 2,...,N.

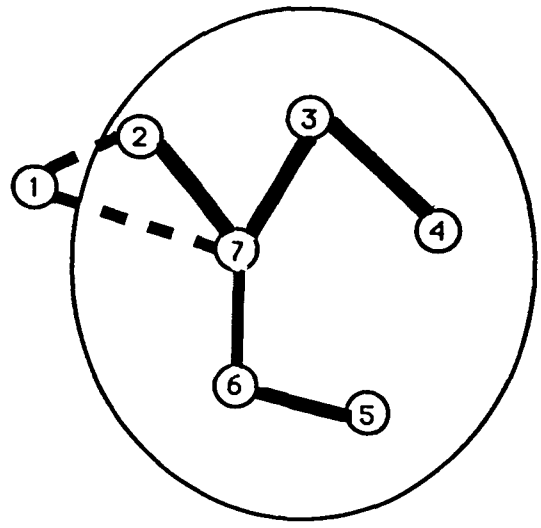


fig. 19 *Un 1-arbre est un arbre complet restreint aux points 2,...,N ; le point 1 est relié à ses deux plus proches voisins.*

Un 1-arbre de longueur minimale satisfait aux conditions suivantes :

- son coût est une borne inférieure au coût optimal du TSP
- si le 1-arbre minimal est une tournée, c'est la solution optimale du TSP.

Dans l'algorithme de Held et Karp, un 1-arbre de longueur minimale est déterminé non pas sur la matrice des connections (l_{ij}) mais sur $(l_{ij}+\lambda_i+\lambda_j)=(l_{ij})'$. L'énergie du 1-arbre minimal avec cette nouvelle matrice de distances est donc une borne inférieure de l'énergie du TSP avec la matrice $(l_{ij})'$ qui, elle, est égale à

$$L'_{TSP} = L_{TSP} + 2\Sigma_i\lambda_i \rightarrow L'_{1\text{-arbre}} - 2\Sigma_i\lambda_i \leq L_{TSP} \qquad (2.15)$$

Dans l'approche de Held et Karp, cette borne inférieure est *maximisée* en fonction des paramètres λ_i

$$L_{TSP} \geq \max_\lambda \{ \min L'_{1\text{-arbre}} - 2\Sigma_i\lambda_i \} \qquad (2.16)$$

(les λ_i sont des paramètres de Lagrange, puisqu'ils pénalisent des points dans le graphe qui ne sont pas reliés à exactement deux autres points). Il est possible de déterminer de cette manière de très bonnes bornes inférieures au TSP. Dans le cas du TSP aléatoire nous trouvons dans la limite $N \rightarrow \infty$

$$L_{TSP} \geq 2.039 \pm 3\cdot10^{-3} \quad (N\rightarrow\infty) \qquad (2.17)$$

Le 1-arbre déterminé avec cette méthode est 'presque' une tournée : avec le

choix optimal des paramètres λ_i, il n'existe qu'un petit nombre de points ($O(\sqrt{N})$ qui ne sont pas connectés à exactement deux autres.

La comparaison des résultats numériques avec la théorie porte en premier lieu sur la longueur de la tournée optimale, pour laquelle la théorie prédit la valeur L=2.0415..., tandis que le calcul numérique assure que L est compris dans la fenêtre $2.039\pm3\cdot10^{-3} < L < 2.21\pm1.3\cdot10^{-2}$. Les deux valeurs sont donc compatibles. La ressemblance du résultat théorique avec les simulations utilisant l'algorithme de Held et Karp est très frappante.

Nous avons également trouvé un très bon accord entre les prédictions théoriques et numériques pour la distribution des liens occupés dans la solution optimale (voir fig. 20).

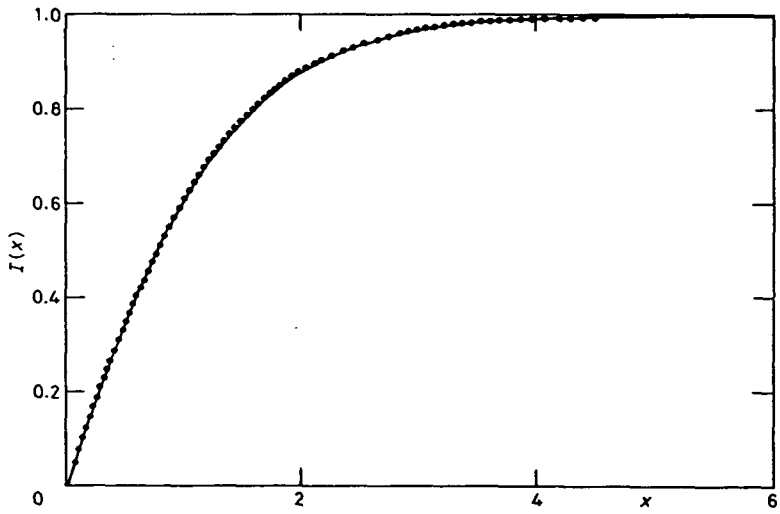


***fig. 20** Prédiction théorique pour la distribution intégrée I(x) des longueurs dans la solution optimale du TSP (en tirets) et données numériques de l'algorithme de Lin-Kernighan pour N=800 (pointillés) en fonction de la longueur renormalisée x (x=1 correspond à la longueur moyenne d'un lien dans chaque échantillon). (voir [VI])*

4. Conclusion

Nous avons esquissé la théorie symétrique dans les répliques du voyageur de commerce aléatoire, qui mène à des prédictions précises sur la longueur de la configuration optimale et la distribution des longueurs dans cette configuration. A la précision actuelle de nos simulations, ces résultats sont en accord avec les données numériques. Actuellement il est donc toujours possible que la théorie symétrique dans les répliques soit correcte.

La vérification de cette hypothèse mérite sans doute des études - analytiques et numériques - plus profondes. Sur le plan analytique il serait très intéressant de pouvoir analyser localement la stabilité de la solution symétrique (c'est-à-dire d'effectuer un calcul du genre deAlmeida-Thouless [44]), tandis que sur le plan numérique il serait sûrement possible d'améliorer la précision des données, utilisant un des algorithmes exacts développés en Recherche Opérationnelle ([45], [46]).

La possibilité d'avoir une solution symétrique n'est donc pas à exclure, et les conséquences - liées au fait qu'il y aurait dans ce cas un seul état thermodynamique jusqu'à la température zéro - seraient spectaculaires. Cette possibilité, évidemment, poserait plus de questions qu'elle n'en résoudrait : est-il possible de trouver un algorithme générant une solution quasi-optimale dans un temps polynomial ? Existe-t-il un choix de mouvements élémentaires pour le recuit simulé adapté au problème ? Existe-t-il d'autres modèles (peut-être même constructifs) du TSP qui exposeraient *toute* la difficulté supposée de ce problème ? Ces questions (même pour le TSP aléatoire) sont très importantes aussi bien en physique statistique qu'en Recherche Opérationnelle ([47], [45]).

Quoi qu'il en soit, les calculs théoriques et numériques à faire sur le TSP aléatoire n'auront rien de simple, tout comme ceux qui ont déjà été faits. Pour se détendre après (ou avant) le passage sur ce terrain difficile, nous voudrions donc reproduire un calcul simple et dépourvu de toute rigueur pour estimer la valeur moyenne de la longueur optimale du TSP aléatoire. Nous calculons L_{opt} pour N=3 et pour N=4 (voir fig. 21).

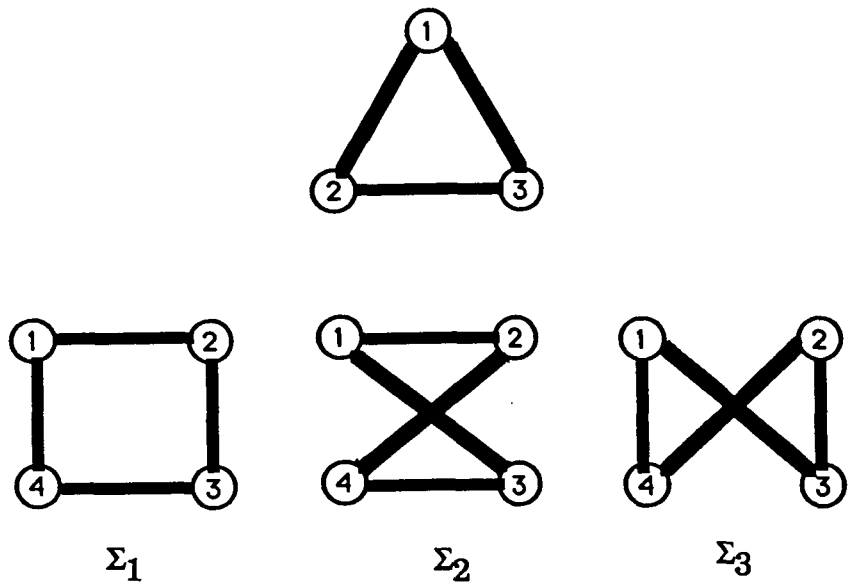


fig. 21 Le TSP avec 3 points (en haut) et 4 points (en bas). Le chemin Σ_1 est plus court que Σ_2 et Σ_3 si $l_{14}+l_{23} < l_{24}+l_{13}$ et si $l_{12}+l_{34} < l_{13}+l_{24}$.

On s'aperçoit facilement que pour N=3 il n'existe qu'une seule tournée possible, dont la longueur moyenne est $L_3 = < L_{12} + L_{23} + L_{13} > = 3/2$. Dans le cas N=4, la longueur optimale est celle de la plus courte tournée de Σ_1, Σ_2, et Σ_3. Sa valeur moyenne se calcule exactement :

$$L_4 = \frac{12}{\pi^2} \int \Pi_{i<j} dl_{ij} l_{12}\, \Theta(l_{24}+l_{13}-l_{14}-l_{23})\, \Theta(l_{13}+l_{24}-l_{12}-l_{34}) \qquad (2.18)$$

La représentation intégrale des fonctions Θ permet d'intégrer sur la distribution des liens et d'écrire l'éq. (2.18) comme suit :

$$L_4 = 12/\pi^2 \iint dxdy\, f(x/2)f(x)\{f((x+y)/2)f(y/2)\}^2 f(y) \cdot \qquad (2.19)$$
$$\cdot \{(-f(x/2)^2/2+f(x))\cos(x/2+y) + g(x)\sin(x/2+y)\}$$

avec

$$f(z)=(\sin z)/z \qquad g(z) = (-f(z)+\cos z)/z \qquad (2.20)$$

L'intégrale (2.19) a été évaluée numériquement

$$L_4= 1.65000... \qquad (2.21)$$

(un calcul analytique est sans doute possible). Il est amusant de constater qu'une 'extrapolation' en 1/N à partir des deux valeurs L_3 et L_4 donne comme valeur pour la longueur moyenne de la solution du TSP

$$L_\infty = 4 \cdot L_4 - 3 \cdot L_3 = 2.10 \qquad (2.22)$$

donc une valeur tout à fait comparable à celles calculées avec beaucoup plus de peine dans les sections 2 et 3 du présent chapitre.

Références

[1] Berlin T. H. et Kac H., Phys. Rev. **86**, 821 (1952)

[2] Hopfield J. J., Proc. Natl Acad. Sci. USA **79**, 2554

[3] Turing A., P. Lond. Math. Soc. (2) **42**, 230 (1936), ibid **43** ,544 (1937)

[4] Hopcroft J. E. et Ullman J. D., *Introduction to Automata Theory, Languages, and Computation* (Reading, MA : Addison-Wesley 1979)

[5] Hebb D., *The organization of behavior* (New York : Wiley 1949)

[6] Amit D., Gutfreund H., et Sompolinsky H., Phys. Rev. Lett. **55**, 1530 (1985)

[7] Pöppel G. et Krey U., Europhys. Lett. **4**, 979 (1987)

[8] Gardner E., Stroud N, et Wallace D. J., Edinburgh Preprint (1987)

[9] Kohonen T., *Self Organization and Associative Memory* (Berlin : Springer 1984)

[10] Personnaz L., Guyon I. et Dreyfus G., J. Physique **16**, L359 (1985)

[11] Minsky M. et Papert S., *Perceptrons* (Cambridge MA : MIT Press 1969)

[12] Duda R. O. et Hart P. E., *Pattern recognition and scene analysis* (New York : Wiley 1973)

[13] Diederich S. et Opper M., Phys. Rev. Lett. **58**, 949 (1987)

[14] Cover T. M., IEEE transactions **EC14** 3, 326 (1965)

[15] Gardner E., J. Phys. A: Math. Gen. **21**, 245 (1988)

[16] Gardner E., Europhys. Lett. **4**, 481 (1987)

[17] Gardner E. et Derrida B., J. Phys. A **21**, 271 (1988)

[18] Kepler T. et Abbott L., J. de Physique **49**, 1657 (1988)

[19] Opper M., Phys. Rev. A **38**, 3824 (1988)

[20] Landau L. D. et Lifshitz,Physique Statistique (Moscou : Mir 1967)

[21] Edwards S. et Anderson P. W., J. Phys. F **5**, 965 (1975)

[22] Mézard M., Parisi G., et Virasoro M. A., *Spin Glass theory and beyond* (Singapore: World Scientific 1987)

[23] Forrest B. M., J. Phys. A: Math. Gen. **21**, 245 (1988)

[24] Gardner E., Derrida B. et Mottishaw P., J. Physique **48**, 741 (1987)

[25] Derrida B, Gardner E. et Zippelius A., Europhys. Lett. **4**, 167 (1987)

[26] Parga N. et Virasoro M. A., J. Physique **47**, 1857 (1986)

[27] Reingold E. M., Nievergelt J., et Deo N., *Combinatorial Algorithms: Theory and Practice* (Englewood Cliffs, NJ: Prentice Hall 1977)

[28] Gardner E. et Derrida B., Saclay preprint SPhT/88-198

[29] Parisi G., J. Phys. A **13**, L115 (1980)

[30] Sherrington D. et Kirkpatrick S., Phys. Rev. Lett. **35**, 1792 (1975)

[31] Barahona F., J. Phys. A **15**, 3241 (1982)

[32] Kirkpatrick S., Gelatt C. D., et Vecchi M. P., Science **220**, 671 (1983)

[33] Mézard M. et Parisi G., J. Physique Lett. **46** , L771 (1985)

[34] Mézard M. et Parisi G., J. Physique **47**, 1285 (1986)

[35] Mézard M. et Parisi G., Europhys. Lett. **2**, 913 (1986)

[36] Held R. M. et Karp R. M., J. SIAM 10, 196 (1962)

[37] Gilmore P. C., Lawler E. L., Shmoys D. B., dans réf. [41], p. 87

[38] De Gennes P. G., Phys. Lett. A **38**, 336 (1980)

[39] Lin S., Bell System Tech. J. **44**, 2245 (1965)

[40] Kirkpatrick S., Toulouse G., J. Physique **46**, 1277 (19853)

[41] Lawler E. L., Lenstra J. K., Rinnooy Kan A. H. G., et Shmoys D. B. (eds) *The Traveling Salesman Problem* (Chichester: Wiley 1985)

[42] Lin S. and Kernighan B. W., Oper. Res. **21**, 498 (1973)

[43] Held R. M. et Karp R. M., Oper. Res. **18**, 1138 (1970)

[44] de Almeida J. R. L. et Thouless D. J., J. Phys. A **11**, 983 (1978)

[45] Balas E. et Toth P., dans réf. [41], p. 361

[46] Padberg M. W. et Grötschel M., dans réf. [41], p. 251

[47] Karp R. M. et Steele J. M., dans réf. [41], p. 181

[48] Kosterlitz J. M., Thouless D. J. et Jones R. C., Phys. Rev. Lett. **36**, 1217 (1976)

PUBLICATIONS

J. Phys. A: Math. Gen. 20 (1987) L745-L752. Printed in the UK

LETTER TO THE EDITOR

Learning algorithms with optimal stablilty in neural networks

Werner Krauth†‡ and Marc Mézard†

† Laboratoire de Physique Théorique de l'Ecole Normale Supérieure, Université de Paris-Sud, 24 rue Lhomond, 75231 Paris Cédex 05, France

‡ Department de Physique de l'Ecole Normale Supérieure, Université de Paris-Sud, 24 rue Lhomond, 75231 Paris Cédex 05, France

Received 19 May 1987

Abstract. To ensure large basins of attraction in spin-glass-like neural networks of two-state elements $\xi_i^\mu = \pm 1$, we propose to study learning rules with optimal stability Δ, where Δ is the largest number satisfying $\Delta \leqslant (\Sigma_j J_{ij}\xi_j^\mu)\xi_i^\mu$; $\mu = 1, \ldots, p$; $i = 1, \ldots, N$ (where N is the number of neurons and p is the number of patterns). We motivate this proposal and provide optimal stability learning rules for two different choices of normalisation for the synaptic matrix (J_{ij}). In addition, numerical work is presented which gives the value of the optimal stability for random uncorrelated patterns.

In the last few years, spin-glass models of neural networks have evolved into an active field of research. Much effort has been invested towards the understanding of the Hopfield model (Hopfield 1982) and its generalisations (see recent reviews such as those by Amit and Sompolinsky cited by van Hemmen and Morgenstern (1987)).

These models consist of a network of N neurons (taken to be two-state elements $S_i = \pm 1$) connected to each other through a synaptic matrix (J_{ij}). The network evolves in time according to a given dynamical rule, often taken to be a zero-temperature Monte Carlo process:

$$S_i(t+1) = \operatorname{sgn}\left(\sum_j J_{ij} S_j(t)\right). \tag{1}$$

This is the rule we will adopt in the following.

So far the interest in neural networks has been mainly focused on their properties of associative memories. This works as follows: in a so-called 'learning phase', the network is taught a number p of 'patterns' ξ^μ, $\mu = 1, \ldots, p$ (each pattern being a configuration of the network $\xi^\mu = \xi_1^\mu, \xi_2^\mu, \ldots, \xi_N^\mu$; $\xi_i^\mu = \pm 1$), i.e. the corresponding information is encoded into the matrix (J_{ij}) by means of a given learning algorithm.

In the retrieval phase, the network is started in a certain initial configuration S $(t = 0)$. If this configuration is not too different from one of the patterns, say ξ^μ, it should evolve under the dynamic rule (1) towards a fixed point, which is the pattern itself $S(t = \infty) = \xi^\mu$. We will say then that $S(t = 0)$ lies in the basin of attraction of ξ^μ. A necessary condition for associative memory in this (rather strict) sense is that the patterns be fixed points of (1) (which implies that the system is at least able to recognise the learned patterns). This can be written as

$$\xi_i^\mu = \operatorname{sgn}\left(\sum_j J_{ij}\xi_j^\mu\right) \qquad \mu = 1, \ldots, p;\ i = 1, \ldots, N \tag{2a}$$

0305-4470/87/110745+08$02.50

or, equivalently, as

$$0<\Delta\leqslant\left(\sum_j J_{ij}\xi_j^\mu\right)\xi_i^\mu \qquad \mu-1,\ldots,p;\ i-1,\ldots,N. \tag{2b}$$

An important problem in the context of associative memory is to devise learning rules which lead to large memory capacities, i.e. models whose basins of attraction are as large as possible. This is a difficult problem of 'phase-space gardening' which is the inverse problem of the spin-glass one. So far the only proposed rules (Gardner *et al* 1987, Poeppel and Krey 1987) are iterative improvement methods on the matrix (J_{ij}): if a given configuration does not converge towards the pattern, one tries to modify (J_{ij}) in order to ensure this convergence. Obviously, in order to dig a basin, one must scan a large part of the configurations of the basin and this is very time consuming (the number of configurations which differ in k bits from a pattern grows like $N^k/k!$).

In view of this difficulty, we propose in the present letter to study instead a 'poor man's version' of this problem: the network should have optimal stability Δ. As we shall see, this enables one to guarantee at least a certain minimal size of the basins of attraction; in addition, we will be able to solve this simplified problem, i.e. to provide learning algorithms which compute synaptic couplings resulting in optimal stability of the network.

A network with the dynamical rule (1) is invariant under a rescaling of the J_{ij}, and our criterion makes sense only if one has chosen a certain scale for these quantities. Let us assume, therefore, that the synaptic connections satisfy

$$|J_{ij}|\leqslant 1/\sqrt{N} \qquad i,j=1,\ldots,N \tag{3}$$

and, further, that one starts from an initial configuration $S(t=0)$ which coincides in all but a number δ of bits (components) with a pattern ξ^α. Conditions (1)-(3) then ensure that

$$S_i(t=1)=\operatorname{sgn}\left(\sum_j J_{ij}S_j(t=0)\right)=\operatorname{sgn}\left(\sum_j J_{ij}\xi_j^\alpha\right)=\xi_i^\alpha \tag{4}$$

provided

$$\delta\leqslant\Delta\sqrt{N}/2. \tag{5}$$

The inequality (5) motivates our strategy: the better the stability Δ of the network, the larger is the size of the region which can be recognised by the network *in one time step*. We will proceed on the assumption that the size of the *whole* basins of attraction of the network will then also be larger. In the absence of analytical methods to calculate basins of attraction, a detailed study of this assumption will require extensive numerical simulations, which we leave for future work. It is to be noted that our criterion is too crude to distinguish the details of the dynamical rules (parallel and sequential updating processes lead to the same result (5)) while it is sensitive to different choices of the normalisation on the synaptic matrix, which will be discussed later.

In the following we will not assume that the matrix (J_{ij}) is symmetric. The inequalities (2) then decouple into N systems, each of which states the constraints on one row vector of (J_{ij}). On row i, the stability condition can therefore be written as

$$0<\Delta_i\leqslant J_i\cdot\eta_i^\mu \qquad \mu=1,\ldots,p \tag{6}$$

where J_i is the ith row vector of (J_{ij}), and where the η_i^μ are defined by $\eta_i^\mu=\xi_i^\mu\xi^\mu$ if self-interactions $(J_{jj}\neq 0)$ are allowed and $\eta_i^\mu=\xi_i^\mu(\xi_1^\mu,\ldots,\xi_{i-1}^\mu,\xi_{i+1}^\mu,\ldots,\xi_N^\mu)$ otherwise.

We will not distinguish the two possibilities in the following and will treat η_i^μ as a vector with N components which will also be called a 'pattern' and whose row index i will generally be dropped.

We now treat the problem of computing the synaptic strengths of a network with optimal stability Δ, given the normalisation (3). This can easily be formulated as a linear program in the sense of optimisation theory (cf Papadimitriou and Steiglitz 1982). There are $N+1$ variables $(J_1, J_2, \ldots, J_N, \Delta)$ which must satisfy the set of linear inequalities

$$\begin{aligned} &\sum_i J_i \eta_i^\mu - \Delta \geqslant 0 \qquad \mu = 1, \ldots, p \\ &-1/\sqrt{N} \leqslant J_i \leqslant 1/\sqrt{N} \qquad i = 1, \ldots, N \\ &\Delta \geqslant 0 \end{aligned} \tag{7}$$

and the objective function one wants to maximise is just Δ. A feasible solution of (7) is $J = 0$, $\Delta = 0$. Therefore, an optimal solution exists; it can be computed using, e.g., the simplex algorithm (cf Papadimitriou and Steiglitz 1982). If the optimal solution is stable ($\Delta > 0$), it will satisfy $\max_j |J_j| = 1/\sqrt{N}$.

For an actual computation, it is advantageous to start from a dual formulation of (7) (cf Papadimitriou and Steiglitz 1982), in which the special form of the inequalities (7) can be used to obtain an initial basic feasible solution. It seems possible, in addition, that more sophisticated methods of combinatorial optimisation can be brought to bear on this problem to increase the speed of the learning procedure and to make efficient use of the correlations between the η_i in different rows of the matrix (J_{ij}).

Normalisations different from (3) may be of importance, in particular those which allow J_{ij} to take on discrete values only such as $J_{ij} = \pm 1, 0$. Finding optimal stability networks with these normalisations seems, however, to be a more complicated problem. We have rather, in addition to (3), treated the case where the Euclidean norm is fixed: $|J| = 1$. This problem has an interesting geometrical interpretation, in the light of which other, widely used, learning rules can be understood. The problem:

$$\begin{aligned} &\text{maximise } \Delta > 0, \text{ such that } \sum_i J_i \eta_i^\mu - \Delta \geqslant 0 \qquad \mu = 1, \ldots, p \\ &|J| = 1 \end{aligned} \tag{8}$$

corresponds, in a geometrical picture, to finding the symmetry axis J of the most pointed cone enclosing all the vectors η^μ (note that $|\eta^\mu| = \sqrt{N}$, $\mu = 1, \ldots, p$). The patterns for which the inequalities (8) are tight come to lie on the border of the cone. This is a simple geometrical problem but it transpires that finding an algorithm which solves it in a space of large dimension is not completely trivial. As a first algorithm one might choose for J the unit vector in the direction of the weighted centre of the η^μ. This, precisely, is Hebb's learning rule which is used in the Hopfield model. Clearly it has no reason to be optimal and should perform badly when some of the patterns ξ^μ are correlated, explaining a well known phenomenon. A different algorithm, the pseudoinverse method, has been proposed by Personnaz *et al* (1985) (cf also Kanter and Sompolinsky (1987)). In this case a vector J is sought, such that $J \cdot \eta^\mu = 1$, $\mu = 1, \ldots, p$. J is thus the symmetry axis of the cone, on whose border all the patterns are situated. Such a cone exists if the patterns are linearly independent (so that $p \leqslant N$ is a necessary condition). The pseudoinverse method does not result in an optimal stability Δ although it gives good results for a small number of uncorrelated patterns.

To determine a synaptic matrix with optimal stability, we present now an iterative method which is based on the perceptron-type algorithm proposed recently by Diederich and Opper (1987). Consider the following minimum-overlap learning rule, which proceeds in a finite number of time steps $t = 0, \ldots, M$, provided a solution of (8) (with $\Delta > 0$) exists.

At time $t = 0$, set $J^{(0)} = 0$ (*tabula rasa*).

At $t = 0, 1, \ldots$, determine a pattern $\eta^{\mu(t)}$ that has minimum overlap with $J^{(t)}$:

$$J^{(t)} \cdot \eta^{\mu(t)} = \min_{\nu=1,\ldots,p} \{J^{(t)} \cdot \eta^{\nu}\} \tag{9}$$

and if

$$J^{(t)} \cdot \eta^{\mu(t)} \leqslant c \qquad (c \text{ is a fixed positive number}) \tag{10}$$

use it to update $J^{(t)}$ by

$$J^{(t+1)} = J^{(t)} + (1/N)\eta^{\mu(t)} \tag{11}$$

or if

$$J^{(t)} \cdot \eta^{\mu(t)} > c \qquad (t = M)$$

renormalise $J^{(m)}$ to unity
then stop.

The stability Δ_c determined by this algorithm is

$$\Delta_c = \min_{\nu=1,\ldots,p} \{J^{(M)} \cdot \eta^{\nu}\}/|J^{(M)}| \geqslant c/|J^{(M)}|. \tag{12}$$

This algorithm differs from that of Diederich and Opper in two points. We allow c to vary instead of taking $c = 1$ (in fact we shall see that the optimal solution is obtained for $c \gg 1$) and among the patterns which satisfy (10) we choose the one which has the minimal overlap (9) instead of updating sequentially.

We now present three results, as follows.

(i) The minimal-overlap algorithm stops after a finite number M of time steps provided a stable optimal solution of (8) exists.

(ii) If Δ_{opt} is the stability of the optimal solution of (8) then Δ_c satisfies

$$\Delta_c \leqslant \Delta_{opt} \leqslant A\Delta_c \tag{13}$$

where A is a performance guarantee factor which can be measured:

$$A = |J^{(M)}|^2 N/cM \tag{14}$$

and which satisfies

$$1 \leqslant A \leqslant 2 + 1/c. \tag{15}$$

These first two results are simple consequences of a perceptron-type convergence theorem which we shall sketch in appendix 1. They also apply to the algorithm of Diederich and Opper (DO) for which we have thus obtained the performance guarantee $\Delta_{DO} \geqslant \frac{1}{3}\Delta_{opt}$.

(iii) For the minimum-overlap algorithm we have the much stronger result:

$$\text{for } c \to \infty \qquad A \to 1 \qquad \text{so that} \qquad \Delta_c \to \Delta_{opt} (c \to \infty). \tag{16}$$

The proof of (16) is somewhat more complicated; it may be found in appendix 2.

The two algorithms we have presented in this letter clearly work whatever the correlations between patterns. In order to test them and to provide a convenient

reference to other learning rules, we have performed simulations on random patterns for which each of the η_i^μ is ± 1 with equal probability. (With respect to the original network this means that we have taken the diagonal of the synaptic matrix (J_{ij}) equal to 0. Nevertheless we keep on denoting by N the total number of components of each η.) In our simulations we recorded the obtained stabilities Δ for both algorithms according to each normalisation: Δ_1 rescaled with $\max_j|J_j|\sqrt{N}$ and Δ_2 rescaled with $|J|$. The results are presented in figure 1 for a storage ratio $\alpha = p/N = 0.5$. The results for $N = 80$, $p = 40$, show, e.g., that after termination there can typically be at least 3.3 wrong bits in an initial state $S(t=0)$ to guarantee convergence to a pattern in one time step using the linear program algorithm, while the corresponding number for the minimum-overlap algorithm with $c = 10$ is 1.7 wrong bits.

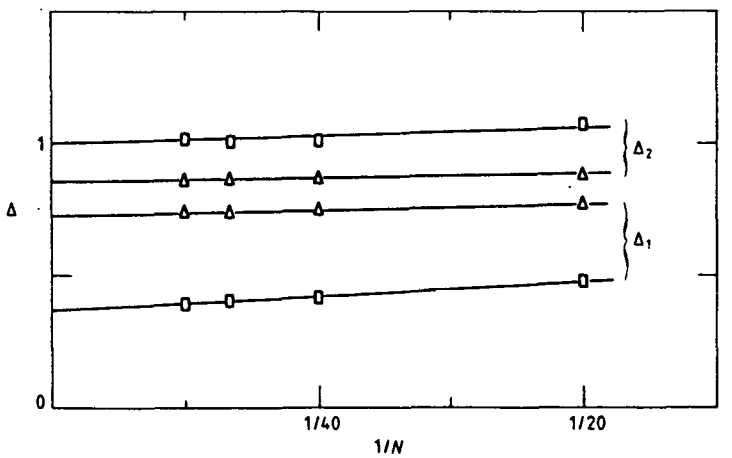


Figure 1. Stabilities Δ_1 and Δ_2 found by the two algorithms in the storage of $p = N/2$ uncorrelated random patterns, with N between 20 and 80. The triangles are the results of the simplex and the squares are the results of the minimal-overlap algorithm with $c = 1$. The upper points give Δ_2. Typical averages over 100 samples have been taken for each value of N. Lines are guides for the eye; error bars are of the size of the symbols.

For random patterns, the optimal value Δ_{opt} as a function of α for $N \to \infty$ has recently been calculated (Gardner 1987). It is the solution of

$$1/\alpha = \int_{-\Delta_{opt}}^{\infty} \frac{dt}{\sqrt{2\pi}} \exp(-\tfrac{1}{2}t^2)(t+\Delta_{opt})^2. \tag{17}$$

Using the formulae (13) and (14), we calculated (with $c = 10$) upper and lower bounds on the value of Δ_{opt} which, after statistical averaging, could be extrapolated to $N \to \infty$. The results confirm Gardner's replica calculations, as shown in figure 2. In the large N limit one can store up to $2N$ random uncorrelated patterns (Venkatesh 1986, Gardner 1987).

Finally, we want to mention a possible extension of our second algorithm and we explain it in analogy to the Hopfield model for which it has been shown that only a number of $N/2 \log N$ random patterns can be stored if one requires stability $\Delta > 0$ (Weisbuch and Fogelman-Soulie 1985), while if one allows a small fraction of wrong bits in the retrieved state then the capacity is $p = 0.14\,N$ (Amit *et al* 1985). A similar situation could occur here: it might be sensible to allow a small number of wrong bits

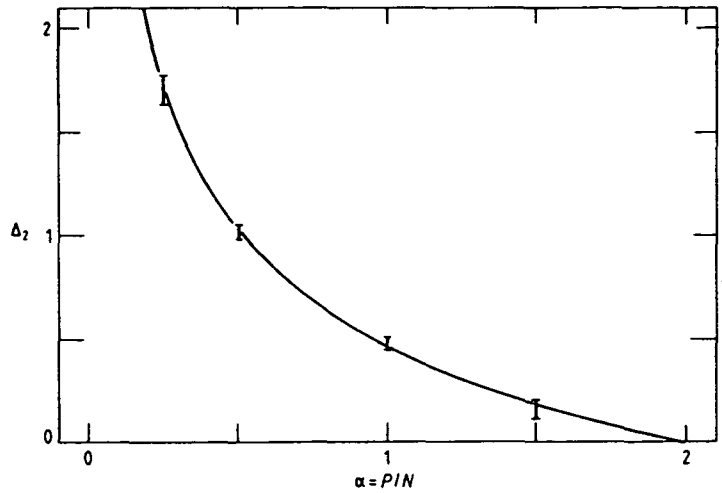


Figure 2. Asymptotic value of the optimal stability Δ_2 for the storage of uncorrelated random patterns in the large-N limit as a function of $\alpha = p/N$. The numerical results have been obtained with the minimal-overlap algorithm with $c = 10$. The error bars take into account the uncertainty on the value of Δ_{opt} due to the fact that c is not infinite (using the bounds (13)), the statistical errors found in averaging over about 100 samples for each size N, and a subjective estimate of the uncertainty of the extrapolation to $N \to \infty$. The curve is the prediction (17).

in order to enlarge the size of the basins of attraction (cf Gardner and Derrida (1987) for an analytical approach to this problem). Preliminary work indicates that quite successful methods might be conceived, using a combination of the minimal-overlap method and a simulated annealing method (Kirkpatrick *et al* 1983) with an energy function of the type $E = -\Sigma_\mu \theta(J \cdot \eta^\mu - \Delta)$. In this case the elementary moves can be those of (9)-(11) but a move is accepted only with a certain probability which depends on the change in this energy for the proposed move. It will certainly be interesting to understand how the storage capacities of uncorrelated patterns can be improved with such a method allowing a small number of errors.

It is a pleasure to thank B Derrida, E Gardner, N Sourlas and G Toulouse for stimulating discussions.

Appendix 1

We prove the convergence of the perceptron-type algorithms and provide bounds on their performance, provided there exists one stable solution. The idea of the proof follows Diederich and Opper (1987).

We assume that there exists an optimal vector J^* such that

$$J^* \cdot \eta^\mu \geqslant c \qquad \mu = 1, \ldots, p$$
$$|J^*| = c/\Delta_{opt}. \tag{A1.1}$$

After M updates with the algorithm (9)-(11), assuming that the pattern η^μ has been used m^μ times for updating ($\Sigma_\mu m^\mu = M$), one has

$$(M/N)c \leqslant (1/N)\sum_\mu m_\mu J^* \cdot \eta^\mu = J^* \cdot J^{(M)} \leqslant (c/\Delta_{opt})|J^{(M)}|. \tag{A1.2}$$

On the other hand an upper bound on $|J^{(M)}|$ is easily provided by

$$|J^{(t+1)}|^2 - |J^{(t)}|^2 = (2/N)J^{(t)} \cdot \eta^{\mu(t)} + 1/N \leq (1/N)(2c+1) \tag{A1.3}$$

which gives

$$|J^{(M)}| \leq [M/N(2c+1)]^{1/2}. \tag{A1.4}$$

Therefore the algorithm converges after a bounded number of steps M

$$M \leq (2c+1)N/\Delta_{\text{opt}}^2 \tag{A1.5}$$

and gives a stability

$$\Delta \geq c/|J^{(M)}| \geqslant \Delta_{\text{opt}}(M/N)c/|J^{(M)}|^2 = \Delta_{\text{opt}}/A \tag{A1.6}$$

where A is defined in (14). Furthermore A can be bounded; from (A1.4) and (A1.5) we obtain

$$A = |J^{(M)}|^2 N/cM \leq (2c+1)/c = 2 + 1/c. \tag{A1.7}$$

Appendix 2

To prove (16) we assume again that there exists an optimal solution J^* which satisfies (A1.1). We decompose $J^{(t)}$:

$$\begin{aligned} J^{(t)} &= a(t)J^* + K^{(t)} \\ K^{(t)} \cdot J^* &= 0 \end{aligned} \tag{A2.1}$$

and reason as in appendix 1, but separately on $K^{(t)}$ and $a(t)$.

In the minimal-overlap algorithm, $\eta^{\mu(t)}$ always has a negative projection on $K^{(t)}$:

$$K^{(t)} \cdot \eta^{\mu(t)} \leq 0 \tag{A2.2}$$

since otherwise the condition

$$\min_{\mu}\{(J^* + uK(t)) \cdot \eta^{\mu(t)}/|J^* + uK(t)|\} \leq \Delta_{\text{opt}} \tag{A2.3}$$

for all u would be violated. As in (A1.3), we can use (A2.2) to show

$$|K^{(t)}| \leq \sqrt{t/N}. \tag{A2.4}$$

If learning stops after M time steps, $a(M-1)$ can be bounded as follows:

$$J^{(M-1)} \cdot \eta^{\mu(M-1)} = a(M-1)J^* \cdot \eta^{\mu(M-1)} + K^{(M-1)} \cdot \eta^{\mu(M-1)} < c \tag{A2.5}$$

which yields

$$a(M-1) < 1 + \sqrt{M}/c. \tag{A2.6}$$

The learning rule (11) ensures that $a(M)$ differs little from $a(M-1)$. In fact

$$a(M) \leq 1 + \sqrt{M}/c + \Delta_{\text{opt}}/\sqrt{N}c. \tag{A2.7}$$

Equations (A2.4) and (A2.7) can now be combined to bound $|J^{(M)}|$ and using (A1.5) (M grows at most linearly with c) we obtain

$$c/\Delta = |J^{(M)}| \to c/\Delta_{\text{opt}} (c \to \infty) \tag{A2.8}$$

which implies the result (16).

Precise bounds and finite c corrections can be obtained using the strategy of appendix 1. They show that the relative precision on Δ is at least of the order of $1/\sqrt{c}$ for c large. In our numerical simulations we have found a precision which improved rather like $1/c$.

References

Amit D, Gutfreund H and Sompolinsky H 1985 *Phys. Rev. Lett.* **55** 1530
Diederich S and Opper M 1987 *Phys. Rev. Lett.* **58** 949
Gardner E 1987 *Preprint* Edinburgh 87/395
Gardner E and Derrida B 1987 to be published
Gardner E, Stroud N and Wallace D J 1987 *Preprint* Edinburgh 87/394
Hopfield J 1982 *Proc. Natl Acad. Sci. USA* **79** 2554
Kanter I and Sompolinsky H 1987 *Phys. Rev.* A **35** 380
Kirkpatrick S, Gelatt C D Jr and Vecchi M P 1983 *Science* **220** 671
Papadimitriou D and Steiglitz K 1982 *Combinatorial Optimization: Algorithms and Complexity* (Englewood Cliffs, NJ: Prentice Hall)
Personnaz L, Guyon I and Dreyfus J 1985 *J. Physique Lett.* **16** L359
Poeppel G and Krey U 1987 *Preprint*
van Hemmen L and Morgenstern I 1987 *Lecture Notes in Physics* vol 275 (Berlin: Springer)
Venkatesh S 1986 *Proc. Conf. on Neural Networks for Computing, Snowbird, Utah*
Weisbuch G and Fogelman-Soulie F 1985 *J. Physique Lett.* **46** L263

J. Phys. A: Math. Gen. 21 (1988) 2995-3011. Printed in the UK

The roles of stability and symmetry in the dynamics of neural networks

Werner Krauth†, Jean-Pierre Nadal‡ and Marc Mézard§

† Département de Physique, Ecole Normale Supérieure, 24 rue Lhomond, 75231 Paris Cedex 05, France

‡ Groupe de Physique des Solides, Ecole Normale Supérieure, 24 rue Lhomond, 75231 Paris Cedex 05, France

§ Laboratoire de Physique Théorique, Ecole Normale Supérieure, 24 rue Lhomond, 75231 Paris Cedex 05, France

Received 28 March 1988

Abstract. In this paper we study the retrieval phase of spin-glass-like neural networks. Considering that the dynamics should depend only on gauge-invariant quantities, we propose that two such parameters, characterising the symmetry of the neural net's connections and the stabilities of the patterns, are responsible for most of the dynamical effects. This is supported by a numerical study of the shape of the basins of attraction for a one-pattern neural network (OPN) model. The effects of stability and symmetry on the short-time dynamics of this model are studied analytically, and the full dynamics for vanishing symmetry is shown to be exactly solvable.

1. Introduction

During the past few years there has been tremendous interest in the theory of neural networks. Outstanding popularity was gained by the Hopfield model (Hopfield 1982, Amit 1987) which, by the symmetry of its interactions and by its Monte Carlo dynamics, has had great appeal for physicists trained in statistical mechanics. Examples have shown, however, that similar behaviour may be reached with quite different architectures and philosophies, such as the asymmetric strict-stability models (Kohonen 1984, Personnaz *et al* 1985, Kanter and Sompolinsky 1987). At present, many questions still appear to be open, concerning architectures, learning rules, storage prescriptions, etc.

Despite the many open issues, all neural networks will, in the retrieval phase, function basically in the same way: started close to a memory state, a given initial configuration will flow towards it, and there will be interference effects due to other memory states which will slow down and possibly inhibit convergence.

In this paper we will exclusively be interested in this retrieval phase. We will present a particular model which allows us to study retrieval without having to specify the details of the learning phase of the system. More specifically, we propose that the symmetry of the net's connections and the stabilities of the memories are responsible for most of the dynamical effects in spin-glass-like neural networks. Our one-pattern neural network (OPN) model highlights the aspects of the stability and of the symmetry, quantities whose definition will be given in the next section.

Working with the OPN model instead of with a complete neural network (with a given set of patterns and a specified learning rule) allows us to make quite general statements, but it involves approximations. It is therefore a crucial step in our argument

0305-4470/88/132995+17$02.50

that it is possible to step back from the complete network to the OPN model. We show in § 3 that, at least for the two learning rules we checked, the OPN approximation preserves the shape of the basins of attraction.

In the large-N (number of neurons) limit, the OPN model becomes quite simple, and independent of many details of the underlying net. We will present analytical calculations (§ 4) which will allow us to gain further insight into the dynamics and, hopefully, into the workings of neural nets. Our conclusions will then be summarised in § 5.

2. Stability, symmetry and gauge transformations

Let us first recall a number of well established facts and fix our notations. In all situations below we regard a network with spins $S_i = \pm 1$ ($i = 1, \ldots, N$) and an ($N \times N$) matrix (J_{ij}) of synaptic couplings J_{ij} with $J_{ii} = 0$. For simplicity we restrict ourselves to parallel dynamics at zero temperature

$$S_i(t+1) = \operatorname{sgn}\left(\sum_j J_{ij} S_j(t)\right) \qquad i = 1, \ldots, N. \tag{2.1}$$

We denote the patterns as $\xi^\mu = (\xi_1^\mu, \xi_2^\mu, \ldots, \xi_N^\mu)$, $\mu = 1, \ldots, p$, and the storing ratio as $\alpha = p/N$.

The 'stabilities' Δ_i, Δ_i^μ, are quantities defined on all rows of the matrix (J_{ij}) as

$$\Delta_i = \min_\mu \Delta_i^\mu \qquad \Delta_i^\mu = \sum_j \xi_i^\mu J_{ij} \xi_j^\mu \left(\sum_j J_{ij}^2\right)^{-1/2}. \tag{2.2}$$

If $\Delta_i^\mu > 0$, $i = 1, \ldots, N$, the pattern ξ^μ is a fixpoint of the dynamics (2.1). Many algorithms have been proposed to compute a matrix (J_{ij}) such that this condition is fulfilled (Minski and Papert 1969, Gardner *et al* 1987b, Pöppel and Krey 1987, Diederich and Opper 1987), and an optimal stability algorithm was published recently (Krauth and Mézard 1987). For random patterns ($\xi_i^\mu = \pm 1$ with probability $\frac{1}{2}$) it is not possible to store more than $2N$ patterns with $\Delta_i > 0$, $i = 1, \ldots, N$ (Venkatesh 1986, Gardner 1987). For fixed α, Δ_i will be constant as $N \to \infty$.

Recently Forrest has shown the importance of the stability for the dynamics of a symmetric neural network, obtained with the learning algorithm of Gardner *et al* (1987b): the greater the stability, the larger the basin of attraction (Forrest 1987). We have shown in addition that, on an asymmetrically diluted network (in the limit of extreme dilution), the typical stability is the only dynamically relevant parameter if the elements of the matrix (J_{ij}) are all of the same order (Mézard *et al* 1988).

We now define the 'symmetry' of the matrix (J_{ij}) as

$$\eta = \sum_{i \neq j} J_{ij} J_{ji} \left(\sum_{i \neq j} J_{ij}^2\right)^{-1} \tag{2.3}$$

η measures the relative weights of the symmetric ($2J_{ij}^{\mathrm{s}} = J_{ij} + J_{ji}$) and the antisymmetric ($2J_{ij}^{\mathrm{a}} = J_{ij} - J_{ji}$) parts of ($J_{ij}$):

$$\eta = \left(\sum_{i \neq j} (J_{ij}^{\mathrm{s}})^2 - \sum_{i \neq j} (J_{ij}^{\mathrm{a}})^2\right)\left(\sum_{i \neq j} J_{ij}^2\right)^{-1}. \tag{2.4}$$

In particular, $\eta = 1$ (respectively -1) for a fully symmetric (respectively fully antisymmetric) matrix. $\eta = 0$ means that symmetric and antisymmetric parts have the same weights, which is, for example, the case if, for $i < j$, J_{ij}^{s} and J_{ij}^{a} are random variables

of the same distribution. The relevance of η to the dynamics of spin-glass-like systems has been noted by several authors (Toulouse 1988, Gutfreund *et al* 1987, Rieger *et al* 1988). For convenience we mention the relation of η with the parameter k of Gutfreund *et al* (1987):

$$\eta = \frac{1-k^2}{1+k^2}.$$

We will several times consider gauge transformations about a state S, defined by

$$J_{ij} \to J_{ij}S_iS_j \qquad \xi_i^\mu \to \xi_i^\mu S_i. \tag{2.5}$$

The transformation (2.5) leaves the dynamics of the network unchanged. Conversely, only gauge-invariant quantities can be important for the dynamics of the neural network. The symmetry η and the stabilities Δ_i are gauge invariant. As pointed out by Derrida (1988), other such quantities are, for example, the higher-order correlations

$$\sum_{j,k,\ldots,m} J_{ij}J_{jk}\ldots J_{mi} \qquad \text{or} \qquad \sum_{j,k,\ldots,m} \xi_i^\mu J_{ij}J_{jk}\ldots J_{mn}\xi_n^\mu \tag{2.6}$$

whose influence on the dynamics of the net cannot be excluded *a priori*. We excluded another such quantity, J_{ii}, from our considerations by setting it equal to zero. On a fully connected network a diagonal term of order 1 (for non-diagonal terms of order $1/\sqrt{N}$) can be shown to influence the convergence properties favourably (Mézard *et al* 1988).

The gauge transformation (2.5) thus has a certain fundamental importance in our context. In addition, we will make use of the transformation (2.5) about pattern ξ^μ, whenever treating ξ^μ explicitly, as a simple change of coordinates. ξ^μ then conveniently transforms into $\mathbf{1} := (1, 1, \ldots, 1)$ and the definitions of magnetisation $q = \Sigma_i S_i/N$ and overlap with pattern $\xi^\mu(\Sigma_i S_i\xi_i^\mu/N)$ become equivalent.

3. One-pattern model

We define a one-pattern neural network (OPN) model to consist of a neural network with one single pattern, $\xi^1 = \mathbf{1}$. The coupling matrix (J_{ij}) does not result from a learning procedure; (J_{ij}) is a random matrix, with elements J_{ij} taken from a given distribution (such as $J_{ij} = \pm 1$), which are subjected to two types of constraints. First, the pattern $\mathbf{1}$ is to be stable with a fixed stability $\Delta_i^1 > 0$, $i = 1, \ldots, N$, so that

$$\sum_j J_{ij} = \Delta_i^1 \left(\sum_j J_{ij}^2\right)^{1/2} \qquad i = 1, \ldots, N. \tag{3.1}$$

Second, the matrix (J_{ij}) is to have a certain degree of symmetry η, as defined in equation (2.4).

While $\mathbf{1}$ is thus the only pattern which is retained explicitly in the OPN model, the quantities η and Δ are meant to represent of the order of N other patterns, and the details of the learning algorithm. By taking the values of all the stabilities greater than zero, we mimicked a strict-stability learning algorithm. Had we allowed fluctuations of the stabilities of the same order as their mean value, for example by putting

$$\sum_j J_{ij} = (\Delta + z_i)\left(\sum_j J_{ij}^2\right)^{1/2} \qquad i = 1, \ldots, N \tag{3.2}$$

with z_i a normalised random variable, a model closely related to the Hopfield model would have resulted. Note that a formula like equation (3.2) with $\Delta = 1/\sqrt{\alpha}$ holds in the Hopfield model. We will elaborate some differences between the strict-stability OPN model and the OPN model with average stability in § 4.2.

It is a key point in this paper that the dynamics of a neural network (with $p = \alpha N$ patterns, and a specific learning rule) close to a given pattern ξ^μ depends mainly on Δ_i^μ and on η, i.e. is similar to the dynamics of the corresponding OPN model (whose matrix (J_{ij}) has unchanged row sums (3.1), identical symmetry η and the same distribution function for the J_{ij}).

3.1. Numerical testing of the OPN model

There are many aspects of the dynamics one could compare, among which we choose to restrict ourselves to a detailed study of the shape of the basins of attraction. Our numerical studies indicate, however, that this comparison may bear also on other points, such as convergence times or short-time dynamics.

We performed computer simulations, making use of two different learning algorithms (see below), and of a randomising procedure which transformed a given matrix (J_{ij}) into a corresponding OPN matrix: this procedure RANDOMISE, given in table 1, scrambles the matrix, keeping the stability of the chosen pattern and establishing the symmetry at a prescribed value.

A numerical simulation run would now proceed as follows.

(i) For a certain number of random patterns, we compute a coupling matrix (J_{ij}) using one specific learning algorithm (see later) and determine its symmetry η.

(ii) We pick an index μ with $1 \leq \mu \leq p$ at random and transform ξ^μ into 1 using equation (2.5).

(iii) For typically 200 arbitrary initial states per magnetisation q_0 we follow the dynamics (equation (2.1)) during 50 time steps, if the state does not get trapped into

Table 1. Symbolic listing of subroutine RANDOMISE (see, for example, Papadimitriou and Steiglitz (1982) for the symbolic programming language used). On output the $(N \times N)$ matrix (J_{ij}) will be a scrambled version of the input matrix, with unchanged row sums and symmetry $\simeq \eta'$ (η' may be positive or negative). The first row of (J_{ij}) remains unchanged in order to eliminate a permutative degree of freedom.

```
procedure RANDOMISE
for i := 2, 3, ..., N do (comment: random permutations)
begin
        J(i, i) ↔ J(i, N)
        for k := N − 1, N − 2, ..., 2 do J(i, k) ↔ J(i, ran(k))
        (comment: ran(k) produces random integers between 1 and k)
        J(i, i) ↔ J(i, N)
end
compute η (comment: see equation (2.3)) and norm := Σ_{i,k} J(i, k)^2
while |η/η'| < 1 do (comment: Monte Carlo step)
begin
        pick a triplet (i, k, l) of mutually different random integers with 2 ≤ i ≤ N,
        1 ≤ k ≤ N, 1 ≤ l ≤ N
        del := (J(i, l) − J(i, k))J(k, i) + (J(i, k) − J(i, l))J(l, i)
        if del sgn(η') > 0 then J(i, k) ↔ J(i, l), η' := η' + del/norm
end
```

a fixed point or a cycle before this point. We record, among other quantities, the probability of perfect recall as a function of q_0 (this way of doing the simulations is due to Forrest (1987)).

(iv) Using the procedure RANDOMISE (see table 1) we scramble the matrix (J_{ij}) and create random matrices with unchanged row sums and various values of the symmetry η', in particular with the original degree of symmetry $\eta' = \eta$.

(v) We repeat (iii) for each value of η'.

3.1.1. Optimal stability algorithm. The first algorithm we considered was the optimal stability learning rule (Krauth and Mézard 1987) which, for a given set of patterns, has been proven to find the matrix (J_{ij}) with maximal values of the Δ_i (see equation (2.2); in the present context this fact is, however, of no importance). We took values of N between 100 and 400, and varied α between $\frac{1}{4}$ and $\frac{1}{2}$. Results of one such run with $N = 100$ and $p = 50$ are shown in figure 1, which shows the probability of perfect recall p_{perf} as a function of the initial overlap q_0. The data thus give the probability of flowing towards pattern 1 when starting with a random state of magnetisation q_0. The two curves result from least-squares fits to functions

$$p_{\text{perf}} \approx \tfrac{1}{2}\{\tanh[a(q_0 - q_c)] + 1\} \tag{3.3}$$

with q_c the critical overlap. In this particular example $0.82 \leqslant \Delta_i^\mu$, $\Sigma_i \Delta_i^\mu / N = 1.15$; the matrix is almost symmetric: $\eta = 0.97$. We see in figure 1 that the OPN model is capable of reproducing qualitatively the dynamics of the optimal stability matrix. The values of a (in equation (3.3)) correspond closely, and the critical overlap q_c differs slightly. OPN matrices with $\eta' = 0.5$ and $\eta' = 0$ give a poorer description of the original dynamical behaviour.

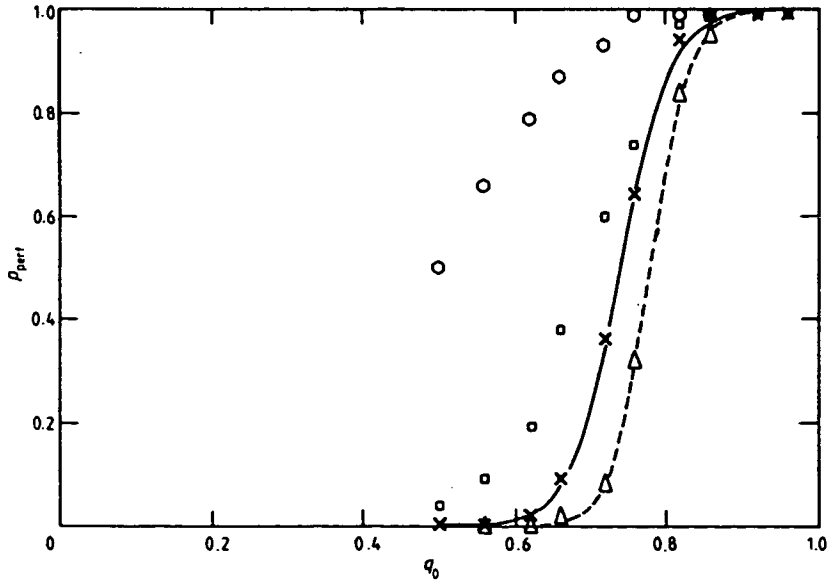


Figure 1. Comparison of the basins of attraction between the optimal stability net (×, $\eta = 0.97$, $\Sigma_i \Delta_i / N = 1.15$) and the corresponding OPN models (△, $\eta' = \eta$; □, $\eta' = 0.5$, ○, $\eta' = 0$) for $N = 100$ and $p = 50$ random patterns. Shown is the probability of perfect recall p_{perf} as a function of the initial overlap q_0 with pattern 1. The full curve is a least-squares fit with function equation (3.3) for the original dynamics and the broken curve corresponds to the OPN model dynamics with $\eta' = \eta$.

Figure 1 is typical of what we have found with the minimum overlap algorithm: there are finite differences between the original and the scrambled matrix dynamics with $\eta' = \eta$. q_c is smaller for the original matrix than for the OPN matrix if α is large, and smaller if α is small, the two dynamics being identical for α of order $\frac{1}{3}$. These differences seem to remain finite as $N \to \infty$, and do not vary much with N for the values of N we considered. Choosing values of η' different from η tends to give larger variances with the original dynamical behaviour; the dependence on η is more pronounced for larger values of the stabilities. In any case, the slope a in equation (3.3) seems to go to infinity with N, indicating the convergence towards a step function ($p_{\text{perf}} = 0$ for $q_0 < q_c$; $p_{\text{perf}} = 1$ for $q_0 > q_c$), in agreement with the results of Forrest (1987).

3.1.2. Simplex-based algorithm. We repeated the same simulations as in § 3.1 for the simplex-based learning rule (Krauth and Mézard 1987). This rule differs from the preceding algorithm in that the quantities

$$\Delta_i' = \sum_j \xi_i^\mu J_{ij} \xi_j^\mu \left(\max_j |J_{ij}| \right)^{-1} \tag{3.4}$$

are optimised. In view of the larger algorithmic complexity of the simplex-based algorithm we restricted ourselves to $20 \leq N \leq 80$ and treated again cases with $\frac{1}{4} \leq \alpha \leq \frac{1}{2}$. Results of two runs are given in figure 2. Surprisingly, there is hardly any difference between the original matrix and the OPN model dynamics (with $\eta' = \eta$), for any value of α. Again, the dynamical properties of the OPN system differ widely for large values of Δ_i, less so for small stability.

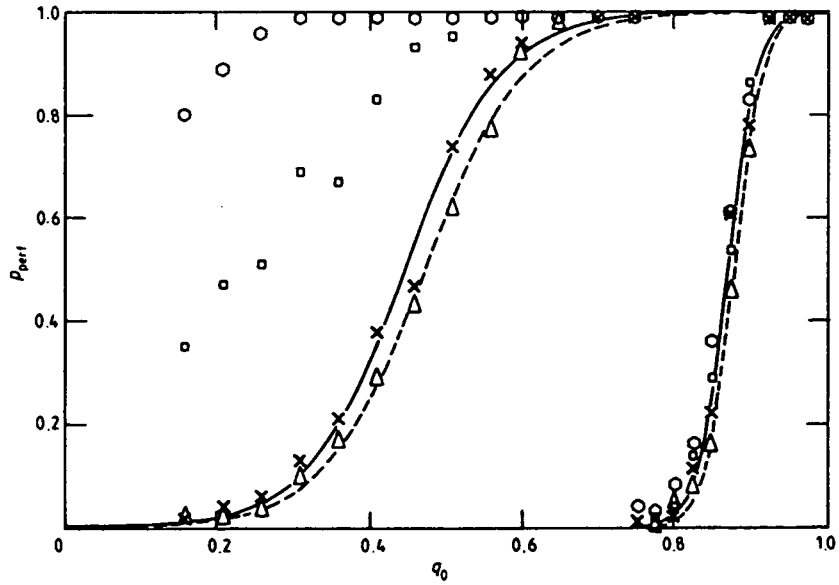


Figure 2. Comparison of neural network dynamics (×, simplex net) for $N = 80$ with $p = 20$ (left: $\eta = 0.78$, $\Sigma_i \Delta_i / N = 1.44$) and $p = 40$ (right: $\eta = 0.76$, $\Sigma_i \Delta_i / N = 0.90$) random patterns with the corresponding OPN model dynamics (△, $\eta' = \eta$; □, $\eta' = 0.5$; ○, $\eta' = 0$). Shown is the probability of perfect recall p_{perf} as a function of initial overlap q_0. The full curve represents the full neural network and the broken curve corresponds to the OPN model dynamics with $\eta' = \eta$. Note the excellent agreement of the two systems for $\eta' = \eta$ and the poor agreement for differing values of the symmetry.

3.2. Comments

In the OPN model we search for a description of the dynamical behaviour close to ξ^μ *not* as a function of all the input patterns and of the learning rule, but as a function of a few gauge-invariant combinations of ξ^μ and of (J_{ij}). It seems plausible that not all of the higher-order correlations in equation (2.6) are relevant. The procedure RANDOMISE allows us to test the hypothesis that η and Δ_i are the most important such quantities.

With a judicious choice of parameters, qualitative agreement between the full matrix model and the OPN model was perhaps to be expected. We are surprised by the excellent agreement for the simplex-based rule, and we have not yet tried to trace the origins of the differences in the case of the minimum overlap algorithm.

Note that the choice of uncorrelated patterns ξ^μ and of the two algorithms is somewhat accidental and of no real significance in the context of this paper.

4. OPN model in the large-N limit

The results of § 3 seem to us to justify an attempt to understand the OPN model in somewhat greater detail. We will present therefore our analytical calculations for a particular OPN model for which we have calculated exactly the dynamics for short times. For the first time step $t=1$, the dynamics depends solely on the stability since the sites are uncoupled. The effects of the symmetry are visible starting from $t=2$. For large t, we expect there to be a sharp transition between a region of strong recall and a region of weak recall. Our calculation up to $t=4$ provides already a good indication of this transition.

The precise OPN model we consider consists of a matrix (J_{ij}) with $J_{ij}=\pm 1$ $(J_{ii}=0)$, for which there is the same ('typical') stability on all rows:

$$\sum_j J_{ij} = \Delta\sqrt{N} \qquad i=1,\ldots,N. \tag{4.1}$$

As before, we impose a certain symmetry η. Beyond these conditions (J_{ij}) is random. This model differs only slightly from the one considered in § 3: the choice of integer values for J_{ij} will be seen to be of no significance in the limit $N\to\infty$ (only the second moment of the J_{ij} distribution enters the calculation), and equation (4.1) differs from equation (3.1) only in that the stabilities Δ_i are replaced by a typical stability Δ.

The symmetry η can be imposed on the matrix (J_{ij}) by splitting J_{ij} into symmetric and antisymmetric parts (see above equation (2.4)) and by choosing independently

$$\begin{aligned} &(J^{\mathrm{s}}_{ij}, J^{\mathrm{a}}_{ij}) = (\pm 1, 0) \qquad &&\text{with probability } \tfrac{1}{4}(1+\eta) \text{ each} \\ &(J^{\mathrm{s}}_{ij}, J^{\mathrm{a}}_{ij}) = (0, \pm 1) \qquad &&\text{with probability } \tfrac{1}{4}(1-\eta) \text{ each} \end{aligned} \tag{4.2}$$

while at the same time keeping the constraints (4.1) enforced by a product of δ functions.

In the limit $N\to\infty$ we determined the dynamics of this model for four time steps, i.e. we calculated the expectation values

$$q_1 = \langle S^{t=1}\rangle, \ldots, q_4 = \langle S^{t=4}\rangle \tag{4.3}$$

and correlation functions such as

$$q_{02} = \langle S^{t=0} S^{t=2}\rangle. \tag{4.4}$$

The mean $\langle\,\rangle$ denotes the trace taken over the couplings (4.2) *and* over initial states with $q_0 = q^{t=0}$ fixed.

4.1. Principle of the calculation, $t = 1$

Our calculations are quite similar to those on the spin-glass model by Gardner *et al* (1987a). They are as cumbersome as well. We choose to sketch only the calculation of q^1 in the main text (using the systematic way that works also for q^t, $t > 1$) in order to make plausible where the problem resides, and we relegate the more involved calculations for $t > 1$ to the appendix. As a matter of fact, the calculation at $t = 1$ can be phrased much more simply, but without any possibility of generalising it (Mézard *et al* 1988).

To determine q_1 we calculate the averaged 'partition function'

$$\bar{Z} = \mathrm{Tr}_J \, \mathrm{Tr}_{S^1} \, \delta\left(\sum_j J_{ij} - \Delta\sqrt{N}\right) \prod_i \left(S_i^1 \sum_j J_{ij} S_j^0\right) \tag{4.5}$$

where Tr_J goes over the four cases in formula (4.2). Introducing integral representations for the Kronecker δ

$$\begin{aligned}\delta(J_{ij} - \Delta\sqrt{N}) &= \int_{-\pi}^{\pi} \frac{\mathrm{d}\gamma_i}{2\pi} \exp\left[\mathrm{i}\gamma_i\left(\sum_j J_{ij} - \Delta\sqrt{N}\right)\right] \\ &= \int_{-\infty}^{\infty} \frac{\mathrm{d}\gamma_i}{2\pi\sqrt{N}} \exp\left[\mathrm{i}\gamma_i\left(\frac{1}{\sqrt{N}}\sum_j J_{ij} - \Delta\right)\right]\end{aligned} \tag{4.6}$$

and for the Heaviside function θ

$$\theta\left(S_i^1 \sum_j J_{ij} S_j^0\right) = \int_{-\infty}^{\infty} \mathrm{d}x_i \int_0^{\infty} \frac{\mathrm{d}\lambda_i}{2\pi} \exp\left[\mathrm{i}x_i\left(\frac{1}{\sqrt{N}}\sum_j J_{ij} S_j^0 - \lambda_i S_i^1\right)\right] \tag{4.7}$$

we arrive at

$$\begin{aligned}\bar{Z} = \mathrm{Tr}_J \prod_i \mathrm{Tr}_{S_i^1} \int_{-\infty}^{\infty} \frac{\mathrm{d}\gamma_i}{\sqrt{N}2\pi} \int_0^{\infty} \frac{\mathrm{d}\lambda_i}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \frac{\mathrm{d}x_i}{\sqrt{2\pi}} \exp(-\mathrm{i}\gamma_i\Delta - \mathrm{i}x_i\lambda_i S_i^1) \\ + \sum_{i<j}\left(\frac{\mathrm{i}}{\sqrt{N}} J_{ij}^{\mathrm{s}}(\gamma_i + \gamma_j + x_i S_j^0 + x_j S_i^0) + \frac{\mathrm{i}}{\sqrt{N}} J_{ij}^{\mathrm{a}}(\gamma_i - \gamma_j + x_i S_j^0 - x_j S_i^0)\right).\end{aligned} \tag{4.8}$$

In order to calculate any useful macroscopic variables (such as q_1), one will have to introduce an additional *ad hoc* source term (e.g. $\exp(h \Sigma_i \sigma_i^1)$; q_1 will then be given by $\partial \log Z(h)/\partial h$ at $h = 0$).

Taking the trace over J_{ij}^{s} and J_{ij}^{a} turns the term inside large round brackets in equation (4.8) into

$$\begin{aligned}(\,) \to \log\left[\frac{1+\eta}{2}\cos\left(\frac{\gamma_i + \gamma_j + x_i S_j^0 + x_j S_i^0}{\sqrt{N}}\right)\right. \\ \left. + \frac{1-\eta}{2}\cos\left(\frac{\gamma_i - \gamma_j + x_i S_j^0 - x_j S_i^0}{\sqrt{N}}\right)\right]\end{aligned} \tag{4.9}$$

which after expanding the trigonometric functions gives

$$\exp \sum_{i<j} (\,) \to \prod_{i \neq j} \exp\left(-\frac{1}{2N}[(\gamma_i + x_i S_j^0)^2 + \eta(\gamma_i + x_i S_j^0)(\gamma_j + x_j S_i^0)]\right). \tag{4.10}$$

All the tediousness of the calculation now stems from the term in η which forces us to introduce Gaussian integrations in order to decouple sites $1, \ldots, N$, and afterwards

to use saddle-point integrations to eliminate the extra variables (order parameters). The results for $t = 1$ are

$$q_1 = \left(\frac{2}{\pi}\right)^{1/2} \int_0^{\Delta q_0/(1-q_0^2)^{1/2}} dz\, e^{-z^2/2} =: \operatorname{erf}\left(\frac{\Delta q_0}{[2(1-q_0^2)]^{1/2}}\right) \tag{4.11}$$

$$q_{01} = q_0 q_1 . \tag{4.12}$$

As expected (Mézard *et al* 1988), q_1 does not depend on the symmetry at all, but only on Δ. There are two cases. For $\Delta \leqslant \sqrt{\pi/2}$ the function q_1 is partly below and partly above the diagonal, while for $\Delta \geqslant \sqrt{\pi/2}$ the mean magnetisation at $t = 1$ is larger than the magnetisation at $t = 0$ for all $q_0 > 0$. The connected correlation function $\langle S^0 S^1 \rangle_c := q_{01} - q_0 q_1$ is always zero. It is for this reason that simpler derivations of equation (4.11) are possible.

4.2. Solution for $\eta = 0$, all times

For $t > 1$ the calculation shown in § 4.1 has to be generalised. The changes consist in additional θ functions and in traces over spin configurations at later times. Gaussian integrations now become inevitable; they serve to decouple a growing number of terms in η (cf equation (4.10)) and, physically, to fix a growing number of order parameters, the connected correlation functions (see the appendix).

There is, however, the special case $\eta = 0$, in which the troublesome term in equation (4.10) is multiplied by 0. In this case of 'zero symmetry', it may be shown by recursion that we can solve the dynamics explicitly for all times: $q^{(1)}$ is an iteration graph and $q^{(t+1)} = q^{(1)}(q^{(t)})$:

$$q_{t+1} = \operatorname{erf}\left(\frac{\Delta q_t}{[2(1-q_t^2)]^{1/2}}\right) \qquad \eta = 0 \qquad t = 0, 1, \ldots \tag{4.13}$$

(cf equation (4.11)). Thus, the case $\eta = 0$ is equivalent to the dynamics on the asymmetrically diluted lattice, which was introduced by Derrida *et al* (1987). The fact that simplifications arise for vanishing symmetry in spin-glass models has been noted by several authors (Gutfreund *et al* 1987, Toulouse 1988, Rieger *et al* 1988, Crisanti and Sompolinsky 1988).

For zero symmetry all the connected correlation functions are zero; the system has no dynamical memory: S'^1 and S'^2 have just the same overlap as two randomly chosen states with magnetisation q'^1 and q'^2. In this case we are able to exactly calculate the critical overlap q_c, which corresponds to the unstable fixpoint of the iteration graph. For $\Delta < \sqrt{\pi/2}$, $0 < q_c < 1$ ($q = 0$, $q = 1$ are stable fixpoints) and for $\Delta > \sqrt{\pi/2}$, $q_c = 0$ (the two smaller fixpoints have merged). There remains a single stable fixpoint of the iteration (4.13), the pattern 1.

It is worth mentioning here that, if the stabilities Δ_i fluctuate, with for example

$$P(\{\Delta_i\},\ i = 1, \ldots, N) = \prod_i P(\Delta_i)$$

which leads to equation (3.2), the dynamics is again solvable at $\eta = 0$. By a similar calculation, one gets that, for $\eta = 0$ and $t \geqslant 1$,

$$q_{t+1} = \int d\Delta\, P(\Delta) \operatorname{erf}\left(\frac{\Delta q_t}{[2(1-q_t^2)]^{1/2}}\right). \tag{4.14}$$

A Hopfield-type case would correspond to a Gaussian distribution:

$$P(\Delta_i) = (2\pi)^{-1/2} \exp[-\tfrac{1}{2}(\Delta - \Delta^*)^2] \tag{4.15}$$

with

$$\Delta^* = 1/\sqrt{\alpha} \tag{4.16}$$

and this gives

$$q_{t+1} = \mathrm{erf}(\Delta^* q_t/\sqrt{2}). \tag{4.17}$$

As expected, $q_\infty < 1$. Interestingly, not all connected correlations vanish. We do have, for any $t > 0$, $q_{t0} = q_t q_0$, but for $0 < t' < t$, $q_{tt'}$ is different from $q_t q_{t'}$ and is a function of $\{q_{t'-1,t-1}, q_{t-1}, q_{t'-1}\}$.

4.3. Results for $2 \leq t \leq 4$, $\eta \neq 0$

In the case of non-zero symmetry we are not allowed to iterate q^1: the correlations between J_{ij} and J_{ji} induce non-trivial correlations between $S(t_1)$ and $S(t_2)$; however, all connected correlation functions between times differing by an odd number of steps remain zero (as in equation (4.11)). This fact is what makes our calculation up to $t = 4$ feasible. Details of the calculation may be found in the appendix and some results are shown in figures 3–6.

For $t = 2$ we find in particular

$$q_2 = \sum_{S_0} \frac{1 + q_0 S_0}{2} \mathrm{erf}\left(\frac{\Delta q_1 + i\eta V_{01}(S_0 - q_0)}{[2(1 - q_1^2)]^{1/2}} \right) \tag{4.18}$$

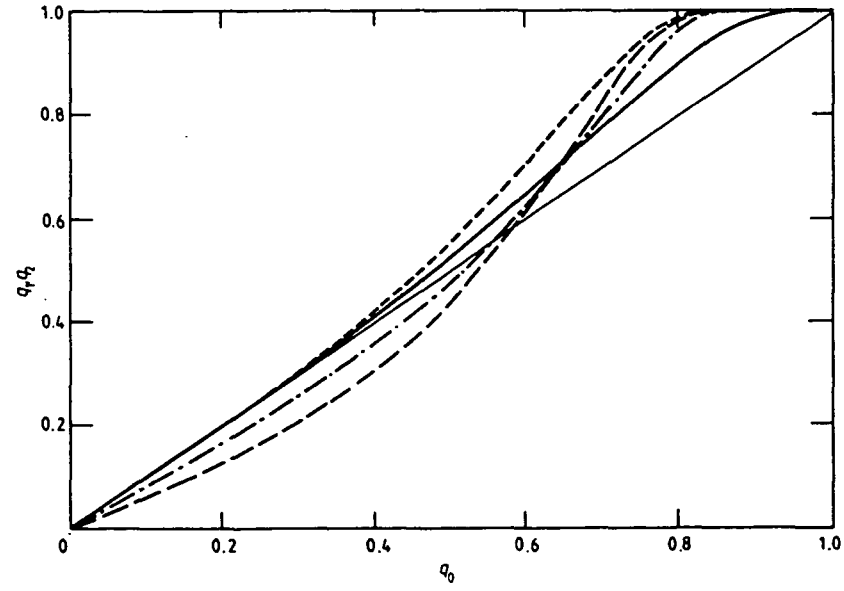


Figure 3. One-step overlap q_1 (——) and two-step overlaps q_2 as a function of initial overlap q_0 for stability $\Delta = \sqrt{\pi/2}$ and symmetry $\eta = -1$ (— — —), 0 (- - -), 1 (— - — -) in the OPN model (q_1 is independent of η).

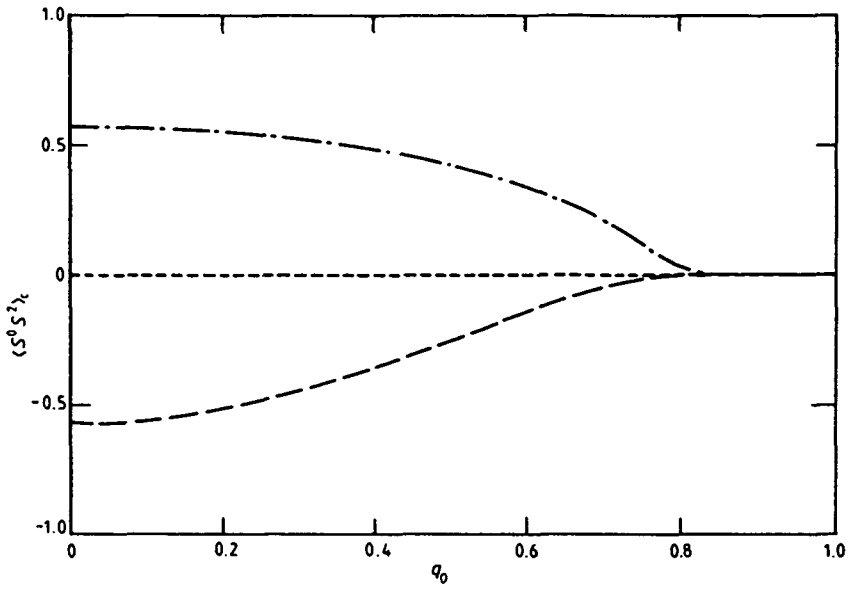


Figure 4. Connected correlation function $\langle S^0 S^2 \rangle_c = q_{02} - q_0 q_2$ for stability $\Delta = \sqrt{\pi/2}$ as a function of initial overlap q_0 for symmetry $\eta = -1$ (— — —), 0 (- - -), 1 (— - — -) in the OPN model.

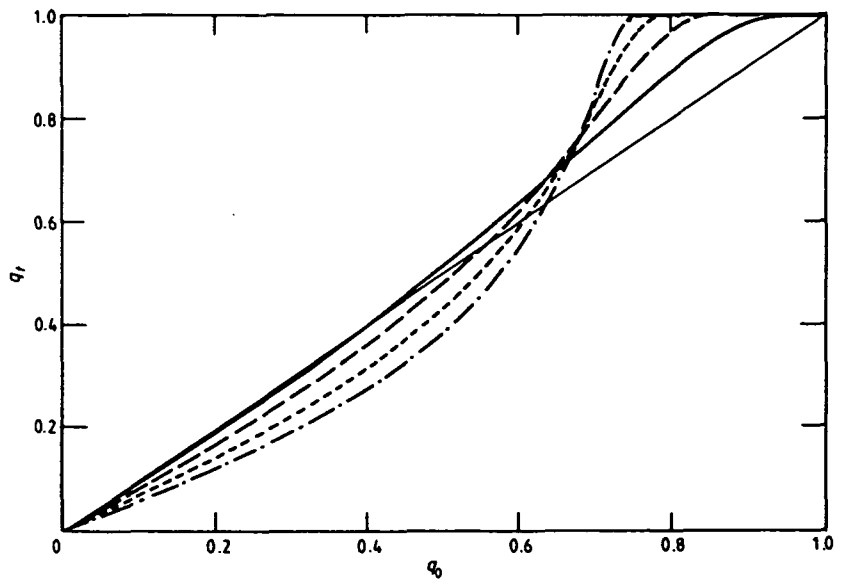


Figure 5. 1, . . . , 4-step overlap $q_1, \ldots, q_4$ as a function of initial overlap q_0 for stability $\Delta = 1.2$ and symmetry $\eta = 0.75$. ——, q_1; — — —, q_2, - - -, q_3. — · —, q_4.

where

$$iV_{01} = \left(\frac{2}{\pi(1-q_0^2)}\right)^{1/2} \exp\left(\frac{-\Delta^2 q_0^2}{2(1-q_0^2)}\right). \tag{4.19}$$

The only difference with q_1, equation (4.11), is a term which depends on η. It introduces correlations between the inital state and the one at $t=2$. Figure 3 illustrates the influence of the symmetry η at $t=2$ for $\Delta=\sqrt{\pi/2}$ (cf preceding section), and figure 4 gives the connected correlation functions for the same values of Δ and η. The high degree of correlation (anticorrelation) for large positive (negative) symmetry translates into a large probability to end up in a cycle of length 2 (of length 4 with inverted second and fourth step), as we have been able to observe numerically.

Our formulae, given in the appendix up to $t=4$, are shown in figure 5 for symmetry $\eta=0.75$ and $\Delta=1.2$. There we see in fact a step function forming, indicating a critical overlap $q_c \approx 0.7$. For a more concise representation of our formulae see figure 6. There we plotted the values of q_0, for which $q_1=q_0, \ldots, q_4=q_0$, for $\eta=0.5$, as a function of Δ. In addition, we included the numerically determined critical overlaps (at time $t=50$, using procedure RANDOMISE).

Figure 6 demonstrates that our calculation up to $t=4$ already provides a fair approximation for the critical overlap and allows at least a qualitative discussion of the OPN model also for large times. There is little influence of the symmetry for small stability, as has already been noted in § 3, and important influence for larger values of the stability. At $\Delta \approx 1.6$, the critical value of the magnetisation q_c becomes zero. This, however, has not to be taken to mean that half of the phase space flows towards 1. It should be realised that in a space of high dimensionality (N) almost all configurations have overlap with 1 which is of the order of $1/\sqrt{N}$. For $q_0=0$, the OPN model

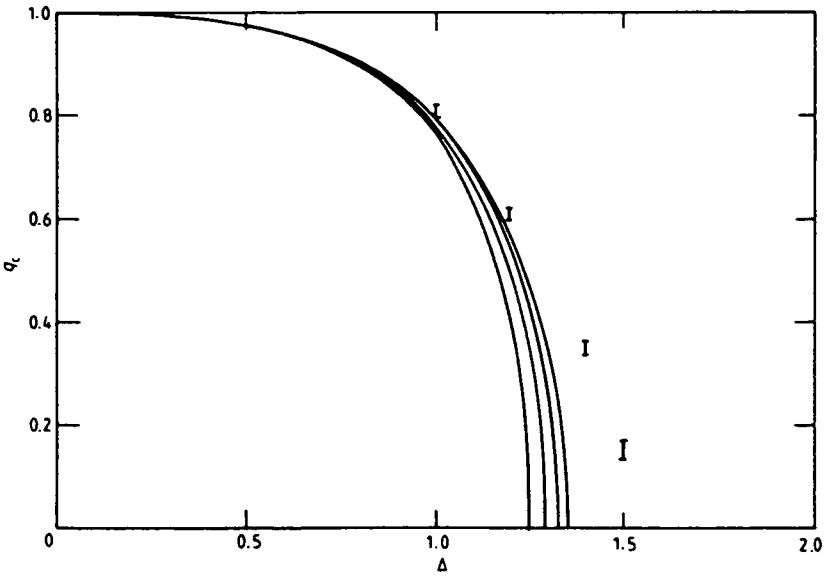


Figure 6. The curves give, from left to right, the analytical 1, 2, ..., 4-step approximations at $\eta=0.5$ for the critical overlap q_c (values of q_0 for which $q_1=q_0, \ldots, q_4=q_0$) as a function of q_0. The points with error bars are numerical values of q_c determined using procedure RANDOMISE. The values of q_c obtained here have to be compared with those at $\eta'=0.5$ shown on figures 1 and 2.

behaves as a spin glass, and our formulae for $q_0 = 0$ and $\eta = 1$ are equivalent to those of Gardner *et al* (1987a).

5. Conclusion

In this paper we presented a one-pattern neural network model with which to describe the retrieval phase of a neural network. This model stresses the importance of the stability and symmetry concepts. Using numerical methods we found it to provide in general a good qualitative description of real neural networks, and even a good quantitative agreement in some cases.

Analytical calculations enabled us to obtain the dynamics at short times, and to demonstrate that for 'zero symmetry' ($\eta = 0$) the dynamics is solvable and simple: the dynamics is equivalent to the one on the extremely diluted asymmetric network. We expect this last result to be quite general. The equivalence of the dynamics of the OPN model for zero initial overlap with the pattern and that of a spin glass provides a rather nice illustration of the similarities between spin glasses and neural networks. Without partial information about the patterns, the latter will, for all finite times, behave as the former.

In the OPN model we have shown that the critical overlap q_c (the border of the basin of attraction) depends on a function of the stability and the symmetry: typically, the larger the stability Δ and the smaller the symmetry η, the smaller q_c. We expect this to be true in a real neural network, where, of course, the two variables are no longer free, and depend on the learning rule and on the number of stored patterns.

To study the influence of the stability and the symmetry on real neural networks, we envisage several possibilities. As an example, one can consider iterative learning procedures with a *tabula non rasa* scheme (Toulouse *et al* 1986, Personnaz *et al* 1986), modulating the values of the stability and of the symmetry by proper choices of the initial matrix. Then one can look, at a given value of α, for the values leading to the largest basins of attraction. As another possibility, one could study relatively sparse networks, in which the degree of symmetry can be reduced by setting an important number of links J_{ij} equal to zero. More generally, it may be that one will have to look for learning schemes which try to reach reasonable values of the stability while trying to keep the symmetry low.

Acknowledgments

We are very grateful to B Derrida for helpful discussions, and especially for pointing out the relevance of gauge-invariant quantities. Discussions with G Toulouse are gratefully acknowledged. WK acknowledges financial support by Carl Duisberg Gesellschaft and by Studienstiftung des deutschen Volkes. The numerical simulations, done on a VAX 750 and a CONVEX C1, were supported by GRECO 70 'Expérimentation Numérique'.

Appendix

We determine the dynamics of the OPN model up to four time steps. To do so, we

calculate the 'partition function' $Z(t)$:

$$Z = \mathrm{Tr}_{S^1,S^2,\ldots,S^t} \prod_{i=1,N} \prod_{l=0,t-1} \theta\left(S_i^{l+1} \sum_j J_{ij} S_j^l\right) \tag{A1}$$

averaged over the distribution of the J_{ij}:

$$\bar{Z} = \mathrm{Tr}_J \prod_i \delta\left(\sum_j J_{ij} - \Delta\sqrt{N}\right) Z / V \tag{A2a}$$

$$V = \mathrm{Tr}_J \prod_i \delta\left(\sum_j J_{ij} - \Delta\sqrt{N}\right) \tag{A2b}$$

where Tr_J denotes the trace equation (4.2). Using the representations (4.6) and (4.7) for the δ and θ functions, respectively, we get

$$\bar{Z} = \frac{1}{V} \mathrm{Tr}_J \prod_i \int \frac{\mathrm{d}\gamma_i}{2\pi\sqrt{N}} \prod_i \prod_{l=0,t-1} \int \mathrm{d}x_i^l \int_0^\infty \frac{\mathrm{d}\lambda_i^l}{2\pi} \exp\left[\mathrm{i} \sum_i \gamma_i \left(\sum_j \frac{J_{ij}}{\sqrt{N}} - \Delta\right)\right]$$
$$\times \exp\left[\mathrm{i} \sum_{l=0,t-1} \sum_{i=1,N} x_i^l \left(\frac{1}{\sqrt{N}} \sum_j J_{ij} S_j^l - \lambda_i^l S_i^{l+1}\right)\right]. \tag{A3}$$

This leads to the following generalisation of equation (4.9) for t time steps:

$$\bar{Z} = \frac{1}{V} \mathrm{Tr}_J \prod_i \int \frac{\mathrm{d}\gamma_i}{2\pi\sqrt{N}} \prod_{i,l} \int \mathrm{d}x_i^l \int_0^\infty \frac{\mathrm{d}\lambda_i^l}{2\pi} \exp\left(-\mathrm{i} \sum_i \gamma_i \Delta - \mathrm{i} \sum_{i,l} x_i^l \lambda_i^l S_i^{l+1}\right)$$
$$\times \exp\left(\frac{\mathrm{i}}{\sqrt{N}} \sum_{i<j} J_{ij}^{\mathrm{s}} (a_{ij} + a_{ji}) + \frac{\mathrm{i}}{\sqrt{N}} \sum_{i<j} J_{ij}^{\mathrm{a}} (a_{ij} - a_{ji})\right) \tag{A4}$$

with

$$a_{ij} = \gamma_i + \sum_{l=0,t-1} x_i^l S_j^l. \tag{A5}$$

Taking the trace on the $J_{ij}^{\mathrm{s,a}}$, and keeping only the dominant terms in N, we get

$$\bar{Z} = \frac{1}{V} \prod_i \frac{\mathrm{d}\gamma_i}{2\pi\sqrt{N}} \prod_l \int \mathrm{d}x_i^l \int_0^\infty \frac{\mathrm{d}\lambda_i^l}{2\pi} \exp\left(-\mathrm{i}\Delta \sum_i \gamma_i - \mathrm{i} \sum_{i,l} x_i^l \lambda_i^l S_i^{l+1}\right)$$
$$\times \exp\left(-\frac{1}{2N} \sum_{i \neq j} (a_{ij}^2 + \eta a_{ij} a_{ji})\right). \tag{A6}$$

After replacing a_{ij} in (A6) by the RHS of (A5), one can perform the integration on the γ_i. Then it is convenient to introduce the macroscopic parameters

$$q_l = \frac{1}{N} \sum_j S_j^l \qquad 1 \leq l \leq t-1 \tag{A7a}$$

$$q_{ll'} = \frac{1}{N} \sum_j S_j^l S_j^{l'} \qquad 0 \leq l' < l \leq t-1 \tag{A7b}$$

$$V_{ll'} = \frac{1}{N} \sum_j x_j^l S_j^{l'} \qquad 0 \leq l' \neq l \leq t-1 \tag{A7c}$$

$$U_l = \frac{1}{N} \sum_j x_j^l S_j^l \qquad 0 \leq l \leq t-1 \tag{A7d}$$

$$T_l = \frac{1}{N} \sum_j x_j^l \qquad 0 \leq l \leq t-1 \tag{A7e}$$

by means of conjugate variables $\hat{q}_l$, $\hat{q}_{ll'}$, $\hat{V}_{ll'}$, $\hat{U}_l$, $\hat{T}_l$, and one obtains

$$\bar{Z} = \int \prod_{l=1}^{t-1} \mathrm{d}q_l \frac{\mathrm{d}\hat{q}_l}{2\pi} \prod_{l=0}^{t-1} \mathrm{d}U_l \frac{\mathrm{d}\hat{U}_l}{2\pi} \mathrm{d}T_l \frac{\mathrm{d}\hat{T}_l}{2\pi} \prod_{l \neq l'} \mathrm{d}V_{ll'} \frac{\mathrm{d}\hat{V}_{ll'}}{2\pi} \prod_{l<l'} \mathrm{d}q_{ll'} \frac{\mathrm{d}\hat{q}_{ll'}}{2\pi} \exp[N(G+f)] \tag{A8}$$

with

$$\begin{aligned} G = & \sum_{l=1}^{t-1} q_l \hat{q}_l + \sum_{l=0}^{t-1} (U_l \hat{U}_l + T_l \hat{T}_l) + \sum_{l<l'} q_{ll'} \hat{q}_{ll'} + \sum_{l \neq l'} V_{ll'} \hat{V}_{ll'} + \mathrm{i}\Delta \sum_l q_l T_l \\ & + \eta \left(\tfrac{1}{2} \sum_l [\eta - q_l^2(1+\eta)] t_l^2 + \frac{1}{l} \sum_{l \neq l'} T_l T_{l'} [\eta(q_{ll'} - q_l q_{l'}) - q_l q_{l'}] \right. \\ & \left. - \tfrac{1}{2} \sum_l U_l^2 - \tfrac{1}{2} \sum_{l \neq l'} V_{ll'} V_{l'l} + \sum_l q_l U_l T_l + \sum_{l \neq l'} q_l T_{l'} V_{ll'} \right) \end{aligned} \tag{A9}$$

and

$$\begin{aligned} f = & \sum_{S_0 = \pm 1} \frac{1+q_0 S_0}{2} \log \left[\sum_{\substack{S_l = \pm 1 \\ l=1,t}} \exp\left(- \sum_{l=1}^{t-1} \hat{q}_l S_l - \sum_{0 \leq l < l' < t-1} \hat{q}_{ll'} S_l S_{l'} \right) \right. \\ & \left. \times \prod_{l=0}^{t-1} \int_{-\infty}^{\infty} \mathrm{d}x_l \int_0^{\infty} \frac{\mathrm{d}\lambda_l}{2\pi} \exp\left(\sum_l \Psi_l \right) \right] \end{aligned} \tag{A10}$$

and, finally,

$$\Psi_l = -\frac{1-q_l^2}{2} x_l^2 - \sum_{l'<l} x_l x_{l'} (q_{ll'} - q_l q_{l'}) - \mathrm{i} x_l \lambda_l S_{l+1} - \hat{T}_l x_l - \hat{U}_l x_l S_l - \sum_{l' \neq l} \hat{V}_{ll'} x_l S_{l'}. \tag{A11}$$

We then take the saddle points. As in Gardner *et al* (1987a), q_l, $q_{l'l}$, $V_{l'l}$ are functions of the time steps $l'' < l$, and many order parameters are zero. Noting that the sum over S_t together with the integral over $\mathrm{d}\lambda_{t-1}$ leads to $\delta(x_{t-1})$, one deduces

$$\begin{aligned} & U_l = T_l = \hat{U}_l = \hat{q}_l = 0 \qquad l \leq t-1 \\ & V_{ll'} = 0;\ \hat{q}_{ll'} = 0;\ \hat{V}_{l'l} = 0 \qquad l' < l \end{aligned} \tag{A12}$$

and

$$\begin{aligned} & \hat{T}_l = -\mathrm{i}\Delta q_l - \eta \sum_{l'<l} V_{l'l} q_{l'} \\ & \hat{V}_{ll'} = \eta V_{l'l} \qquad l' < l. \end{aligned} \tag{A13}$$

At the saddle point, $G = 0$, and f can be written as

$$\begin{aligned} & f = \sum_{S_0} \frac{1+q_0 S_0}{2} \log \sum_{S_1, \ldots, S_t} \prod_{l=0}^{t-1} \int \mathrm{d}x_l \int \frac{\mathrm{d}\lambda_l}{2\pi} \exp\left(\sum_{l=0}^{t-1} \Psi_l \right) \\ & \Psi_l = -\frac{1-q_l^2}{2} x_l^2 - \sum_{0 \leq l' < l} x_l x_{l'} (q_{ll'} - q_l q_{l'}) - \mathrm{i} x_l \lambda_l S_{l+1} + \mathrm{i}\Delta q_l x_l - \eta \sum_{0 \leq l' < l} V_{l'l} (S_{l'} - q_{l'}) x_l \end{aligned} \tag{A14}$$

and q_l, $q_{ll'}$, $V_{ll'}$ are given by

$$\begin{aligned} & q_l = \langle\langle S_l \rangle\rangle \\ & q_{l'l} = \langle\langle S_l S_{l'} \rangle\rangle \qquad l' < l \\ & V_{l'l} = \langle\langle x_{l'} S_l \rangle\rangle \qquad l' < l \end{aligned} \tag{A15}$$

where $\langle\langle\ \rangle\rangle$ is the average taken with (A14).

Once (A15) has been obtained for $l \leqslant t-1$, we can calculate q_t, q_{lt} and V_{lt} for $l < t$ by introducing *ad hoc* source terms $\exp(hS_t)$, $\exp(hS_lS_t)$, and $\exp(hx_lS_t)$, respectively, and by taking the derivative with respect to h at $h=0$.

One can see also that $q_{ll'} - q_lq_{l'}$ is zero if l and l' are of differing parity, and that $V_{l'l}$ is zero if l and l' agree in parity. For $\eta = 0$, $q_{ll'} - q_lq_{l'}$ is always zero.

This leads to the following results for $t \leqslant 4$:

$t=1$

$$q_1 = \left(\frac{2}{\pi}\right)^{1/2} \int_0^{\Delta q_0/(1-q_0^2)^{1/2}} dz\, e^{-z^2/2}$$

$$iV_{01} = \left(\frac{2}{\pi(1-q_0^2)}\right)^{1/2} \exp\left(\frac{-\Delta^2 q_0^2}{2(1-q_0^2)}\right)$$

$t=2$

$$q_2 = \sum_{S_0} \frac{1+q_0S_0}{2}\left(\frac{2}{\pi}\right)^{1/2} \int_0^{x(S_0)} dz\, e^{-z^2/2}$$

$$q_{02} = \sum_{S_0} \frac{S_0+q_0}{2}\left(\frac{2}{\pi}\right)^{1/2} \int_0^{x(S_0)} dz\, e^{-z^2/2}$$

$$iV_{12} = \sum_{S_0} \frac{1+q_0S_0}{2}\left(\frac{2}{\pi(1-q_1^2)}\right)^{1/2} \exp\left(-\frac{1}{2(1-q_1^2)}[\Delta q_1 + iV_{01}\eta(S_0-q_0)]^2\right)$$

with

$$x(S_0) = \frac{\Delta q_1 + \eta i V_{01}(S_0 - q_0)}{(1-q_1^2)^{1/2}}$$

$t=3$

$$q_3 = \sum_{S_1,S_3} S_3 I(S_1, S_3)$$

$$q_{13} = \sum_{S_1,S_3} S_1S_3 I(S_1, S_3)$$

$$iV_{03} = \sum_{S_1,S_3} S_1S_3 \frac{\exp[-\frac{1}{2}(1-\omega_{02}^2)x_0^2]}{[2\pi(1-q_0^2)]^{1/2}} \int_{x_2-\omega_{02}x_0S_1S_3}^{\infty} \frac{dy}{\sqrt{2\pi}} e^{-y^2/2}$$

$$iV_{23} = \sum_{S_1,S_3} S_1S_3 \frac{\exp[-\frac{1}{2}(1-\omega_{02}^2)x_2^2}{[2\pi(1-q_2^2)]^{1/2}} \int_{x_0-\omega_{02}x_2S_1S_3}^{\infty} \frac{dy}{\sqrt{2\pi}} e^{-y^2/2}$$

with

$$I(S_1, S_3) = \int_{x_0}^{\infty} \frac{dy}{\sqrt{2\pi}} \int_{x_2}^{\infty} \frac{dz}{\sqrt{2\pi}} \left(\frac{D_{02}}{(1-q_0^2)(1-q_2^2)}\right)^{1/2} \exp[-\tfrac{1}{2}(y^2+z^2-2\omega_{02}yzS_1S_3)]$$

where

$$D_{02} = (1-q_0^2)(1-q_2^2) - (q_{02}-q_0q_2)^2$$

$$\omega_{02} = (q_{02}-q_0q_2)/[(1-q_0^2)(1-q_2^2)]^{1/2}$$

$$x_0 = -\Delta q_0 S_1 \left(\frac{1-q_2^2}{D_{02}}\right)^{1/2}$$

$$x_2 = -S_3[\Delta q_2 + \eta i V_{12}(S_1 - q_1)]\left(\frac{1-q_0^2}{D_{02}}\right)^{1/2}$$

$t=4$

$$q_4=\sum_{S_0,S_2,S_4}\frac{1+q_0S_2}{2}S_4J(S_0,S_2,S_4)$$

$$q_{l,4}=\sum_{S_0,S_2,S_4}\frac{1+q_0S_0}{2}S_lS_4J(S_0,S_2,S_4)$$

for $l=0$, 2, with

$$J(S_0,S_2,S_4)=\int_{x_1}\frac{dy}{\sqrt{2\pi}}\int_{x_3}\frac{dz}{\sqrt{2\pi}}\left(\frac{D_{13}}{(1-q_1^2)(1-q_3^2)}\right)^{1/2}\exp[-\tfrac{1}{2}(y^2+z^2-2\omega_{13}yzS_2S_4)]$$

where

$$D_{13}=(1-q_1^2)(1-q_3^2)-(q_{13}-q_1q_3)^2$$

$$\omega_{13}=(q_{13}-q_1q_3)/[(1-q_1^2)(1-q_3^2)]^{1/2}$$

$$x_1=-\left(\frac{1-q_3^2}{D_{13}}\right)^{1/2}S_2[(\Delta q_1+\eta i V_{01}(S_0-q_0)]$$

$$x_3=-\left(\frac{1-q_1^2}{D_{13}}\right)^{1/2}S_4[\Delta q_3+\eta i V_{03}(S_0-q_0)+\eta i V_{23}(S_2-q_2)].$$

References

Amit D J 1987 *Heidelberg Colloq. on Glassy Dynamics* ed J L van Hemmen and I Morgenstern (Berlin: Springer) p 430

Crisanti A and Sompolinsky H 1988 *Preprint* Jerusalem

Derrida B 1988 private communication

Derrida B, Gardner E and Zippelius A 1987 *Europhys. Lett.* 4 167

Diederich S and Opper M 1987 *Phys. Rev. Lett.* 58 949

Forrest B M 1988 *J. Phys. A: Math. Gen.* 21 245

Gardner E 1987 *Europhys. Lett.* 4 481

Gardner E, Derrida B and Mottishaw P 1987a *J. Physique* 48 741

Gardner E, Stroud N and Wallace D J 1987b *preprint Edinburgh 87/394*

Gutfreund H, Reger J D and Young A P 1988 *J. Phys. A: Math. Gen.* 21 2775

Hopfield J J 1982 *Proc. Natl Acad. Sci. USA* 79 2554

Kanter I and Sompolinsky H 1987 *Phys. Rev.* A 35 380

Kohonen T 1984 *Self Organization and Associative Memory* (Berlin: Springer)

Krauth W and Mézard M 1987 *J. Phys. A: Math. Gen.* 20 L745

Mézard M, Nadal J P and Krauth W 1988 in preparation

Minsky M and Papert S 1969 *Perceptrons* (Cambridge, MA: MIT Press)

Papadimitriou D and Steiglitz K 1982 *Combinatorial Optimization: Algorithms and Complexity* (Engelwood Cliffs, NJ: Prentice-Hall)

Personnaz L, Guyon I and Dreyfus G 1985 *J. Physique* 16 L359

Personnaz L, Guyon I, Dreyfus G and Toulouse G 1986 *J. Stat. Phys.* 43 411

Pöppel G and Krey U 1987 *Europhys. Lett.* 4 481

Rieger H, Schreckenberg M and Zittartz J 1988 *J. Phys. A: Math. Gen.* 21 L263

Toulouse G 1988 Private communication

Toulouse G, Dehaene S and Changeux J P 1986 *Proc. Natl Acad. Sci. USA* 83 1695

Venkatesh S 1986 *Proc. Conf. on Neural Networks for Computing, Snowbird, Utah* (*AIP Conf. Proc.* 151) ed J S Denker

Complex Systems **2** (1988) 387–408

Basins of Attraction in a Perceptron-like Neural Network

Werner Krauth
Marc Mézard
Jean-Pierre Nadal
Laboratoire de Physique Statistique,
*Laboratoire de Physique Théorique de l'E.N.S.,**
24 rue Lhomond, 75231 Paris Cedex 05, France

Abstract. We study the performance of a neural network of the perceptron type. We isolate two important sets of parameters which render the network fault tolerant (existence of large basins of attraction) in both hetero-associative and auto-associative systems and study the size of the basins of attraction (the maximal allowable noise level still ensuring recognition) for sets of random patterns. The relevance of our results to the perceptron's ability to generalize are pointed out, as is the role of diagonal couplings in the fully connected Hopfield model.

1. Introduction

An important aspect of the physicists' approach to the study of neural networks has been to concentrate on some standard situations which can be described as probability distributions of instances. For these one can then obtain quantitative comparison of the performances of different networks for large numbers of neurons and connections. A typical example is Hopfield's model [1] of associative memory. In order to quantify its performance, it has been calculated how many independent randomly chosen patterns can be stored with such an architecture, in the "thermodynamic limit" where the number N of neurons is large. For unbiased patterns the original Hebb rule allows to store $0.14N$ patterns [2,3], and more sophisticated, but still perceptron-type, rules [3–5] can reach the upper storage limit [7,8] of $2N$ patterns.

While Hopfield's model and variants of it have been studied thoroughly from a statistical physics point of view (for recent reviews see [9,10]),other widely used models such as layered networks [11] have not been analyzed in this way so far.

*Laboratoire Propre du Centre National de la Recherche Scientifique, associé à l'Ecole Normale Supérieure et à l'Université de Paris Sud.

In this paper we shall deal with the simplest such network, namely the perceptron, which consists of two layers (the usual description of a perceptron [12] contains an initial layer which insures some frozen precoding; in this paper we will not consider this first stage). In particular, we study its associative properties, which are interesting, even though the limitations of the perceptron are well known [13]. A recent review of previous studies of associative properties in other two layers networks can be found in [14].

Associativity is an important feature of neural networks as it allows for the correction of errors: even noisy input configurations can be mapped close to the desired output in the sense of Hamming distance. This is a linearly separable problem, and therefore it can be solved by a perceptron, in contrast to, e.g., the parity and the connectivity problems, which fall into a different class of computational problems, where the correlations between input configurations are not naturally related to the Hamming distance, and where the definition of noise would not be appropriate.

Hereafter we shall study the storage capacity of the perceptron, concentrating on the size of the basins of attraction. The basic result is that the size of the basin of attraction of a pattern depends primarily on its stability. (The precise definition of "stability" is given in the next section. For the pattern to be recognizable by the network in the absence of noise, its stability has to be positive.) For independent random patterns (which may be biased or not) we then calculate the typical stabilities of the patterns achieved by two learning rules, the pseudoinverse rule [15,24] and the minimal overlap rule [6] which can reach optimal stability.

Besides fully determining the associative power, knowledge about the stability achieved in the network gives us information about its capacity; an interesting outcome of our analysis is that the optimal capacity (defined as the ratio of the number of stored patterns to the number of neurons in the input layer) tends to infinity when all the output patterns coincide provided the input patterns are correlated. This result can be interpreted as reflecting the perceptron's ability to generalize: it is able to infer a simple rule from a large enough number of examples.

When studying the auto-association in a perceptron (mapping the patterns — and their nearby configurations — onto themselves) we shall see that a second parameter becomes important in order to obtain large basins of attraction: the values of the diagonal elements in the matrix of couplings, which link the neurons to themselves and tend to freeze the configurations. As the problem of auto-association can be regarded as one single parallel update of a Hopfield network, we then emphasize the relevance of these results to the fully connected Hopfield model. We show by numerical simulations that the stability and the strength of the diagonal couplings are indeed two important parameters for the dynamics of the Hopfield net. There exists an optimal value of the diagonal couplings which maximizes the radius of the basins of attraction.

The evolving simple picture — the stability of a perceptron governs its static properties (the storage capacity) as well as its dynamics (associativity)

— becomes considerably more complicated as soon as one allows several iterations of the perceptron's mapping. The correlations of the synaptic strengths start to play an important role, especially the degree of symmetry of the matrix, and it is no longer possible to make as general statements as for the perceptron. These questions have been stressed in another article [16] which is complementary to the present one. Related work on the role of the stability can be found in [17,18].

The plan of this article is as follows: In section 2 we define the network, its dynamics, the notion of attraction basins and the probability distribution of the patterns to be used for quantitative analysis. In section 3 we compute the quality of retrieval for a noisy input for two general classes of coupling matrices. Section 4 contains a detailed comparison of the associative properties of two specific learning rules: the pseudoinverse and the minimum overlap rules. In section 5 the relevance of the results to auto-association in fully connected networks is discussed. Section 6 shows how some of the results can be interpreted as the ability of generalization of the perceptron. Lastly some concluding remarks are given in section 7.

2. Dynamics of a two-layer network

We study a network of the perceptron type which consists of two layers of neurons. The neurons are Boolean units which we write as (Ising-) spins taking values ± 1. The input layer consists of N spins $\vec{\sigma} = \{\sigma_j = \pm 1, j = 1, \ldots, N\}$ and the output layer contains N' spins $\vec{\sigma}' = \{\sigma'_i = \pm 1, i = 1, \ldots, N'\}$. We shall concentrate on the limiting case where the numbers of neurons N and N' both go to infinity.

The coupling (synapse) between neuron σ_j of the input layer and the neuron σ'_i of the output layer is denoted by J_{ij} so that the coupling matrix (J_{ij}) is of size $(N' \times N)$. The output corresponding to a given input configuration is given by a (zero-)threshold automaton rule

$$\sigma'_i = \text{Sign}\,(\sum_{j=1,N} J_{ij}\sigma_i), \qquad i = 1, \ldots, N' \tag{2.1}$$

The network is taught (through the determination of the J_{ij}) to map each of the $p = \alpha N$ input patterns $\vec{\xi}^\mu = \{\xi_j^\mu = \pm 1, j = 1, \ldots, N\}$ onto a certain output pattern $\vec{\xi}'^\mu = \{\xi_i'^\mu = \pm 1, i + 1, \ldots, N\}$. We shall distinguish between two different cases: hetero-association, in which input and output patterns differ and auto-association, in which they are identical. In the latter case we have $N' = N$, and the coupling matrix is square. In this case a special role will be played by the diagonal coupling matrix elements J_{ii} which connect corresponding neurons (i) on the input and on the output layer.

Whenever we need to specialize to a specific distribution of patterns (mostly in section 4), we shall consider the case where the patterns are chosen randomly following the prescription

$$\xi'^{\mu}_j = \begin{cases} +1 \text{ with probability } & (1+m)/2 \\ -1 \text{ with probability } & (1-m)/2 \end{cases} \quad j = 1, \ldots, N \tag{2.2}$$

The probabilities are adjusted so that the patterns carry a mean magnetization $m \equiv 1/N \sum_j \xi^{\mu}_j$ (the parameter m is related to the activity of the neuron). In the case of hetero-association the output patterns are similarly chosen randomly with magnetization m'. This type of bias — and its generalization to more structured hierarchically correlated patterns — has been studied in the case of the Hopfield model [19–21].

For associativity we need that configurations close to $\vec{\xi}^{\mu}$ also be mapped close to $\vec{\xi}'^{\mu}$. To give this notion a precise meaning we shall suppose that the input configuration $\vec{\sigma}$ is chosen randomly, but with a fixed overlap q:

$$q \equiv 1/N \sum_j \xi^{\mu}_j \sigma_j \tag{2.3}$$

with the pattern $\vec{\xi}^{\mu}$ under study. This is achieved by the following choice:

$$\sigma_j = \begin{cases} +\xi^{\mu}_j \text{ with probability } & (1+q)/2 \\ -\xi^{\mu}_j \text{ with probability } & (1-q)/2 \end{cases} \quad j = 1, \ldots, N \tag{2.4}$$

i.e. we assume the noise on different neurons in the input layer to be uncorrelated and of equal strength. The average over the realizations of the noise (2.4) will be denoted by $\langle\,\rangle$.

The perceptron works as an associator, which means that configurations $\vec{\sigma}$ having a large overlap q with $\vec{\xi}^{\mu}$ should also be mapped onto $\vec{\xi}'^{\mu}$. However in the cases we consider this will be exactly true only if the input overlap q is of the order $q = 1 - O(1/\sqrt{N})$. In contrast, the noise will be reduced for a much larger number of configurations with an input overlap of the order of $q = 1 - O(1)$. This means that the output overlap obtained from equation (2.1),

$$q' \equiv 1/N' \sum_{i=1,N'} \xi'^{\mu}_i \sigma'_i \tag{2.5}$$

will be greater than q. In order to characterize this noise reduction by a number r (the "radius" of the basin of attraction), we will therefore choose a cutoff q'_c on the retrieval quality. A noisy input configuration will be said to lie inside the basin of attraction of $\vec{\xi}^{\mu}$ if $q' \geq q'_c$. As we will see in the next section, this will happen with probability 1 when q is larger than a critical value q_c. The radius of the basin of attraction is then defined as $r = 1 - q_c$.

3. Basins of attraction and stability

In order to calculate the basins of attraction of one pattern it turns out that we need rather little information on the elements of the synaptic matrix (J_{ij}). We distinguish two typical situations.

3.1 Equilibrated matrix of synaptic connections

A rather general case is that all the J_{ij} are of the same order of magnitude (i.e. $1/\sqrt{N}$). As we shall see explicitly in the study of the two learning rules, this is the typical situation for random patterns in hetero-association, or in auto-association when the diagonal couplings are set to zero. If this condition is fulfilled, the calculation of moments:

$$\begin{aligned} \langle \xi_i'^{\mu} h_i \rangle &= q\, \xi_i'^{\mu} \sum_j J_{ij} \xi_j \mu \qquad (3.1) \\ \langle (\xi_i'^{\mu} h_i - \langle \xi_i'^{\mu} h_i \rangle)^2 \rangle &= (1-q^2) \sum_j J_{ij}^2 \\ \langle (\xi_i'^{\mu} h_i - \langle \xi_i'^{\mu} h_i \rangle)^4 \rangle &= 3[(1-q^2) \sum_j J_{ij}^2]^2 (1 + O(1/N)) \\ &\cdots \end{aligned}$$

shows explicitly that $\xi_i'^{\mu} h_i$, and therefore $\xi_i'^{\mu} h_i / \sqrt{\sum_j J_{ij}^2}$, are Gaussian random variables with respect to the realization of the input noise (2.4). It is this latter quantity which we refer to as the stability of pattern μ on site i:

$$\Delta_i^{\mu} = \sum_j J_{ij}\, \xi_i'^{\mu} \xi_j^{\mu} / \sqrt{\sum_j J_{ij}^2} \qquad (3.2)$$

It follows from equation (3.1) that this Gaussian random variable has a width equal to $\sqrt{(1-q^2)}$. From equations (2.1) and (2.5) we now find that the average output overlap q' on pattern μ is related to the input overlap by

$$q' = \int P(\Delta) d\Delta \int_0^{\Delta q/\sqrt{1-q^2}} dz \sqrt{\frac{2}{\pi}}\, e^{-z^2/2} \qquad (3.3)$$

where $P(\Delta)$ reflects the site to site fluctuations of the stability of the pattern under study:

$$P(\Delta) = 1/N' \sum_i \delta(\Delta - \Delta_i^{\mu}) \qquad (3.4)$$

Equation (3.3) is the basic result of this section. It shows that the quality of retrieval of one pattern, as measured by the output overlap q', is the better the larger the stability parameters Δ_i^{μ}. The condition of perfect retrieval ($q' = 1$ when q goes to 1) is that almost all stabilities be non-negative.

Now, depending on the learning rule and the choice of patterns, the values of Δ_i^{μ} may fluctuate from site to site, and this will affect the final result for q'. Let us first suppose that the stability parameters are all equal in the thermodynamic limit:

$$\Delta_i^{\mu} = \Delta \quad i = 1, \ldots, N' \qquad (3.5)$$

In this case the radius of the basin of attraction $r = 1 - q_c$ is a function of one single parameter, the stability Δ, given implicitly by

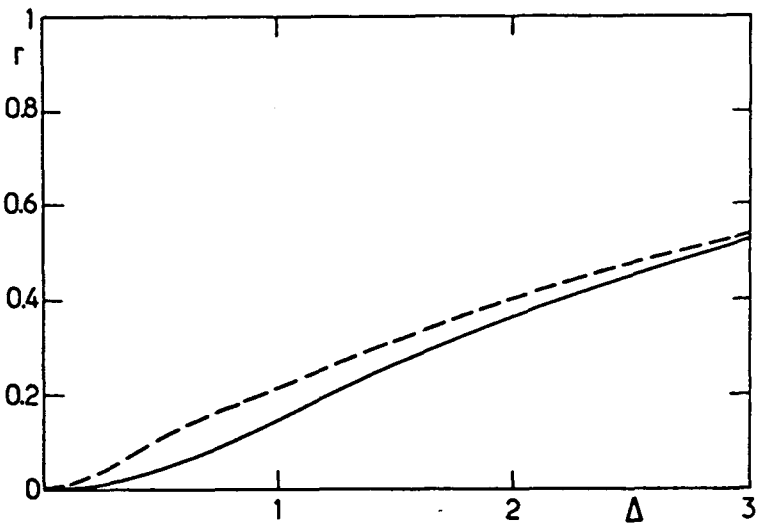


Figure 1: The radius $r = 1 - q_c$ of the basins of attraction as a function of the stability Δ. The cutoff on the output overlap is $q_c' = 0.9$ (see (3.6)). Full curve: zero diagonal couplings $J_{ii} = 0$; dashed curve: optimal diagonal couplings ($J_{ii} = J_{\text{opt}}$), in the case of autoassociation.

$$q_c' = \text{erf}\left(\frac{\Delta(1-r)}{\sqrt{2r(2-1)}}\right) \tag{3.6}$$

This function is plotted in figure 1 for $q_c' = 0.9$ (i.e. less than 5% wrong bits in the output).

As a simple example showing how site-to-site fluctuations of Δ can ruin this result, let us consider the case of Hebb's rule:

$$J_{ij} = 1/N \sum_{\mu} \xi_i'^{\mu} \xi_j^{\mu} \tag{3.7}$$

for unbiased *random* patterns ($m = m' = 0$). Then a simple calculation shows that the distribution of stabilities is Gaussian with mean $1/\sqrt{\alpha}$ and width 1. Thus the fluctuations and the mean value of the stability are of the same order. In this special case of Hebb's rule, the fluctuations over configurations and fluctuations from site to site combine in such a way that $\sum_j J_{ij} \xi_i'^{\mu} \xi_j^{\mu}$ is a Gaussian with unit width (independent of q). The final resulting noise on the output pattern is independent of the input overlap:

$$q' = \int_0^{q/\sqrt{\alpha}} dz \sqrt{\frac{2}{\pi}}\, e^{-z^2/2} \tag{3.8}$$

The most important qualitative difference between equation (3.8) and equation (3.3) is that for q approaching 1 the output overlap q' does not tend to 1: the memorized pattern differs slightly from the correct output pattern, as in Hopfield's model [1].

3.2 Unequilibrated matrix of synaptic connections

In this section we shall explore the case which is of importance for the perceptron in auto-associative mode and for the Hopfield model. There the diagonal elements of the matrix of couplings play a special role since ξ_i^μ and $\xi_i'^\mu$ are identical. It often happens that these diagonal elements are much larger than the off-diagonal elements. This special role of J_{ii} has been recognized also by Kanter and Sompolinsky [22] (see also [14] and reference therein).

Let us therefore assume that $J_{ii} = J_0(\sqrt{\sum_{j(\neq i)} J_{ij}^2})$, while $J_{ij} = O(1/\sqrt{N})$ $(j \neq i)$. Then the terms J_{ii} in equation (2.1) must be treated separately, and the formula generalizing equation (3.3) is

$$q' = \int P(\Delta) d\Delta \sum_{\tau=\pm 1} \frac{1+q\tau}{2} \int_0^{\frac{\Delta q + J_0 \tau}{\sqrt{1-q^2}}} dz \sqrt{\frac{2}{\pi}}\, e^{-z^2/2} \tag{3.9}$$

The quality of retrieval now depends both on the stability and the diagonal coupling. A well chosen value of J_0 can increase the basin of attraction. Supposing again that all the stabilities are equal to Δ, it is easy to see that the slope evaluated at $q = 1$ is zero if $J_0 < \Delta$, but it is equal to one if $J_0 > \Delta$; if the diagonal coupling is too large, the network cannot flow towards the correct output pattern, even when started from a configuration very close to the input pattern, and its noise will not be reduced. For fixed Δ there exists an optimal value of the diagonal coupling, between 0 and Δ, which maximizes the basin of attraction. The plot of the optimal value as a function of Δ is given in figure 2, and the corresponding new value of the radius of the basin of attraction (evaluated for any Δ, and for J_0 taken at its optimal value) is given in figure 1.

4. Comparison of learning rules: pseudoinverse and minimum overlap

4.1 Definition of the learning rules

Several learning rules have been proposed to choose the J_{ij}'s which allow for the memorization of a given set of patterns $\{\vec{\xi}^\mu, \vec{\xi}'^\mu, \mu = 1, p\}$. A necessary condition for perfect memorization is

$$\xi_i'^\mu = \mathrm{Sign}\Big(\sum_{j=1,N} J_{ij}\xi_j^\mu\Big), \quad i = 1, \ldots, N' \tag{4.1}$$

which is equivalent to having $\Delta_i^\mu > 0$, $i = 1, \ldots, N'$.

One efficient learning rule, the pseudoinverse (P.I.), has been proposed by Kohonen [15] in the context of linear networks. The idea is to look for a matrix (J_{ij}) which is the solution of the equation

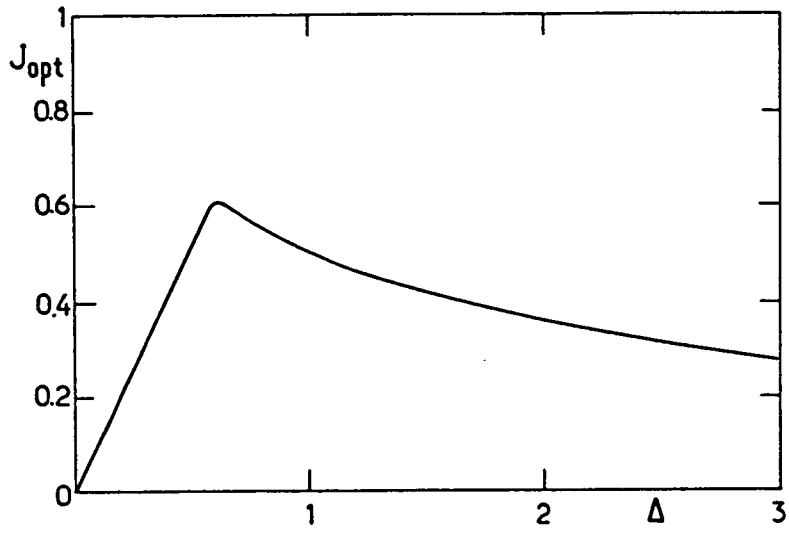


Figure 2: Value of the optimal diagonal couplings $J_{ii} = J_{opt}$ (for autoassociation) as a function of the stabilities Δ.

$$t_i^\mu \equiv \xi_i'^\mu \sum_{j=1,N} J_{ij}\xi_j{}^\mu = 1,\ \mu = 1,\ldots,p,\ i = 1,\ldots,N' \tag{4.2}$$

(or more generally the (J_{ij}) which minimizes $\sum_\mu (1 - t_i^\mu)^2$, for each i). When the input patterns are linearly independent, which will always be the case in the situations we study below, an explicit form of the matrix (J_{ij}) is

$$J_{ij} = 1/N \sum_{\mu,\nu} (Q^{-1})_{\mu\nu} \xi_i'^\mu \xi_j{}^\mu \tag{4.3}$$

where $Q_{\mu\nu}$ is the matrix of overlaps of the input patterns:

$$Q_{\mu\nu} = 1/N \sum_{j=1,N} \xi_j^\mu \xi_j^\nu \tag{4.4}$$

(The most general solution of (4.2) contains an additional arbitrary projector onto the subspace orthogonal to the input patterns. For definiteness we keep to the case where this term vanishes).

Another important family of learning rules uses error correcting algorithms to find iteratively a matrix (J_{ij}) such that all the stabilities are positive, the convergence of these algorithms being assured (if a solution exists) by the famous perceptron convergence theorem [12]. Recently this has been refined in order to obtain optimal stability parameters [6]. The corresponding "minimal overlap" (M.O.) algorithm finds a matrix of couplings (J_{ij}) which guarantees that the smallest stability parameter is maximized:

$$K_i = \inf_\mu \Delta_i^\mu \text{ is maximized}, i = 1,\ldots,N' \tag{4.5}$$

We will now calculate the values of the stability which can be reached with these algorithms for sets of random patterns introduced in section 2 (see formula (2.2)).

4.2 Hetero-association

We begin with the case of hetero-association, and with the minimal overlap algorithm. There we follow [8]. For each output neuron i we calculate the fraction of the volume of the N-dimensional space Ω of couplings such that $\Delta_i^\mu \geq K$ for all the patterns $\mu = 1,\ldots p = \alpha N$. Given α there is a critical value of K above which this fraction of Ω has zero volume for $N \to \infty$: this is the maximal value of K, K_{opt}, which can possibly be reached by any algorithm, and which *is* reached by the M.O. algorithm. The calculation of K_{opt} is a mild generalization of the work of Gardner [8] which we shall not reproduce here. The result is that K_{opt} is related to α by the equation:

$$\frac{1}{\alpha} = \sum_{\tau=\pm 1} \frac{1 + m'\tau}{2} \int_{\frac{-K_{\text{opt}} + mM\tau}{\sqrt{1-m^2}}}^{\infty} dz \frac{1}{\sqrt{2\pi}} e^{-z^2/2} \left(\frac{K_{\text{opt}} - mM\tau}{\sqrt{1-m^2}} + z \right)^2 \tag{4.6}$$

where M is an auxiliary parameter fixed by the condition that it should make the above expression stationary:

$$\sum_{\tau=\pm 1}(1+m'\tau)\tau\int_{\frac{-K_{\rm opt}+mM\tau}{\sqrt{1-m^2}}}^{\infty}dz\frac{1}{\sqrt{2\pi}}e^{-z^2/2}\left(\frac{K_{\rm opt}-mM\tau}{\sqrt{1-m^2}}+z\right)=0 \quad (4.7)$$

The resulting dependence of $K_{\rm opt}$ as function of α, for various values of m and m', is plotted in figure 3. The maximal storage capacity for this network is $p=\alpha_c N$, where α_c is the critical value for which one gets $K_{\rm opt}=0$. The M.O. algorithm can in fact produce a coupling matrix for which Δ_i^μ is strictly larger than $K_{\rm opt}$ for a few sites i and a few patterns μ. So the basins of attraction will always be larger than (and almost equal to) the values obtained under the assumption that $\mathcal{P}(\Delta)=\delta(\Delta-K_{\rm opt})$. These values can be taken from figures 1 and 3 and the result is plotted in figure 4.

If one wants an exact measure of the radius one must calculate the distribution of stabilities $\mathcal{P}(\Delta)$ reached by the M.O. algorithm. As has been noted by Kepler and Abbott [18] this can be done using the same kind of replica formalism which has been used to determine the value of $K_{\rm opt}$: of the space of couplings such that $\Delta_i^\mu \geq K$ for all the patterns there is a subspace of volume Ω' such that the stability of pattern 1 is $\Delta_i^1=\Delta$. Then

$$\mathcal{P}(\Delta)=\lim_{K\to K_{\rm opt}}\overline{(\Omega'/\Omega)} \quad (4.8)$$

where the $\overline{(\,)}$ means an average over the realizations of random patterns. This is in turn calculated as

$$\mathcal{P}(\Delta)=\lim_{K\to K_{\rm opt}}\lim_{n\to 0}\overline{(\Omega'\,\Omega^{n-1})} \quad (4.9)$$

which allows a replica calculation analogous to the one of Gardner [8]. We shall not reproduce the details of this calculation here, but just quote the results: for the M.O. algorithm, the distribution of stabilities is

$$\begin{aligned}\mathcal{P}(\Delta) &= \sum_{\tau=\pm 1}\frac{1+m'\tau}{2}\{\theta(\Delta-K_{\rm opt})\frac{e^{\frac{-(\Delta-mM\tau)^2}{2(1-m^2)}}}{\sqrt{2\pi(1-m^2)}} \\ &+ \delta(\Delta-K_{\rm opt})\int_{\frac{-K_{\rm opt}+mM\tau}{\sqrt{1-m^2}}}^{\infty}\frac{dz}{\sqrt{2\pi}}e^{-z^2/2}\}\end{aligned} \quad (4.10)$$

where $K_{\rm opt}$ and M are related to the capacity α by equations (4.6) and (4.7). This formula has been derived independently in reference [25]. Using equations (4.10) and (3.3) one can determine the radius of the basins of attraction, which is plotted in figure 4 and differs very little from the one obtained by putting $\mathcal{P}(\Delta)=\delta(\Delta-K_{\rm opt})$.

Let us now turn to the pseudoinverse rule. As we have seen before this leads to a stability which is equal to one before normalization, so that, in order to calculate the parameters Δ_i^μ it is necessary to calculate the normalization of the couplings. From equation (4.3) we have:

$$\sum_j J_{ij}^2=\frac{1}{N}\sum_{\mu,\nu=1,\alpha N}\xi_i^{l\mu}\xi_i^{l\nu}(Q^{-1})_{\mu\nu}\equiv A \quad (4.11)$$

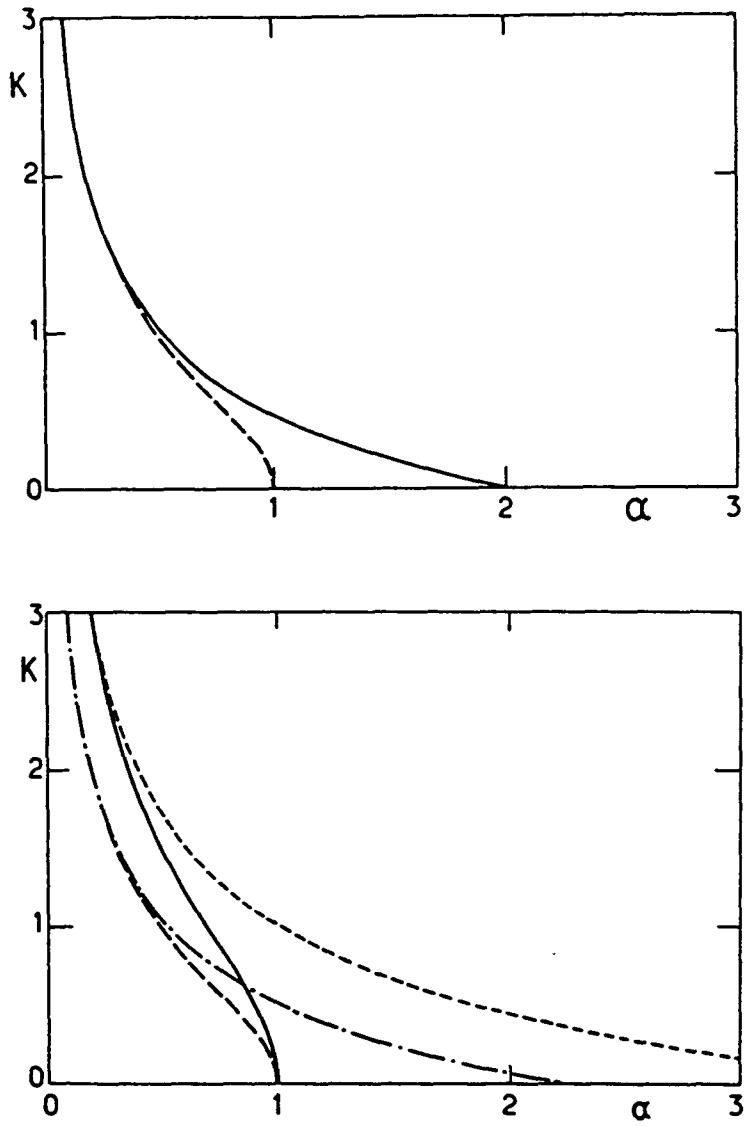


Figure 3: Lower bound on the stabilities, K, as a function of the number of stored patterns per input neuron, $\alpha = p/N$. a) Unbiased input and output patterns ($m = m' = 0$). Full curve: K_{opt}, optimal K (M.O. algorithm); dashed curve: K reached by the P.I. algorithm. b) Biased input ($m = 0.4$) and output ($m' = 0.4$ and 0.8) patterns. Full curve: P.I. algorithm ($m' = 0.8$); dashed curve: P.I. algorithm ($m' = 0.4$); dashed-dotted curve: M.O. algorithm ($m' = 0.4$); dotted curve: M.O. algorithm ($m' = 0.8$).

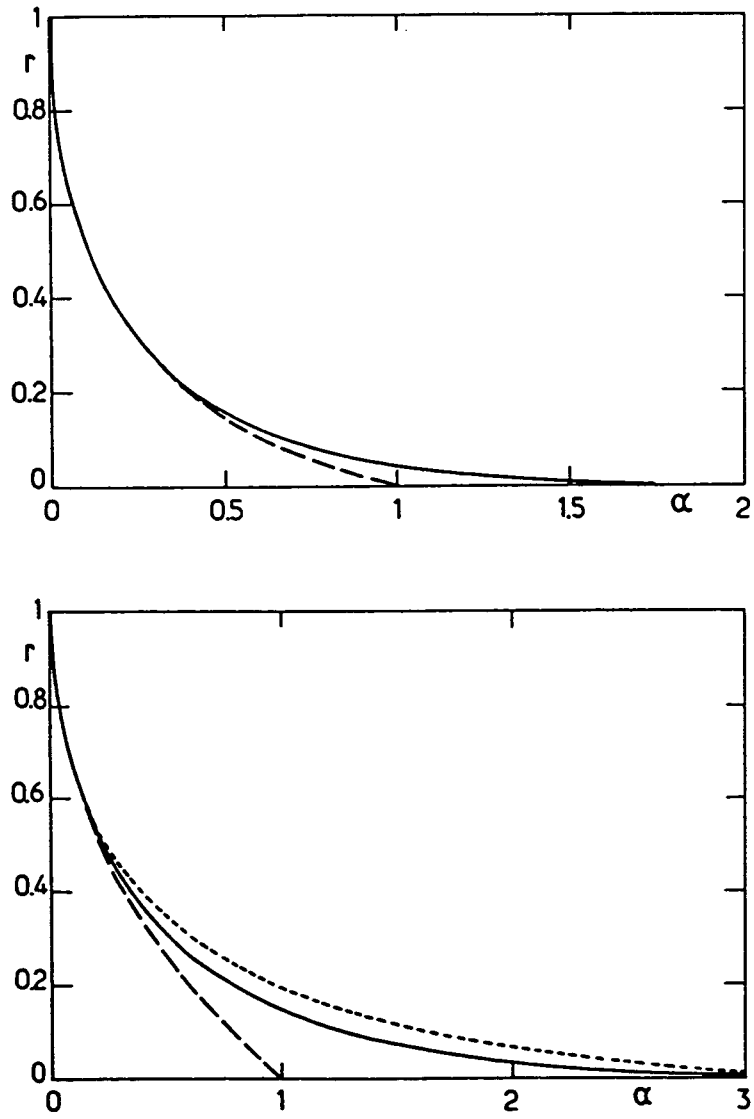


Figure 4: Radius of the basins of attraction r as a function of the number of stored patterns per input neuron, $\alpha = p/N$, in hetero-association. Dashed curve: P.I. algorithm; full curve: theoretical result which would be obtained in a network with fixed stabilities ($\mathcal{P}(\Delta) = \delta(\Delta - K_{\rm opt})$); dotted curve: result for the M.O. algorithm, taking into account the fluctuations of the stabilities above $K_{\rm opt}$. a) Unbiased input and output patterns ($m = m' = 0$). The full curve and dotted curve are essentially indistinguishable on this scale. b) Biased input ($m = 0.4$) and output ($m' = 0.8$) patterns.

where as before Q is the overlap matrix of the input patterns. As the input and output patterns are mutually uncorrelated, it will not be surprising to find that A self-averages to:

$$\bar{A} = \frac{m'^2}{N} \sum_{\mu \neq \nu} \overline{(Q^{-1})}_{\mu\nu} + \frac{1}{N} \sum_{\mu} \overline{(Q^{-1})}_{\mu\mu} \tag{4.12}$$

In order to prove this self averageness and to compute A we write

$$e^{\frac{\lambda^2}{2}A} = \frac{Z(\lambda)}{Z(0)} \tag{4.13}$$

where we have introduced the partition function

$$Z(\lambda) = \int \Pi_{\mu=1}^{P} \frac{dx_\mu}{\sqrt{2\pi}} e^{-\frac{1}{2}\sum_{\mu,\nu} x_\mu Q_{\mu\nu} x_\nu + \frac{\lambda}{\sqrt{N}} \sum_\mu x_\mu \epsilon_i'^\mu} \tag{4.14}$$

The calculation of $\overline{Z(\lambda)}$ and $\overline{(Z(\lambda)/Z(0))}$ can be done using standard techniques from the statistical physics of disordered systems. It is sketched in appendix A. One finds that A is self averaging and

$$\begin{aligned} \overline{(Q^{-1})}_{\mu\mu} &= \frac{\alpha}{(1-\alpha)(1-m^2)} \\ \mu \neq \nu : \overline{(Q^{-1})}_{\mu\nu} &= \begin{cases} 0 & \text{if } m = 0 \\ \frac{-\alpha}{(1-\alpha)(1-m^2)} & \text{if } m \neq 0 \end{cases} \end{aligned} \tag{4.15}$$

So, finally, the stability parameter for the P.I. rule is

$$\begin{aligned} \Delta_{\text{P.I.}} &= \sqrt{\tfrac{1-\alpha}{\alpha}} && \text{if } m = 0 \\ \Delta_{\text{P.I.}} &= \sqrt{\tfrac{1-\alpha}{\alpha}} \sqrt{\tfrac{1-m^2}{1-m'^2}} && \text{if } m \neq 0 \end{aligned} \tag{4.16}$$

(the crossover region which decides if m is close enough to zero is $m < 1/\sqrt{N}$). The P.I. rule realizes the case (3.5), where all stabilities are equal in the thermodynamic limit.

$\Delta_{\text{P.I.}}$ is plotted as a function of α in figure 3. Formulas (3.6) and (4.16) (or figures 1 and 3) allow one to obtain the radius of the basins of attraction for a given number of stored patterns. The result is plotted in figure 4.

4.3 Auto-association

The case of auto-association does not call for changes in the case of the M.O. algorithm, the optimal stability is given exactly by equation (4.6), with $m = m'$.

For the pseudoinverse rule one proceeds as in section 4.2. This requires now the calculation of

$$\sum_{j \neq i} J_{ij}^2 = B - B^2 \tag{4.17}$$

where

$$B = \frac{1}{N} \sum_{\mu,\nu} \xi_i^\mu (Q^{-1})_{\mu\nu} \xi_i^\nu \tag{4.18}$$

Using the same method as before (see appendix B), we find that $\sum_{j(\neq i)} J_{ij}^2$ is self averaging and that for any i it tends toward the limit $\alpha(1-\alpha)$ when $N \to \infty$. On the other hand the stability (3.2) (without the contribution of the diagonal coupling) is

$$\Delta_{\text{P.I.}} = \sqrt{\frac{1-\alpha}{\alpha}} \tag{4.19}$$

This formula coincides with (4.16) for $m' = m$. Therefore the dependence of Δ as a function of α can be read from figure 3.

It is interesting to notice that the M.O. algorithm (as any perceptron-type algorithm) uses the correlations between patterns to increase the storage capacity, while the pseudoinverse method in some way orthogonalizes the patterns, so that its capacity remains the same whatever the correlations between patterns.

5. Relevance to Hopfield-type models

The dynamics of the perceptron in auto-association can be considered as the evolution after one time-step of a Hopfield-type network for associative memory (using parallel updating). So the value of q' is the overlap on the pattern at time $t = 1$. Unfortunately, the reasoning which led to equation (3.3) cannot be reapplied to calculate q_{t+1} as function of $q_t (t \geq 1)$, because in general the noise of the configuration will no longer be Gaussian, the spins on various sites become correlated. Derrida et al. [23] have invented a special type of strongly diluted lattice, on which these correlations can be neglected, to which our results on auto-association with $J_{ii} = 0$ can be applied. Formula (3.3) gives, for parallel updating, the evolution of $q(t+1) = f(q(t))$:

$$q(t+1) = \int P(\Delta) d\Delta \int_0^{\Delta q(t)/\sqrt{1-q(t)^2}} dz \sqrt{\frac{2}{\pi}} e^{-z^2/2} \tag{5.1}$$

and the generalizations to thermal noise and asynchronous updating are straightforward. The radius of the basin of attraction is given by the unstable fixed point $q^* = f(q^*)$. We have therefore found that the dynamics at all times on this strongly diluted lattice is governed only by the stability, irrespective of the special learning rule used. The introduction of diagonal coupling ($J_{ii} \neq 0$) reintroduces correlations even on this lattice.

Nevertheless, some of the conclusions of the previous section can be used as hints for what happens in Hopfield's model. There the importance of the stability has been investigated in previous papers [16–18] even though the connection between stability and the basins of attraction there depends also on the correlations of the synaptic matrix (J_{ij}), especially on the symmetry of the matrix [16]. To test the role of diagonal couplings, we have performed numerical simulations on the fully connected Hopfield net with $100 \leq N \leq 400$

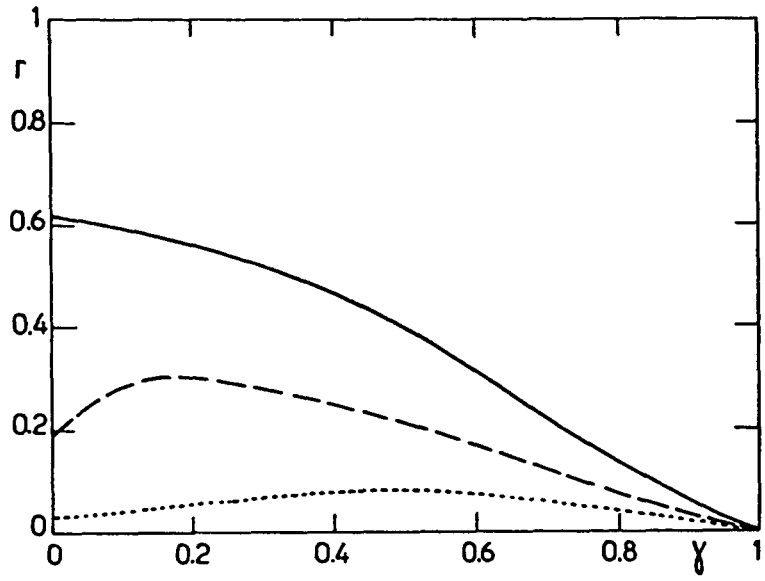


Figure 5: Radius of the basins of attraction r as a function of $\gamma = J_{ii}/\sum_{j\neq i} J_{ij}\xi_i^\mu\xi_j^\mu$, in a fully connected Hopfield model. The non-diagonal couplings are obtained by the P.I. algorithm. Full curve: $\alpha = 1/4$; dashed curve: $\alpha = 1/2$; dotted curve: $\alpha = 3/4$.

and $p = \alpha N (1/4 \leq \alpha \leq 3/4)$ unbiased random patterns. The non-diagonal couplings were chosen with the pseudo-inverse rule while the diagonal ones were all taken to be equal $J_{ii} = \gamma \sum_{j\neq i} J_{ij}\xi_i^\mu\xi_j^\mu$. The results (radius r as a function of γ) are presented in figure 5. They show clearly that varying the strength of the diagonal couplings has a strong effect on the convergence properties of the fully connected net (cf. Kanter and Sompolinsky [22]) and that the best choice of J_{ii} may not be $J_{ii} = 0$. We find e.g that choosing $\gamma \cong 0.15$ instead of $\gamma = 0$ can increase the radius of the basins of attraction by about 50%, for $\alpha = 1/2$.)

6. Using a perceptron to generalize: A simple case

The formulas we have found in section 4.2 show that, for the two algorithms, the size of the basins of attraction increases with the correlations in the output state, i.e. with the value of m', but *only* if the input correlation m is nonzero. Any perceptron-type algorithm can even achieve an infinite capacity in the limit $m' \to 1 (\lim_{m'\to 1} \alpha = \infty)$, provided $m \neq 0$. This is an interesting result which sheds some light on the perceptron's ability to generalize. If $m' \to 1$ there is an unique output state, whatever the input

state. Let us suppose, e. g., that the input correlation is $m > 0$. Then the memorization of p input patterns can be considered from a totally different point of view: the network must send all inputs with positive magnetization towards the unique output state which has been given to it (and by symmetry the input configuration with negative magnetization must be sent to the reversed output state). It is taught to learn this task through the presentation of p examples (the patterns). In this context the fact that the capacity is infinite simply means that the M.O. algorithm is able to learn this task [25]: in fact it is clear that for $p \gg N$ the set of coupling constants reached by the system is stable: each (magnetized) new pattern is automatically memorized, so that it does not lead to a change in the couplings.

This ability to generalize is a nice property. We have pointed it out here as a simple consequence of the results of the previous section, its detailed study is, however, beyond the scope of the present paper.

Let us however point out the relationship with the usual language of data analysis. The problem considered here is a simple case of classification: all patterns are distributed into two classes, depending on the sign of the magnetization. In classification tasks one usually looks for some distance criterium such that patterns belonging to a same class are grouped into a cluster of nearby elements, and patterns of different classes are as far apart as possible (see for example [26]). In the commonly used Discriminant Analysis method, one looks for axes (in state space) such that the projection on these axes optimally distributes the patterns into clusters. In the neural network language, the directions of the N' axes are given by the lines of the coupling matrix of a Perceptron with N' output units [27].

On one given axis i, that is for one given output unit i, the distances between the patterns are in fact directly given by the stabilities as defined in (3.2). Indeed, on axis i defined by the vector $\vec{J}_i = (J_{ij})_{j=1,N}$, the abscissa X_i^μ of the projection of a pattern μ is plus or minus its stability Δ_i^μ — depending on the class it belongs to: in the above example, $X_i^\mu = \Delta_i^\mu$ for a positive magnetization and $X_i^\mu = -\Delta_i^\mu$ for a negative magnetization. The choice of the Pseudo-Inverse rule corresponds to a typical choice in Discriminant Analysis method, which result in the minimization of the dispersion within classes [27]. Indeed, we have seen that all the patterns have the same stability $\Delta_{\text{P.I.}}$. On the axis, the two clusters are reduced to two points, distant of $2\,\Delta_{\text{P.I.}}$. The choice of the M.O. algorithm corresponds to the maximization of the distance between the two clusters: the distance between any two patterns belonging to different classes is at least equal to $2K_i$ (see (4.5)). But now the elements within a cluster are distributed according to the distribution (4.10).

7. Conclusion

We have calculated the storage capacity and the size of the basins of attraction for a perceptron-type network storing random pattern. In the case of hetero-association the important parameter which determines this size is

the stability which is maximized by the M.O. algorithm. When one considers auto-association, another parameter allows to improve the performance of the network: the diagonal couplings. This is also a useful parameter in Hopfield's network.

The dynamics which has been studied in this article is one-step (as the net is feed-forward and consists of two layers). It would be very useful to understand how this may be extended to the dynamics at several time steps, either in fully connected models, or in a multilayered feed forward architecture.

Acknowledgments

We acknowledge helpful discussions with L. F. Abbott, H. Gutfreund, O. Lefèvre, and M. Virasoro. This work has been partially supported by the European program "BRAIN", contract number ST2J-0422-C (EDB), and the numerical work has been performed in part on Sun workstations provided by DRET. WK acknowledges financial support by Studienstiftung des deutschen Volkes.

Appendix A.

We first calculate the average over the choices of random patterns of the partition function $Z(\lambda)$ defined in equation (4.14). Using the definition (4.4) of the matrix Q, we have

$$Z(\lambda) = \int \Pi_{\mu=1}^{P} \frac{dx_\mu}{\sqrt{2\pi}} \Pi_{j=1}^{N} \frac{dt_j}{\sqrt{2\pi}}$$

$$\exp\left(-\frac{1}{2}\sum_j t_j^2 + \frac{i}{\sqrt{N}} \sum_{j,\mu} t_j \xi_j^\mu x_\mu + \frac{\lambda}{\sqrt{N}} \sum_\mu \xi_i'^\mu x_\mu\right) \tag{A.1}$$

The average with respect to the input pattern is

$$\overline{Z(\lambda)} = \int \Pi_{\mu=1}^{P} \frac{dx_\mu}{\sqrt{2\pi}} \Pi_{j=1}^{N} \frac{dt_j}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}\sum_j t_j^2 + \frac{i}{\sqrt{N}} m \sum_j t_j \sum_\mu x_\mu\right.$$

$$\left. -\frac{1}{2N}(1-m^2)\sum_j t_j^2 \sum_\mu x_\mu^2 + \frac{\lambda}{\sqrt{N}} \sum_\mu \xi_i'^\mu x_\mu\right) \tag{A.2}$$

Writing $X = 1/N \sum_\mu x_\mu^2$ and enforcing this constraint through an auxiliary parameter $\hat{X}$, we have after integrating over the t_j:

$$\overline{Z(\lambda)} = \int dX \int_{-i\infty}^{i\infty} \frac{d\hat{X}}{2\pi} \int \Pi_{\mu=1}^{P} dx_\mu$$

$$\exp\left\{\frac{1}{2}\hat{X}(NX - \sum_\mu x_\mu^2) - \frac{N}{2}\log(1 + X(1-m^2))\right.$$

$$-\frac{m^2}{2}\frac{1}{1+X(1-m^2)}(\sum_\mu x_\mu)^2 + \frac{\lambda}{\sqrt{N}}\sum_\mu \xi^{\prime\mu}{}_i x_\mu\Big\} \tag{A.3}$$

The integral over the x_μ is Gaussian with a quadratic form

$$e^{-\frac{1}{2}\sum_{\mu,\nu} x_\mu M_{\mu\nu} x_\nu} \tag{A.4}$$

where

$$M_{\mu,\nu} = \hat{X}\delta_{\mu,\nu} + \frac{m^2}{1+X(1-m^2)} \tag{A.5}$$

Performing the integrations over x_μ' s we find

$$\overline{Z(\lambda)} = \int dX \int_{-i\infty}^{i\infty} \frac{d\hat{X}}{2\pi} e^{N\left\{\frac{X\hat{X}}{2} - \frac{1}{2N}\log\det M - \frac{1}{2}\log(1+X(1-m^2))\right\}}$$

$$\times\, e^{\frac{\lambda^2}{2N}\sum_{\mu,\nu}\xi_i^{\prime\mu}(M^{-1})_{\mu\nu}\xi_i^{\prime\nu}} \tag{A.6}$$

The determinant and the inverse of M are

$$\det M = \hat{X}^{p-1}\left[\hat{X} + p\frac{m^2}{1+X(1-m^2)}\right] \tag{A.7}$$

$$(M^{-1})_{\mu\nu} = \frac{1}{\hat{X}}\left[\delta_{\mu\nu} - \frac{m^2}{pm^2 + \hat{X}(1+X(1-m^2))}\right] \tag{A.8}$$

For p and N going to infinity with fixed capacity $\alpha = p/N$, one can perform the integrals over X and $\hat{X}$ by saddle point. It is easy to see from equation (A.8) that the last term in equation (A.6) does not contribute to the saddle point. The solution to the saddle point equations are

$$X = \frac{\alpha}{\hat{X}} = \frac{\alpha}{(1-\alpha)(1-m^2)} \tag{A.9}$$

so that finally

$$\begin{aligned} \overline{Z(\lambda)} &= \overline{Z(0)}e^{\frac{\lambda^2}{2}\frac{\alpha}{1-\alpha}\frac{1-m^2}{1-m^2}} \quad \text{if } m \neq 0 \\ \overline{Z(\lambda)} &= \overline{Z(0)}e^{\frac{\lambda^2}{2}\frac{\alpha}{1-\alpha}} \qquad\quad \text{if } m = 0 \end{aligned} \tag{A.10}$$

This is not yet enough to prove the announced result since we have computed so far $\overline{Z(\lambda)}/\overline{Z(0)}$ instead of $\overline{Z(\lambda)/Z(0)}$. In order to compute this last quantity we could use the replica method (see e.g. Mézard et al. [10])

$$\overline{\left(\frac{Z(\lambda)}{Z(0)}\right)} = \lim_{n\to 0} \overline{Z(\lambda)Z(0)^{n-1}} \tag{A.11}$$

$Z(\lambda)Z(0)^{n-1}$ being calculated by the introduction of n copies of the variables x_μ.

However, in the present case, this is not even necessary, because $Z(\lambda)$ turns out to be self averaging: we can directly calculate $\overline{Z^2(\lambda)}$ using the same techniques as before. We introduce two types of x-variables x_μ and x'_μ and write

$$X = 1/N \sum_\mu x_\mu^2, X' = 1/N \sum_\mu x'^2{}_\mu, Q = 1/N \sum_\mu x_\mu x'_\mu,$$

together with the auxiliary parameters $\hat{X}$, $\hat{X}'$, $\hat{Q}$. Then

$$\overline{Z^2(\lambda)} = \int dX \frac{d\hat{X}}{2\pi} dX' \frac{d\hat{X}'}{2\pi} dQ \frac{d\hat{Q}}{2\pi} \int \Pi_{\mu=1}^P \frac{dx_\mu}{\sqrt{2\pi}} \frac{dx'_\mu}{\sqrt{2\pi}} \tag{A.12}$$

$$\times \exp\left\{\frac{\hat{X}}{2}(NX - \sum_\mu x_\mu^2) + \frac{\hat{X}'}{2}(NX' - \sum_\mu x'^2_\mu) + \hat{Q}(NQ - \sum_\mu x_\mu x'_\mu)\right\} \tag{A.13}$$

$$\times \exp\left\{-\frac{N}{2}\log D - \frac{m^2}{2}\frac{1 + X'(1-m^2)}{D}(\sum_\mu x_\mu)^2 - \frac{m^2}{2}\frac{1 + X(1-m^2)}{D}(\sum_\mu x'_\mu)^2\right\} \tag{A.14}$$

$$\times \exp\left\{m^2 Q \frac{1-m^2}{D}(\sum_\mu x_\mu)(\sum_\mu x'_\mu) + \frac{\lambda}{\sqrt{N}} \sum_\mu \xi_i'^\mu (x_\mu + x'_\mu)\right\}$$

where D is a notation for the determinant

$$D = \{1 + X(1-m^2)\}\{1 + X'(1-m^2)\} - Q^2(1-m^2)^2 \tag{A.15}$$

The Gaussian integrals over x_μ and x'_μ lead as before to a determinant of the corresponding quadratic form. Let us begin with $\lambda = 0$. We obtain

$$\overline{Z^2(0)} = \int dX \frac{d\hat{X}}{2\pi} dX \frac{d\hat{X}'}{2\pi} dQ \frac{d\hat{Q}}{2\pi}$$

$$\times \exp \frac{N}{2}\left\{X\hat{X} + X'\hat{X}' + 2Q\hat{Q} - \log D - \alpha \log(\hat{X}\hat{X}' - \hat{Q}^2)\right\} \tag{A.16}$$

A careful examination of the saddle point equations of this integral shows that the dominant saddle point is always at

$$Q = \hat{Q} = 0 \tag{A.17}$$

$$X = \frac{\alpha}{\hat{X}} = \frac{\alpha}{(1-\alpha)(1-m^2)}; \qquad X' = \frac{\alpha}{\hat{X}'} = \frac{\alpha}{(1-\alpha)(1-m^2)}$$

so that

$$\overline{Z^2(0)} \overset{N\to\infty}{\sim} \overline{Z(0)}^2 \tag{A.18}$$

As the term in λ in equation (A.12) is not of the leading order in N, it cannot change the saddle point and one finds

$$\overline{Z^2(\lambda)} \overset{N\to\infty}{\sim} \overline{Z(\lambda)}^2 \tag{A.19}$$

This shows that, for any λ, the fluctuations of $Z(\lambda)$ can be neglected; therefore we obtain from equation (A.10)

$$\overline{\left(\frac{Z(\lambda)}{Z(0)}\right)} = \overline{\left(e^{\frac{\lambda^2}{2}A}\right)} \overset{N\to\infty}{\sim} \begin{cases} e^{\frac{\lambda^2}{2}\frac{\alpha}{1-\alpha}\frac{1-m'^2}{1-m^2}} & \text{if } m' \neq 0 \\ e^{\frac{\lambda^2}{2}\frac{\alpha}{1-\alpha}} & \text{if } m' = 0 \end{cases} \tag{A.20}$$

which is the announced result.

Appendix B.

We calculate the stability parameters for the P.I. rule in the case of auto-association. From equation (4.18) we need to calculate

$$B = \frac{1}{N}\sum_{\mu,\nu=1}^{P} \xi_i^\mu \xi_i^\nu (Q^{-1})_{\mu\nu} \tag{B.1}$$

As in appendix A we write $e^{\frac{\lambda^2}{2}B} = \overline{\left(\frac{Z(\lambda)}{Z(0)}\right)}$, with

$$\begin{aligned} Z(\lambda) &= \int \Pi_{\mu=1}^{P} \frac{dx_\mu}{\sqrt{2\pi}} \Pi_{j=1}^{N} \frac{dt_j}{\sqrt{2\pi}} \exp(-\frac{1}{2}\sum_j t_j^2) \times \\ &\times \left[\exp\left(\frac{i}{\sqrt{N}}\sum_{j,\mu} t_j \xi_j^\mu x_\mu + \frac{\lambda}{\sqrt{N}}\sum_\mu \xi_i^\mu x_\mu\right)\right] \end{aligned} \tag{B.2}$$

We proceed as in appendix A: we first average over the ξ_j^μ $(j \neq i)$, then integrate over the t_j $(j \neq i)$ and finally integrate over the x_μ, with $X = 1/N \sum_\mu x_\mu^2$, fixed by an auxiliary parameter $\hat{X}$. The result is

$$\begin{aligned} Z(\lambda) &= \int dX d\hat{X} \exp N\left(\frac{\hat{X}X}{2} - \frac{\alpha}{2}Ln\hat{X} - \frac{1}{2}Ln[1 + X(1-m^2)]\right) \times \\ &\times \left[\int \frac{dt_i}{\sqrt{2\pi}} \exp\left(-\frac{t_i^2}{2} + \frac{1}{2N}\sum_{\mu,\nu}(M^{-1})_{\mu\nu}(\lambda + it_i)^2 \xi_i^\mu \xi_i^\nu\right]\right. \end{aligned} \tag{B.3}$$

where M and M^{-1} are given in equations (A.5–8). As before it is easily seen that the last term does not affect the saddle point (A.9) on X and $\hat{X}$, so that

$$\frac{\overline{Z(\lambda)}}{\overline{Z(0)}} = \frac{\int \frac{dt}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} e^{\frac{1}{2}\frac{\alpha}{1-\alpha}(\lambda+it)^2}}{\int \frac{dt}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} e^{\frac{1}{2}\frac{\alpha}{1-\alpha}(it)^2}} = e^{\frac{\lambda^2}{2}\alpha} \tag{B.4}$$

The self averageness of $Z(\lambda)$ derived in appendix A also applies here, so that

$$\overline{\left(\frac{Z(\lambda)}{Z(0)}\right)} \underset{\sim}{N\to\infty} e^{\frac{\lambda^2}{2}\alpha} \tag{B.5}$$

Hence B self averages to

$$\lim_{N\to\infty} B = \alpha \tag{B.6}$$

and

$$\sum_{j(\neq i)} J_{ij}^2 = \alpha(1-\alpha) \tag{B.7}$$

References

[1] J. J. Hopfield, "Neural networks and physical systems with emergent collective computational abilities", *Proc Nat Acad Sci USA* , **79** (1982) 2554.

[2] D. J. Amit, H. Gutfreund, and H. Sompolinsky, "Storing infinite number of patterns in a spin-glass model of neural network", *Phys Rev Lett*, **55** (1985) 1530.

[3] A. Crisanti, D. J. Amit, and H. Gutfreund, "Saturation level of the Hopfield model for neural network", *Europhys Lett*, **2** (1986) 337.

[4] E. Gardner, N. Stroud, and D. J. Wallace, "Training with noise and the storage of correlated patterns in a neural network model", Preprint Edinburgh 87/394 (1987).

[5] S. Diederich and M. Opper, "Learning of correlated patterns in spin-glass networks by local learning rules", *Phys Rev Lett*, **58** (1987) 949.

[6] W. Krauth and M. Mézard, "Learning algorithms with optimal stability in neural networks", *J Phys A: Math Gen*, **20** (1987) L745.

[7] S. Venkatesh, in *Neural Networks for Computing, AIP Conference Proceedings*, **151**, ed. J. S. Denker (Am. Inst. Phys., New York, 1986) 440.

[8] E. Gardner, "The space of interactions in neural networks models", *J. Phys. A*, **21** (1988) 257; "Maximum storage capacity in neural networks", *Europhys Lett*, **4** (1987) 481.

[9] D. J. Amit, "The properties of models of simple neurals networks", in *Heidelberg Colloquium on Glassy Dynamics*, eds. J. L. van Hemmen and I. Morgenstern (Springer, Berlin, 1987).

[10] M. Mézard, G. Parisi, and M. A. Virasoro, *Spin Glass Theory and Beyond* (World Scientific, Singapore, 1987).

[11] D. E. Rumelhart and J. L. McClelland, *Parallel Distributed Processing*, vol.1 and 2 (Bradford Books, Cambridge MA, 1986).

[12] F. Rosenblatt, *Principles of Neurodynamics* (Spartan Books, New York, 1962).

[13] M. Minsky and S. Papert, *Perceptrons* (MIT Press, Cambridge, MA, 1969).

[14] F. Fogelman Soulié, P. Gallinari, Y. Le Cun, and S. Thiria, "Automata networks and artificial intelligence" in *Automata Networks in Computer Science*, eds. F. Fogelman Soulié, Y. Robert, and M. Tchuente (Manchester Univ. Press, 1987) 133.

[15] T. Kohonen, *Self Organization and Associative Memory* (Springer, Berlin, 1984).

[16] W. Krauth, J. P. Nadal, and M. Mézard, "The roles of stability and symmetry on the dynamics of neural networks", to appear in *J Phys A: Math Gen*, **21** (1988) 2995.

[17] B. M. Forrest, "Content addressability and learning in neural networks", *J. Phys. A: Math Gen.*, **21** (1988) 245.

[18] T. B. Kepler and L. F. Abbott, "Domains of attraction in neural networks", Brandeis University preprint (1988).

[19] N. Parga and M. A. Virasoro, "The ultrametric organization of memories in a neural network", *J. Physique* (Paris) **47** (1986) 1857.

[20] D. J. Amit, H. Gutfreund, and H. Sompolinsky, "Neural networks with correlated patterns: towards pattern recognition", *Phys. Rev.*, **A35** (1987) 2293; H. Gutfreund, "Neural Networks with Hierarchically Correlated Patterns", *Phys. Rev. A*, **37** (1988) 570.

[21] M. V. Feigelman and L. B. Ioffe, "The augmented models of associative memory, asymmetric interaction and hierarchy of patterns", *Int. J. of Mod. Phys.*, **B1** (1987) 51.

[22] I. Kanter and H. Sompolinsky, "Associative recall of memory without errors", *Phys. Rev.*, **A35** (1987) 380.

[23] B. Derrida, E. Gardner, and A. Zippelius, "An exactly soluble asymmetric neural network model", *Europhys Lett*, **4** (1987) 167.

[24] L. Personnaz, I. Guyon, and G. Dreyfus, "Information storage and retrieval in spin-glass like neural networks", *J de Physique*, **L16** (1985) 359.

[25] P. Delgiudice, S. Franz, and M. A. Virasoro, Preprint 605, Rome University.

[26] R. O. Duda and P. E. Hart, *Pattern Classification and Scene Analysis* (Wiley, 1973).

[27] P. Gallinari, S. Thiria, and F. Fogelman Soulié, "Multilayer Perceptrons and Data Analysis," to appear in ICNN 88, IEEE Annual International Conference on Neural Networks (San Diego, CA, July 24-27, 1988).

1

Critical Storage Capacity of the J = ±1 Neural Network

Werner Krauth[1] and Manfred Opper[2,†]

[1]Laboratoire de Physique Statistique de l'Ecole Normale Supérieure
24, rue Lhomond; 75231 Paris Cedex 05; France
[2]Institut für theoretische Physik der Justus Liebig Universität
Heinrich Buff Ring 16, D-6300 Giessen; FRG

Abstract

For neural networks in which the couplings J_{ij} are allowed to take on the values $J_{ij} = 1$ or $J_{ij} = -1$, we determine numerically the critical storage capacity for random unbiased patterns as a function of the stability. We use an exact enumeration scheme based on the Gray code and a continuous distribution for the patterns to control finite size effects. Results are presented for $N \leq 25$; they indicate an optimal storage capacity of $\alpha_c \cong 0.82$ $(N \to \infty)$.

PACS. 05.20. - Statistical mechanics.

† presently at Laboratoire de Physique Théorique de l'Ecole Normale Supérieure, Unité propre du CNRS, associé à l' Ecole Normale Supérieure et à l'Université de Paris-Sud.

2

An important result in the 'modern' theory of neural networks is due to Gardner [Gardner 1987, 1988]. By means of a replica calculation she determined the typical volume of real-valued interactions J_{ij} ($i,j=1,...,N$; $(N\to\infty)$) with $\Sigma_j J_{ij}^2=N$ which solve

$$\xi_i^\mu \Sigma_j J_{ij} \xi_j^\mu \geq \kappa \sqrt{N};\ i=1,...,N;\ \mu = 1,...,p \qquad (1)$$

In eq. (1), the ξ_i^μ are fixed independent random patterns, which take on the values $\xi_i^\mu=+1$ and $\xi_i^\mu=-1$ with probability $(1+m)/2$ and $(1-m)/2$, repectively (m is called bias).

This problem is in some respect the inverse of the mean field theory of spin glasses, where, e.g., the properties of a Hamiltonian

$$H = -\Sigma_{i,j} J_{ij} S_i S_j \qquad (2)$$

are considered. There, (J_{ij}) is a fixed (symmetric) random matrix, and the spins are allowed to vary, i.e. the roles of couplings and spins are interchanged. The spherical condition $\Sigma_j J_{ij}^2 = N$ for the neural network then corresponds to the spherical model of spin glasses ($\Sigma_i S_i^2=N$) (Kosterlitz et al 1976).

The calculation of E. Gardner marked a significant progress in theory because it showed that the replica method can be used in the phase space of couplings J_{ij} and, more generally, for systems in which these couplings are not necessarily symmetric ($J_{ij}\neq J_{ji}$). In addition, her results were of much practical importance as they settled the problem of the optimal storage capacity in Hopfield-type neural networks under the zero temperature dynamics

$$s_i(t+1) = \mathrm{sgn}\ \Sigma_j J_{ij} s_j(t) \quad i=1,...,N \qquad (3)$$

The stability condition for all patterns $\vec{\xi}^\mu$ $(=(\xi_1^\mu, \xi_2^\mu,..., \xi_N^\mu))$ under the dynamics eq. (3) is $\xi_i^\mu = \mathrm{sgn}(\Sigma_j J_{ij} \xi_j^\mu)$ $i=1,...,N$; $\mu=1,...,p$, this is equivalent to eq. (1) with stability $\kappa=0$. The optimal storage capacity $\alpha_c=\alpha_c(\kappa=0)$ appears in Gardner's elegant formulation as the value of $\alpha=p/N$ at which the typical fractional volume of the sphere $\Sigma_j J_{ij}^2=N$ which solves eq. (1) goes to zero, at $\kappa=0$.

3

The results are by now well known: $\alpha_c = 2$, for unbiased patterns (cf. Cover 1965, Venkatesh 1986); for strongly biased patterns the result is $\alpha_c \sim 1/((m-1)\log(1-m))$ $(m \to 1)$.

Gardner's calculation for this 'spherical' model $\Sigma_j J_{ij}^2 = N$ is unambiguous: replica symmetry has to be assumed but can be shown to remain intact. Thus, the solution she proposed is at least locally stable. The predictions for $\alpha_c(\kappa)$ have been confirmed numerically with an optimal stability algorithm (Krauth and Mézard 1987).

In this letter we will be concerned with a system in which the coupling matrix consists not of arbitrary real variables, but of Ising-type variables $J_{ij} = +1$ or $J_{ij} = -1$. This system may be thought to model more faithfully practical neural networks in which the precision of the J_{ij} is fixed. Conceptually, the $J = \pm 1$ model offers some advantages over the spherical model, especially since the information content of the coupling matrix can be determined. It follows from information theory that the capacity of this model should be smaller than or equal to one [Gardner and Derrida 1988].

This $J = \pm 1$ model of neural networks is related to the Sherrington-Kirkpatrick model of spin glasses in the same way as the $\Sigma_j J_{ij}^2 = N$ model to the spherical one of Kosterlitz et al. This suggests that the two network models might have quite different properties as well.

There has been some study of this model before. A replica symmetric calculation is still possible [Gardner and Derrida 1988] (it gives $\alpha_c = 4/\pi$), but the replica symmetry is broken above a deAlmeida-Thouless (AT) line, which passes through $\alpha \cong 1.03$ for $\kappa = 0$. Thus, the critical capacity is basically not known.

The simulation of the $J_{ij} = \pm 1$ neural network is more complicated also. (See [Amaldi and Nicolis, 1988, Gardner and Derrida, 1988], who obtained rough estimates of α_c.) Since there is no perceptron-type learning algorithm (guaranteed to converge if there exists a solution), not to speak of a well-behaved optimal stability algorithm, we resort to an exact enumeration method. We are thus avoiding the drawbacks of Monte Carlo simulations. To cope with the equally annoying problem of small system sizes, we use here a

sophisticated algorithm, which allows us to reach rather large sizes and a continuous probability distribution for the patterns, to reduce finite size effects.

As in the work of Gardner we consider just one row of the matrix (J_{ij}) (with fixed index i). For one sample of patterns we define the variable $\kappa_J = \min_\mu\{\Sigma J_{ij}\eta_j^\mu/\sqrt{N}\}$, and the optimal stability as $\kappa_{opt}=\max_J\{\kappa_J\}$. The critical storage capacity $\alpha_c(\kappa)$ is then found as the inverse of $\kappa_{opt}(\alpha)$. Calculating the optimal stability involves, for one given sample of random patterns $\eta^\mu=\xi_i^\mu\,\xi^\mu$, to check all the 2^N possible vectors $J=(J_{i1}, J_{i2},\ldots,J_{iN})$. This is less of a brute force method than it may seem: there is room for algorithmic subtleties, and the use of them in combination with a powerful vector computer allows us to consider (with good statistics) systems up to the size N=25.

After generation of a set of random patterns (we restrict ourselves to unbiased patterns with m=0), the calculation of the optimal stability is not complicated. One just determines for each of the 2^N possible vectors the variable κ_J. Since we are free to choose the order in which the possible vectors J are scanned, we use the minimal change order provided by the Gray code [Nievergelt et al 1977]. In the Gray code one vector J is derived from the previous one by flipping just one of the J_{ij}'s. This simplifies enormously the calculation of the stabilities $\Delta^\mu=\Sigma_j J_{ij}\eta_j^\mu$ at each step, compared to an order of the J's analogous to counting in the binary system.

The resulting algorithm is amazingly fast. For our largest systems with N=25 it enumerates all possible couplings in 400 s. Of this time it spends about 1/7 on counting (Gray code) and 6/7 on other computing (calculating stabilities, taking minima and maxima, bookkeeping). This excellent ratio is due to the fact that the numerical part can be performed in parallel.

As we are interested in the optimal stability κ_{opt} for large systems it may seem natural to calculate $\max_J\{\min_\mu\Sigma J_{ij}\,\eta_j^\mu\}/\sqrt{N}$, to average over many samples and to try one's luck with an extrapolation to $N\to\infty$. This has in fact been done, but without success. The fact that the possible values of the stability are discrete with a spacing of $2/\sqrt{N}$ for binary valued patterns seems to

preclude this approach. In fact, the resulting curves are rather erratic, and there is a large parity effect (see fig. 1).

It is for this reason that we add one more trick: Instead of restricting the values of the η_j^μ's to binary values ± 1 we use random variables η_j^μ with a continuous distribution. For the present model it becomes clear from the theoretical treatment that the statistical properties depend only on the first two moments of the distribution for $N \to \infty$ (cf. the situation in the Sherrington-Kirkpatrick model, in which a Gaussian disorder was used for convenience only).

The model with continuous η's has the advantage to yield continuous values for the stabilities as well. As we will see, the limit $N \to \infty$ is much smoother and an extrapolation becomes possible. For concreteness we choose the normalized Gaussian distribution

$$p(\eta_j^\mu) = 1/\sqrt{2\pi}\ \exp(-1/2\ \eta_j^{\mu\,2}) \qquad (4)$$

The mean values of the optimal stabilities $\kappa_{opt}(N, \alpha)$ are found by averaging over a large number of samples (e.g. 10000 at N=10, 500 samples at N=25) for values of α ranging from 0.25 to 2.0. The results for α=0.75, 0.8, 0.857, 1. are given in fig.1. An extrapolation in 1/N seems to pose no problem, and leads to rather precise predictions for the value of $\alpha_c(\kappa)$ in the thermodynamic limit $N \to \infty$. To show what we have gained by using continuous patterns, we display for comparison the results of earlier simulations with patterns $\xi_i^\mu = \pm 1$ for α=1. (We stress again that in the limit $N \to \infty$ both distributions of patterns must yield the same $\alpha_c(\kappa)$).

The extrapolated values of $\alpha_c(\kappa)$ are given in fig.2. We use the opportunity to display our numerical results together with the results for $\alpha_c(\kappa)$ in the replica symmetric approximation and for the AT-line. The critical capacity for the spherical model is included also.

The evolving picture of our numerical investigation is the following: the critical storage ration α_c of the $J=\pm 1$ neural network seems to be close to 0.82. For all values of κ we find results for $\alpha_c(\kappa)$ which are below the AT-line and, as it should be, below the critical capacity of the spherical model. It is needless to caution that these statements (for $N \to \infty$) critically depend on the validity of our

finite-size scaling hypothesis, which should be checked at larger values of N and which should be given a theoretical foundation. Certainly this model merits further theoretical study.

We would expect that the efficient enumeration provided by the Gray code may be used in different circumstances, as for exhaustive studies of spin glasses.

Acknowledgments

We would like to thank J. Anlauf, M. Mézard, and J.-P. Nadal for helpful discussions. The calculations were performed on a Convex C1 vector computer supported by GRECO 70 'Expérimentation Numérique'. We acknowledge financial support from Studienstiftung des deutschen Volkes (W. K.) and from Deutsche Forschungsgemeinschaft (M. O.).

7

Figure captions

Fig.1: Values of $\kappa_{opt}(N, \alpha)$ vs 1/N for (from above) α=3/4, 4/5, 6/7, and 1, and extrapolation to $N\to\infty$. For comparison we display the corresponding values from earlier simulations with $\xi_i^\mu=\pm1$ for α=1 ([]).

Fig.2: Critical lines for Hopfield-type models. The points on the lowest curve give the values of $\alpha_c(\kappa)$ as determined numerically for α=1/2, 4/5, 6/7, 1, 3/2, 2 (the curve is a guideline to the eye). The other curves give, from above, the critical storage capacity $\alpha_c(\kappa)$ for the 'spherical' model ($\Sigma_j J_{ij}^2=N$), the storage capacity for the $J=\pm1$ model in the replica-symmetric approximation, and the deAlmeida-Thouless line, above which replica symmetry is broken.

8

References

Amaldi E and Nicolis S 1988, Rome University preprint # 642

Cover T M 1965, IEEE transactions **EC14** 3, 326

Gardner E 1987, Europhys. Lett **4**, 481

Gardner E 1988, J. Phys. A: Math. Gen. **21**, 257

Gardner E and Derrida B 1988, Saclay preprint SPhT/88-198

Kosterlitz J M, Thouless D J, and Jones R C 1976; Phys Rev Lett **36**, 1217

Krauth W and Mézard M 1987, J. Phys. A: Math. Gen. **20** L745

Reingold E M, Nievergelt J, and Deo N 1977, Combinatorial Algorithms: Theory and Practice (Englewood Cliffs, NJ: Prentice Hall)

Sherrington D and Kirkpatrick S 1975; Phys Rev Lett **36**, 1217

Venkatesh S 1986, Proc. Conf. on Neural Networks for Computing, Snowbird, Utah (AIP Conf. Proc. **151**) ed Denker J S

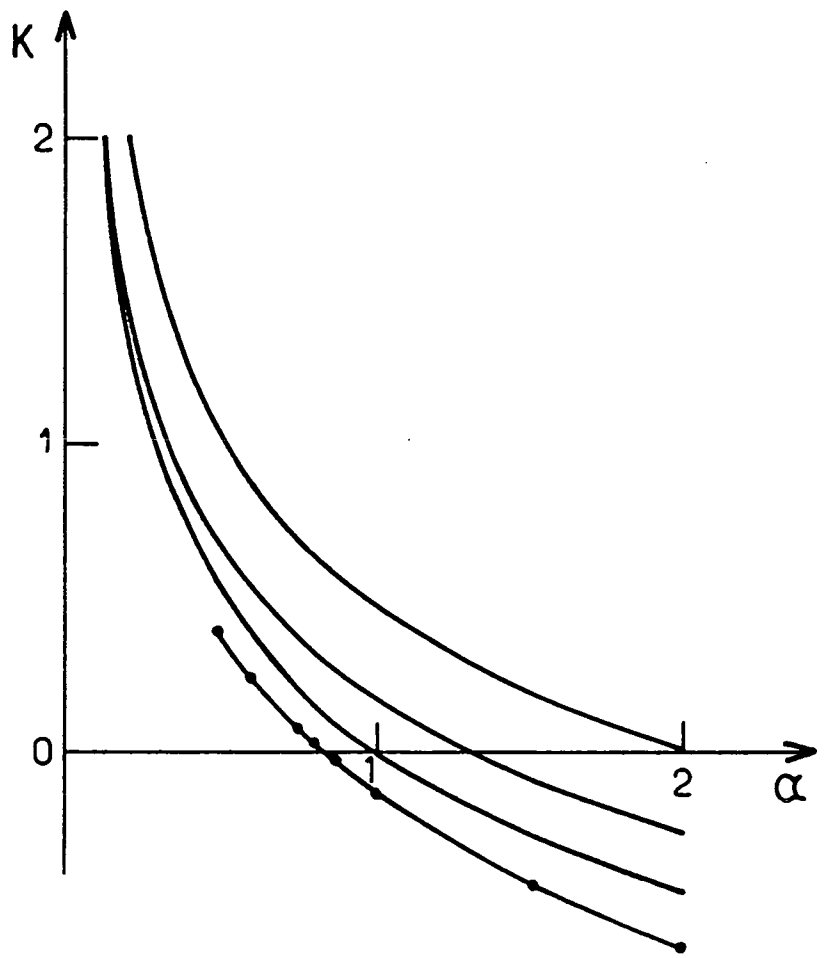


fig 2

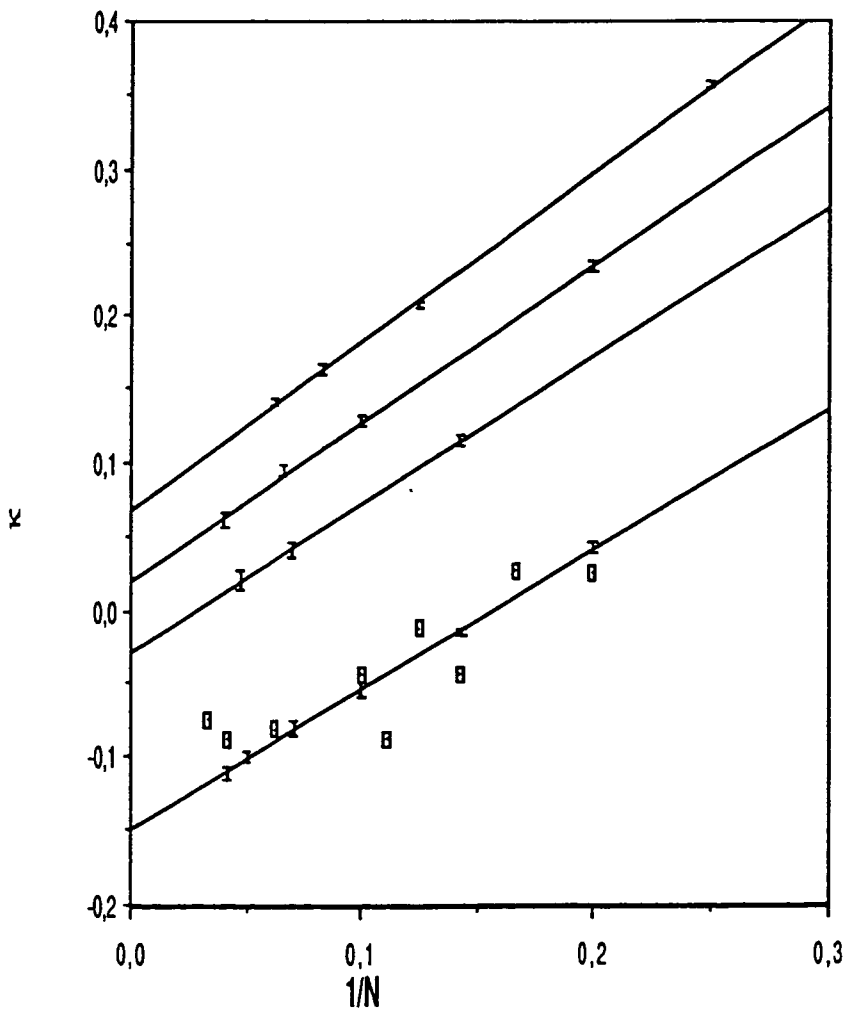


Fig 1

Storage capacity of memory networks with binary couplings

Werner Krauth and Marc Mézard (*)

Laboratoire de Physique Statistique de l'Ecole Normale Supérieure†

24, rue Lhomond, 75231 Paris Cedex 05, France

Abstract ;

We study the number p of unbiased random patterns which can be stored in a neural network of N neurons used as an associative memory, in the case where the synaptic efficacies are constrained to take the values ± 1. We find a solution with one step of replica symmetry breaking à la Parisi. This solution gives a critical capacity $\alpha_c = p/N \sim 0.83$ which seems to agree with known numerical results.

LPS preprint 95

† Laboratoire associé au CNRS et à l'Université Pierre et Marie Curie (URA 1306)

(*) Laboratoire de Physique Théorique de l'ENS.

1. INTRODUCTION

Consider a network of N binary neurons $\sigma_i = \pm 1$, coupled through synapses J_{ij}, and evolving in time according to the zero temperature Monte-Carlo dynamics (sequential or parallel) :

$$\sigma_i^{t+1} = Sign\left(\sum_j J_{ij}\, \sigma_j^t \right) \qquad (1)$$

The network is used as an associative memory to store p patterns which are N bit words, $\xi^\mu_i = \pm 1$, $i = 1,..., N$; $\mu = 1,...,p$. An important question which immediately arises is what is the capacity of such a network ? This is usually quantified by the number of random patterns which can be memorised in the limit $N \to \infty$. A lot of work has already been devoted to this capacity problem, firstly in the case of some specific learning rules. For instance with Hebb's rule $J_{ij}=(1/N)\sum_\mu \xi_i^\mu \xi_j^\mu$ the asymptotic capacity is $\alpha_c = p/N \sim .14$ [1,2].

A major breakthrough in this problem was achieved by Elizabeth Gardner who showed how to calculate the optimal capacity α, that is the number of patterns per neuron which can be stored by the best learning rule [3]. Gardner's calculation imposed no constraint on the couplings J_{ij} and gave an optimal capacity for unbiased random patterns of $\alpha_c = 2$.This result confirms an old analysis of Cover [4] and allows to extend it to more complicated distributions of patterns with bias or correlations [3,5].

Another direction in which a similar result is clearly needed is the case where the synapses J_{ij} are constrained. The extreme case where J_{ij} can take only two values $J_{ij} = \pm 1$ is of particular interest for at least two reasons :

- first of all one can count the number of bits used in the synapses, N^2, and compare it to the number of stored bits which is αN^2 for unbiased random patterns,

- secondly, it is clear that for practical applications one does not want to require an infinite precision for the J_{ij}'s but one wants to code them on a small number of bits. The present case is just the extreme one of one bit per synapse [F1]. For the specific case of the clipped Hebb's rule $J_{ij} = \text{Sign}\,(\Sigma_\mu \xi_i^\mu \xi_j^\mu)$ Sompolinsky [6] found a capacity of $\alpha \sim .1$.

In this paper we consider the problem of the learning capacity for unbiased random patterns, with J_{ij} constrained to the values ± 1, independently of any specific choice for the learning rule. This problem has already been considered by Gardner and Derrida [5] who found a replica symmetric solution giving $\alpha_c = 4/\pi$ but showed that replica symmetry must be broken. The solution we give here uses one step of replica symmetry breaking in Parisi's hierarchical scheme and predicts $\alpha_c = .83$. Thus the effects of replica symmetry breaking are drastic. This is due to a first order phase transition.

2. REPLICA FORMULATION

The starting point of the computation follows previous approaches [5]. In a N+1 neuron system we consider the N couplings $J_{0i} = J_i$, i = 1, ..., N, incoming onto a given neuron i = 0. For a pattern μ to be learnt one needs that it be a fixed point of the dynamics :

$$\xi_0^\mu = \text{Sign}\left(\sum_{k=1}^{N} J_k \xi_k^\mu\right) \qquad (2)$$

which we write as $\Sigma_k J_k \eta_k^\mu > 0$, where $\eta_k^\mu = \xi_0^\mu \xi_k^\mu$ are independent random variables, taking values ± 1 with probability 1/2. It is useful to generalize this condition to $\Sigma_k J_k \eta_k^\mu > \kappa\sqrt{N}$ where κ is a stability parameter. When κ is positive any configuration which differs from ξ^μ in less than $\kappa\sqrt{N}/2$ bits will flow towards it.

In fact it has been shown that the radius of the attraction basin is much larger (of order N) and increases with κ [7,8].

For each set of couplings we define the energy as the number of patterns which are not memorized :

$$E[J,\eta] = \sum_{\mu=1}^{P} \Theta\left(\kappa - \frac{1}{\sqrt{N}} \sum_{i=1}^{N} J_i \eta_i^{\mu}\right) \qquad (3)$$

Following Gardner and Derrida [5] we introduce, for a given set of patterns η_k^{μ}, the partition function at temperature $T = 1/\beta$:

$$Z = \sum_{\{J_i\}} e^{-\beta E[J,\eta]} \qquad (4)$$

The physical properties of the problem defined by this partition function nicely map onto the desired properties of the learning problem in the zero temperature limit:

- the internal energy at zero temperature is the number of ill-memorized patterns (those with $\Sigma_k J_k \eta_k^{\mu} < \kappa\sqrt{N}$) for the optimal configuration of J_i 's.

- the patterns can be learnt if and only if the zero temperature internal energy vanishes; in this case the zero temperature entropy is the logarithm of the number of coupling configurations which can learn all the patterns with a stability larger than κ.

As we shall see, although we are mostly interested in the zero temperature limit, it is extremely useful to consider the problem also at finite temperatures.

As we want to compute extensive thermodynamic quantities, we expect that these will be self-averaging, so that the properties of one typical sample will be obtained in the thermodynamic limit by the quenched average of Log. Z. This is computed with the replica method [9] which gives in the present case [5] (denoting by a bar the average over the distribution of patterns) :

$$\overline{Z^m} = \prod_{1 \le a < b \le m} \left(\int_{-i\infty}^{i\infty} \frac{d\hat{Q}_{ab}}{2\pi/N} \int_{-\infty}^{+\infty} dQ_{ab} \right) e^{N\left[-\sum_{a<b} \hat{Q}_{ab} Q_{ab} + G_1(\hat{Q}) + \alpha G_0(Q)\right]} \qquad (5)$$

where

$$G_1(\hat{Q}) = \mathrm{Log}\left(\sum_{\{J^a = \pm 1\}} e^{\sum_{a<b} \hat{Q}_{ab} J^a J^b} \right) \tag{6a}$$

$$G_0(Q) = \mathrm{Log}\left(\prod_{a=1}^{n} \left[\int_{-i\infty}^{i\infty} \frac{d\lambda^a}{2\pi} \left(e^{-\beta} \int_{-\infty}^{+\infty} dt^a + (1-e^{-\beta}) \int_{\kappa}^{\infty} dt^a \right) \right] e^{\sum_{a=1}^{n} \left(\lambda^a t^a + \frac{1}{2}(\lambda^a)^2 + \frac{1}{2} \sum_{a \neq b} Q^{ab} \lambda^a \lambda^b \right)} \right) \tag{6b}$$

Q_{ab} is the natural order parameter which measures the overlap of the configurations of couplings in two replicas : $Q_{ab} = (1/N)\Sigma_i J_i^a J_i^b$ .The interpretation of $\hat{Q}_{ab}$ involves the probability of one constraint to be strict $\Sigma_k J_k \eta_k^\mu = \kappa\sqrt{N}$ and is more complicated [10].

Starting from the representation (5,6), one can study various types of approximations or Ansätze for the form of the order parameters Q_{ab} and $\hat{Q}_{ab}$ on the saddle point of (5). Hereafter we shall consider successively the annealed approximation, the replica symmetric solution, and the replica symmetry breaking (r.s.b) solution.

3. ANNEALED APPROXIMATION

This is obtained by taking n = 1 and approximating the quenched free energy density $F/N = -(T/N)\overline{\mathrm{Log}\, Z}$ by the annealed average:

$$\frac{F_{ann}}{N} = -\frac{T}{N} \mathrm{Log}\left(\overline{Z}\right) = -T\left[\mathrm{Log}\, 2 + \alpha\, \mathrm{Log}\left(e^{-\beta} + (1-e^{-\beta}) H(\kappa)\right)\right] \tag{7}$$

where H(κ) is an error function defined as :

$$H(K) = \int_K^\infty \frac{dt}{\sqrt{2\pi}} e^{-t^2/2} \tag{8}$$

From (7) we see that the density of internal energy vanishes in the zero temperature limit, while the density of entropy at zero temperature is :

$$\frac{S_{ann}(T=0)}{N} = Log\, 2 + \alpha\, Log\, H(K) \tag{9}$$

As is well known, because of the concavity the logarithm, the annealed free energy is a lower bound to the correct quenched free energy. From this and the fact that the quenched free energy cannot be negative we obtain a lower bound on the zero temperature energy density [11]. This implies in particular, using [9], that :

$$\alpha_c \leq \frac{Log\, 2}{-Log\, H(K)} \equiv \alpha_A(K) \tag{10}$$

For stability $\kappa = 0$, this gives $\alpha_c \leq 1$ which corresponds to the bound from the theory of information.

4. REPLICA SYMMETRIC SOLUTION :

In order to compute the correct quenched free energy in (5-6) we want to take the $n \to 0$ limit. To make this step we need an Ansatz on the form of Q_{ab} and $\hat{Q}_{ab}$ on the saddle point. The simplest approximation is the replica symmetric one where $Q_{ab} = q$ and $\hat{Q}_{ab} = \hat{q}$. This has been studied by Gardner and Derrida [5] and gives

$$G_{r.s.}(q,\hat{q},\beta) \equiv \frac{1}{Nm} \overline{\log Z^m} = \frac{1}{2} q\hat{q} - \frac{\hat{q}}{2} +$$

$$+ \int Dz \log\left(2\cosh z\sqrt{\hat{q}}\right) + \qquad (11)$$

$$+ \alpha \int Dz \log\left[e^{-\beta} + (1-e^{-\beta})\, H\left(\frac{K + z\sqrt{q}}{\sqrt{1-q}}\right)\right]$$

where Dz is a Gaussian integration measure :

$$Dz \equiv \frac{dz}{\sqrt{2\pi}} e^{-z^2/2} \qquad (12)$$

and q, and $\hat{q}$ are obtained by solving the saddle point equations $\partial G_0/\partial q = \partial G_0/\partial \hat{q} = 0$.

Solving these equations one finds the following typical properties. The zero temperature energy is zero for $\alpha \leq \alpha_E(\kappa)$, and becomes strictly positive above. $\alpha_E(\kappa)$ is the point where the value of the overlap q goes to one at zero temperature. From this argument one would get a capacity $\alpha_E(0) = 4/\pi$. However there exists a second characteristic line $\alpha_S(\kappa) < \alpha_E(\kappa)$ such that, for α larger than $\alpha_S(\kappa)$, the zero temperature entropy becomes negative. Firstly this implies that the replica symmetric approximation is wrong at least for $\alpha > \alpha_S(\kappa)$, since the entropy cannot become negative in such a problem with a finite number (2^N) of possible configurations. Secondly this gives a second possible value of the capacity, namely $\alpha_S(0) \equiv .83$, since the vanishing of the zero temperature entropy just means that there are no longer any accessible configurations of the couplings for T going to zero.

As the replica symmetric approximation gives an incorrect solution, it is natural to study its stability against replica symmetry breaking, i.e. to look whether the replica

symmetric saddle point is locally stable. It was already noticed by Gardner and Derrida [5] that for $\kappa = 0$ and $\alpha = \alpha_E(0) = 4/\pi$ the replica symmetric solution is unstable. In fact, in the zero temperature limit we find that the replica symmetric solution is locally unstable when $\alpha \geq \alpha_{AT}(\kappa)$, where $\alpha_{AT}(0) \cong 1.015$

5) PHYSICAL DISCUSSION

Concerning the capacity problem, given by the behaviour at T = 0, the picture which we get from the previous two sections is somewhat confusing. There are four special values of α. Keeping for simplicity to the case $\kappa = 0$, we encounter starting from small α the following special points :

- $\alpha_S \cong .83$: For $\alpha > \alpha_S$ the zero temperature entropy of the replica symmetric saddle point is negative. This is one candidate value for the critical capacity. Beyond this point, the replica symmetric solution, although locally stable, is surely wrong.

- $\alpha_A = 1$: Upper bound of the critical capacity obtained from the annealed approximation.

- $\alpha_{A\,T} \cong 1.015$: For $\alpha > \alpha_{AT}$ the replica symmetric solution becomes locally unstable.

- $\alpha_E = 4/\pi \cong 1.27$: For $\alpha > \alpha_E$ the zero temperature energy becomes stricly positive and $q \to 1$. This is a second candidate value for the critical capacity.

We shall see below from a physical argument that the transition to a non-zero ground state energy must take place at a value of q which is much lower than q = 1. This calls for a first order transition to a r.s.b. solution with two parameters q_0 and q_1

instead of q, where q_0 is close to the replica symmetric value, and q_1 is close to one.

In order to understand this effect, one can resort to some other kind of approximation scheme : if we relax the constraint $J_i = \pm 1$ and keep the J_i free with the only global constraint $\sum_i J_i^2 = N$, we go to a spherical model. This is the model which has been solved exactly by Gardner [3]. The replica symmetric solution is exact and gives a critical capacity $\alpha_c(\kappa)$, with $\alpha_c(0) = 2$, which has been confirmed numerically [12]. The basic idea of Gardner's calculation is as follows : on the surface of the sphere $J^2 = \sum_i J_i^2 = N$, the vectors J with zero energy (defined in (3)) constitute a subspace f , the surface of which is a fraction f of the total surface of the sphere. The typical value of f is then computed with the replica method. In our case of binary couplings the correct sets of couplings are the points which belong both to f and to the hypercube $J_i = \pm 1$. As this hypercube has 2^N points, we can expect that the critical α will be obtained when f contains typically one point of the hypercube, which should occur around $f \sim 1/2^N$. From Gardner's paper this happens (for $\kappa = 0$) at $\alpha \cong .85$; furthemore it is most interesting to notice that for this value of α, the typical overlap of two spherical {J} configurations in f is $q \sim .5$.

It is reasonable to expect that the above argument gives a good approximation to the critical capacity, because the correlation of f with the hypercube can be neglected. (Taking a continuous distribution of patterns with $\overline{\xi_i^\mu} = 0$ and $\overline{(\xi_i^\mu)^2} = 1$, the hyperplanes defining f are no longer correlated with the hypercube, but one gets the same formulas (5-6) and therefore the same capacity [F2]). The only reason which prevents this prediction $\alpha_c \cong .85$ to be strictly exact is the fact that the shape of f itself is not totally random (for $\alpha < 1$, there are no constraints in the subspace orthogonal to all the patterns). A somewhat similar approximation has been attempted in the case of the SK model [13].

The main result we get from this analysis is that when f becomes of order 2^{-N} the typical overlap of two configurations is q ~.5, therefore much lower than 1. This same effect is found in the replica symmetric approximation: when $\alpha = \alpha_S$ (0) (vanishing of the zero temperature entropy), we get a replica symmetric q ≅.56, which is not too far from the above value of .5.

6. REPLICA SYMMETRY BREAKING

We have looked for solutions of (5-6) with one stage of replica symmetry breaking in the hierarchical scheme of Parisi [14]. In such a solution ergodicity is broken, the Gibbs state is decomposed into pure states α of relative weights P_α, the Gibbs expectation value of an observable O being :

$$\langle O \rangle = \sum_\alpha P_\alpha \langle O \rangle_\alpha \tag{13}$$

The breaking à la Parisi involves five order parameters $q_0, q_1, \hat{q}_0, \hat{q}_1$, m; the physical interpretation of q_0, q_1, m, is :

$$\begin{aligned} q_0 &= \frac{1}{N} \overline{\sum_i \langle J_i \rangle_\alpha \langle J_i \rangle_\beta} \qquad \alpha \neq \beta \\ q_1 &= \frac{1}{N} \overline{\sum_i \langle J_i \rangle_\alpha^2} \\ m &= 1 - \overline{\sum_\alpha P_\alpha^2} \end{aligned} \tag{14}$$

In terms of these order parameters we get the solution at 1st stage of replica symmetry breaking $G^{(1)}_{rsb}$

$$G^{(1)}_{rsb}(q_0, \hat{q}_0, q_1, \hat{q}_1, m) \equiv \frac{1}{Nm} \overline{\log Z^m} =$$

$$= \frac{1}{2}\left[m q_0 \hat{q}_0 + (1-m) q_1 \hat{q}_1 - \hat{q}_1\right] +$$

$$+ \frac{1}{m} \int Dz_0 \log\left(\int Dz_1 \left[2 \cosh\left(z_0 \sqrt{\hat{q}_0} + z_1 \sqrt{\hat{q}_1 - \hat{q}_0}\right)\right]^m \right) +$$

$$+ \frac{\alpha}{m} \int Dz_0 \log\left(\int Dz_1 \left[e^{-\beta} + (1-e^{-\beta}) \cdot H\left(\frac{K + z_0 \sqrt{q_0} + z_1 \sqrt{q_1 - q_0}}{\sqrt{1-q_1}} \right) \right]^m \right) \qquad (15)$$

The five parameters q_0, q_1, $\hat{q}_0$, $\hat{q}_1$, m, must be computed through the saddle point equations:

$$0 = \frac{\partial G^{(1)}_{rsb}}{\partial q_0} = \frac{\partial G^{(1)}_{rsb}}{\partial q_1} = \frac{\partial G^{(1)}_{rsb}}{\partial \hat{q}_0} = \frac{\partial G^{(1)}_{rsb}}{\partial \hat{q}_1} = \frac{\partial G^{(1)}_{rsb}}{\partial m} \qquad (16)$$

Stationarity with respect to m is not necessarily required if one looks eventually for a continuous order parameter function q(x) [9], but it is useful at first step, and it is absolutely necessary in the case where there is a first order transition as will happen here.

We have tried to solve these equations numerically. For $q_1 < 1$, we have not found a solution to all the five equations, except at $q_0 = q_1$, or m = 0 (replica symmetric solution). For fixed m (that is, forgetting the last equation in (16)) there can be one or two r.s.b. solutions. If $\alpha > \alpha_{AT}(\kappa)$, there are two solutions, one of which

($q_0^{(1)}$, $q_1^{(1)}$) bifurcates continuously from the replica symmetric one at $\alpha = \alpha_{AT}(\kappa)$, while the second one ($q_0^{(2)}$, $q_1^{(2)}$) satisfies $q_0^{(2)} < q_0^{(1)} < q_1^{(1)} < q_1^{(2)}$. This second solution still exists below $\alpha_{AT}(\kappa)$. Adding the last equation $\partial G^{(1)}_{rsb}/\partial m = 0$ in order to find the optimal m, we find no solution except in a region where $q_1 \to 1$ at finite T.

We have therefore studied directly, what happens for $q_1 = 1$ (at finite temperature). The saddle point equation then implies that $\hat{q}_1 = \infty$. It turns out that in this limit the expression for $G^{(1)}_{rsb}$ simplifies to :

$$G^{(1)}_{rsb}(q_0, \hat{q}_0, 1, \infty, m, \beta) = \frac{1}{m} G_{r.s.}(q_0, m^2\hat{q}_0, \beta m) \qquad (17)$$

where $G_{r.s.}$ is the replica symmetric expression (11) of 1/(Nn) Log Z^n . We can easily solve the saddle point equations: stationarity with respect to q_0 and $\hat{q}_0$ implies that $q_0 = q$ and $\hat{q}_0 = \hat{q}/m^2$, where q and $\hat{q}$ are the replica symmetric order parameters at inverse temperature βm.

Stationarity with respect to m gives :

$$S_{r.s.}(q_0, m^2\hat{q}_0, \beta m) = S_{r.s.}(q, \hat{q}, \beta m) = 0 \qquad (18)$$

which means that βm must be equal to $1/T_c$, where T_c is the temperature where the replica symmetric entropy vanishes.Therefore we get :

- if $\alpha < \alpha_S(k)$ there is no such solution and only the replica symmetric solution exists.
- if $\alpha > \alpha_S(k)$: if $T > T_c$ there is no solution with $0 \le m \le 1$ and we should keep to the replica symmetric solution. (The restriction $0 \le m \le 1$ comes from the physical interpretation of m (14).)

if $T < T_c$ there exists a solution at one step of replica symmetry breaking ; it is defined by :

$$m = \frac{T}{T_c} \;,\; q_0 = q \;,\; \hat{q}_0 = \frac{\hat{q}}{m^2} \;,\; q_1 = 1 \;,\; \hat{q}_1 = \infty \qquad (19)$$

and corresponds to the following thermodynamical behaviour : the free energy is independent of T and equal to the replica symmetric free energy at T_c (see fig.1). The entropy vanishes and the energy is constant, equal to the replica symmetric energy at T_c. We notice that this thermodynamic behaviour is exactly identical to what is found in the simplest spin glass-random energy model [15,16]. However, unlike in this model, the energy levels are not independent random variables. This can be seen from the behaviour in the high temperature phase $T > T_c$ which is not given by the annealed average. Therefore we have a somewhat new behaviour where, although the energy levels are correlated, a total freezing occurs at T_c. The transition is first order in the sense that the order parameter function is not continuous, but the free energy and its first derivatives are still continuous at T_c. As in the simplest spin glass, the order parameter function is a sum of two delta functions [16] :

$$P(Q) = \sum_{\alpha,\beta} P_\alpha P_\beta \; \delta\left(Q - \frac{1}{N}\sum_i \langle J_i \rangle_\alpha \langle J_i \rangle_\beta\right) = \\ = \frac{T}{T_c}\,\delta(Q-q) + \left(1 - \frac{T}{T_c}\right)\delta(Q-1) \qquad (20)$$

where q is the replica symmetric overlap at T_c.

As far as the original problem of capacity is concerned, the present solution predicts a critical capacity $\alpha = \alpha_S(k)$ where $\alpha_S(k)$ is the point where the replica symmetric entropy vanishes. The corresponding curve is plotted in fig. 2.

7. DISCUSSION AND PERSPECTIVES

The replica symmetry breaking solution of the previous section is consistent (no negative entropy, a unique answer for the critical capacity), and in good agreement with the physical picture of sect. 5. It also agrees with the numerical results obtained so far in the literature [17-19]. (In this respect it should be noticed that the result $\alpha_c(\kappa) = \alpha_S(\kappa)$ is in close agreement with the numerical findings of Krauth and Opper over a wide range of values of α ; for instance it has been tested successfully for $\alpha = 10$ where it gives a value of $\kappa = -1.508$ (theory) and $\kappa = -1.52 \pm 0.02$ (simulations)).

An important question is whether this solution is exact or whether it is just a good approximation. As we have seen before, the stability analysis is not necessarily a good indication when there are first order transitions like here. We have rather decided to look at the problem with a second step of replica symmetry breaking à la Parisi [14]. There are then seven order parameters q_0, q_1 q_2, $\hat{q}_0$, $\hat{q}_1$ $\hat{q}_2$, m_1, m_2. In the limit $q_2 = 1$, $\hat{q}_2 = \infty$, we find a result similar to (17), in the sense that the expression $G^{(2)}_{rsb}$ of $1/(Nn)\log Z^n$ reduces to the expression $G^{(1)}_{rsb}$ we had with one step of replica symmetry breaking :

$$G^{(2)}_{rsb}(q_0, \hat{q}_0, q_1, \hat{q}_1, 1, \infty, m_1, m_2, \beta) =$$

$$= \frac{1}{m_2} G^{(1)}_{rsb}\left(q_0, m_2^2 \hat{q}_0, q_1, m_2^2 \hat{q}_1, \frac{m_1}{m_2}, \beta m_2\right) \quad (21)$$

Optimizing with respect to m_1 is equivalent to optimizing with respect to m in the solution of the previous section. As we said before, we have not found any solution of the equation $\partial G^{(1)}_{rsb}/\partial m = 0$ except if $q_1 = 1$ or $q_1 = q_0$, which means that no

new replica symmetry breaking occurs.This has been found numerically for $\kappa = 0$ and essentially in the region where $\alpha_S \leq \alpha \leq \alpha_{AT}$ where most of our numerical work on the saddle point equations was done.Therefore we except that at least in this region the one step r.s.b. solution of the previous section could be exact. Unfortunately this statement depends on some difficult numerical work for solving the saddle point equations at one step r.s.b. Also we have not been able so far to study the r.s.b. to all orders, as had been done in the simplest spin glass problem [16].

To summarize, we have found a one step replica symmetry breaking solution which predicts that the critical α is the one for which the zero temperature replica symmetric entropy vanishes. This solution might be exact, but some more analytic and numerical work to check this is still needed. Extension to biased patterns will be quite interesting.

Let us note finally that the critical line seems to be on the boarder of the replica symmetric region. Perhaps the prospects of finding a workable optimum stability algorithm are not totally bleak.

Acknowledgements :

We thank G. Toulouse for useful discussions. W. K. acknowledges financial support by Studienstiftung des deutschen Volkes.

REFERENCES :

[1] J.J. Hopfield, Proc. Natl. Acad. Sc. USA **79** (1982) 2554.

[2] D.J. Amit, H. Guttreund and H. Sompolinsky, Phys. Rev. Lett. **55** (1985) 1530.

[3] E. Gardner, Europhys. Lett. 4 (1987) 481, and J. Phys. A **21** (1987) 257.

[4] T.M. Cover, I.E.E.E. transactions EC **14**,3 (1965) 326.

[5] E. Gardner and B. Derrida, J. Phys. A **21** (1988) 271.

[6] H. Sompolinsky, in *Heidelberg colloquium on glassy dynamics*, eds. L. van Hemmen and I. Morgenstern, Lecture Notes in Physics **275** (1987) 485, (Springer Verlag).

[7] B.M. Forrest, J. Phys. A **21** (1988) 245.

[8] W. Krauth, J.P. Nadal and M. Mézard, J. Phys. A **21** (1988) 2995.

[9] For a review, see M. Mézard, G. Parisi and M.A. Virasoro, "Spin glass theory and beyond", World Scientific (Singapore) 1987.

[10] M. Mézard, L.P.T.E.N.S. preprint 89/1.

[11] G. Toulouse and J. Vannimenus, Phys. Rep. **67** (1980) 47.

[12] W. Krauth and M. Mézard, J. Phys. A **20** (1987) L745.

[13] G. Toulouse, private communication.

[14] G. Parisi, J. Phys. A **13** (1980), L115, 1101 ; see also [9].

[15] B. Derrida, Phys. Rev. B **24** (1981) 2613.

[16] D.J. Gross and M. Mézard, Nucl. Phys. B **240** (1984) 431.

[17] E. Amaldi and S. Nicolis, Rome University preprint (642 (1988).

[18] W. Krauth and M. Opper, J. Phys. A, to appear.

[19] E. Gardner and B. Derrida, Saclay preprint S Ph T / 88 - 198

FOOT NOTES :

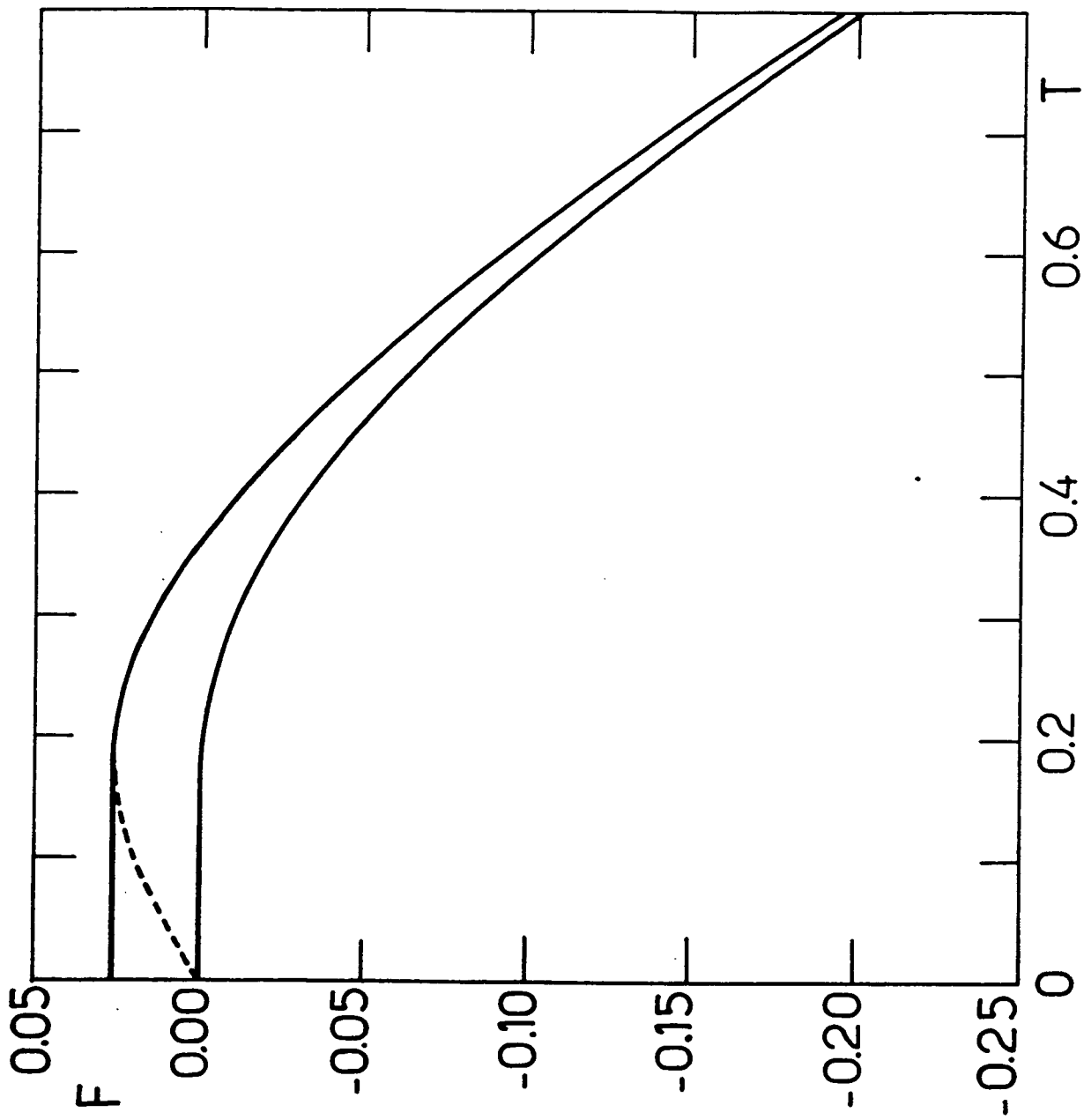
F1 : Intermediate cases have been studied by Gutfreund et al. (private communication).

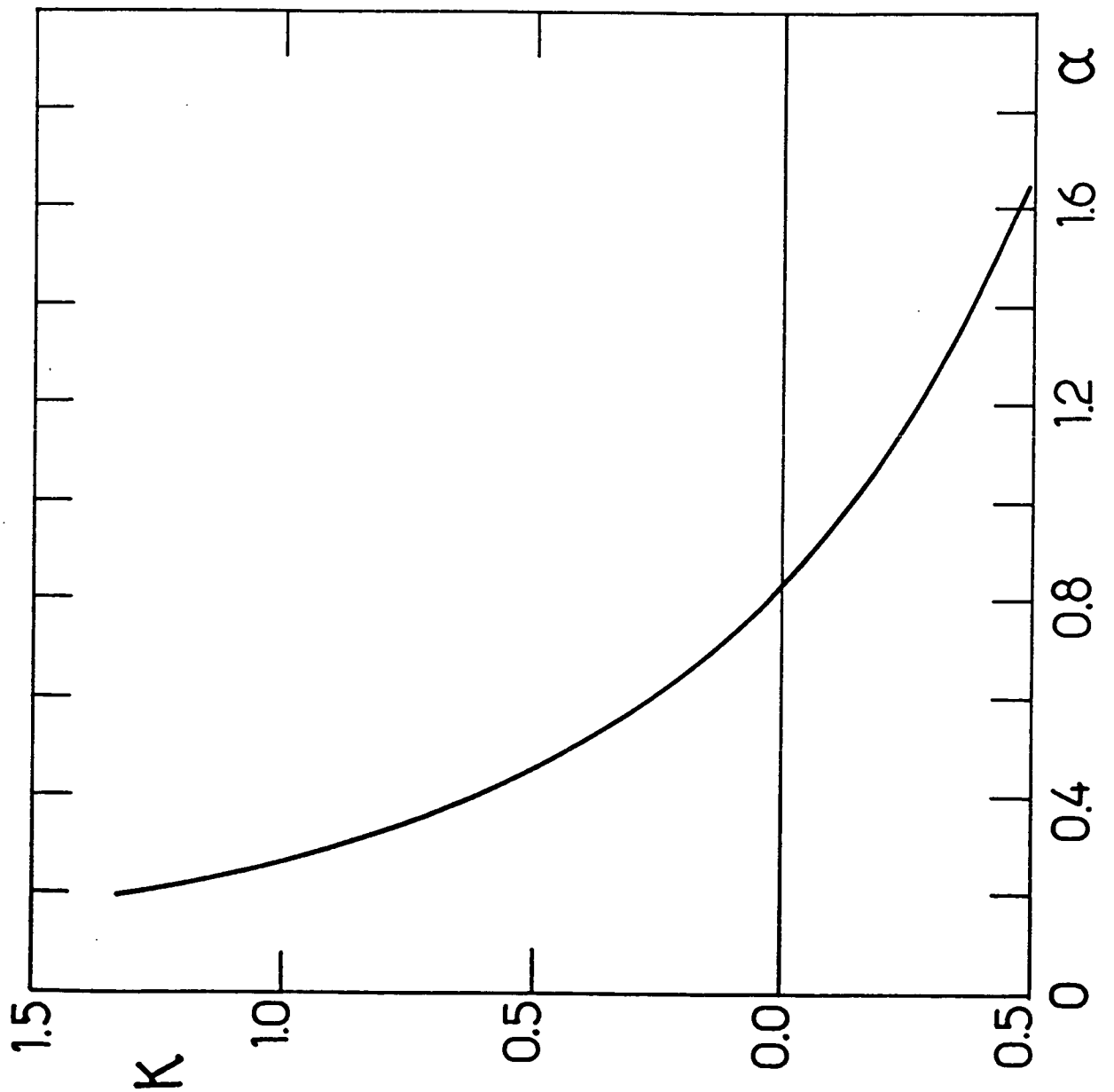
F2 : This has been used to diminish the finite size effects in numerical simulations of this problem with small values of N [18].

FIGURE CAPTIONS :

Fig. 1 : Free energy versus temperature, for $\alpha = 1$, $\kappa = 0$. The lowest curve is the annealed approximation. The upper dotted curve is the replica symmetric solution and the upper full curve is the replica symmetry breaking solution.

Fig. 2 : Phase diagram predicted by the one step replica symmetry breaking solution of section 6. Shown is the critical stability κ_c versus the number of stored patterns per neuron $\alpha = p/N$. The system can learn in the region $\kappa \leq \kappa_c(\alpha)$. The critical capacity α_c is given by $\kappa_c(\alpha_c) = 0$.


0.05
0.00
-0.05
-0.10
-0.15
-0.20
-0.25
F
0
0.2
0.4
0.6
T


K
1.5
1.0
0.5
0.0
0.5
0
0.4
0.8
1.2
1.6
α

EUROPHYSICS LETTERS 1 February 1989

Europhys. Lett., 8 (3), pp. 213-218 (1989)

The Cavity Method and the Travelling-Salesman Problem.

W. KRAUTH(*) and M. MÉZARD

Laboratoire de Physique Théorique de l'E.N.S.(**)
24 rue Lhomond, 75231 Parix Cedex 05, France

(received 12 September 1988; accepted in final form 21 November 1988)

PACS. 05.20 – Statistical mechanics.
PACS. 75.50 – Studies of specific magnetic materials.

Abstract. – For the random link travelling-salesman problem we solve the zero-temperature cavity equations, assuming that there is only one pure state. We get precise predictions for the length of the optimal tour and the probability distribution of links in this tour. These are compared with numerical simulations using the Lin-Kernighan algorithm and the one-tree relaxation of Held and Karp.

Among the several problems of combinatorial optimization which have been studied from the point of view of statistical physics [1, 2], the travelling salesman occupies a choice place: it is considered as an important laboratory for testing ideas on *NP*-complete problems. The problem consists in the following: given N points $i = 1, ..., N$, and the distances l_{ij} between points i and j, one must find the shortest closed line of N links going through all the points. For N large (thermodynamic limit) this can be seen as a problem in statistical physics. One introduces a temperature T, and each tour ϑ of length l_ϑ is weighted by a Boltzmann factor $\exp[-l_\vartheta/T]$. One can then try to predict the asymptotic value (for $N \to \infty$) of the length of the optimal tour, for some sets of samples. One such set which has received much attention is the random link problem in which the l_{ij} are independent random variables, with a probability distribution $\rho(l)$ (here we keep to the symmetric case $l_{ij} = l_{ji}$).

There have been several studies of this problem in the literature before [1-8], mainly numerical ones. On the analytical side, the TSP can be written as a self-avoiding walk interacting through random couplings. It has been studied with the replica method [3] within the replica symmetric approximation. However, the equations satisfied by the order parameters were very complicated. An estimate of the length of the optimal tour was obtained only in one specific case of «flat distances»: a uniform distribution of distances on [0, 1].

On the numerical side, physicists have concentrated on the use of simulated annealing [1, 4, 5] and on the study of finite-temperature properties.

(*) Laboratoire de Physique Statistique.

(**) Laboratoire Propre du Centre National de la Recherche Scientifique, associé à l'Ecole Normale Supérieure et à l'Université de Paris-Sud.

Our analytic approach does not use replicas, but the cavity method [2]. The cavity equations have been written down in [6], within the hypothesis that there exists only one pure state. Adding a new point $i = 0$ to a system of N points $i = 1, \ldots, N$, the magnetization of the new point is $m_0 = 2A_1/((A_1)^2 - A_2)$, where $A_k = \sum_{i=1}^{N} T_{0i}^k m_i^k$ and m_i is the magnetization on site i before the addition of the new point. The coupling constant is $T_{0i} = \exp[-\beta N^\delta l_{0i}]$, with β the inverse temperature. For a length distribution $\rho(l)$ scaling as $\rho(l) \sim l^r/r!$ $(l \to 0)$, $\delta = 1/(r+1)$ must be chosen to have a good thermodynamic limit [7].

As we are interested in the zero-temperature limit $\beta \to \infty$, it is natural to parametrize $m_i = \exp[\beta\varphi_i]$, and the cavity equation for m_0 simplifies at zero temperature to

$$\varphi_0 = N^\delta l_{02} - \varphi_2 . \tag{1}$$

In eq. (1) we reordered the points $i = 1, \ldots, N$ in such a way that $N^\delta l_{01} - \varphi_1 \leqslant \leqslant N^\delta l_{02} - \varphi_2 \leqslant \ldots \leqslant N^\delta l_{0N} - \varphi_N$.

Denoting by $\overline{(\ )}$ the average over the distribution of links, eq. (1) implies a self-consistent equation for the probability distribution of the φ_i:

$$P(\varphi) = \overline{\delta(\varphi - \varphi_0)} = \overline{\delta(\varphi - \varphi_i)}, \quad i = 1, \ldots, N. \tag{2}$$

To write this self-consistency equation explicitly, we first deduce from eq. (2) the distribution $\Pi(\chi)$ of $\chi = N^\delta l - \varphi$

$$\Pi(\chi) = \frac{1}{N} \int_0^\infty \frac{l^r}{r!} \mathrm{d}l\, P(l - \chi). \tag{3}$$

(For N finite the integral has a cut-off, which diverges with N.) In the cavity method the N random variables $\chi_i = N^\delta l_{0i} - \varphi_i$ are independent, so that the distribution of the second smallest of all the χ_i's (see eq. (1)) is easily derived. Using eqs. (2), (3), this leads to

$$P(\varphi) = N(N-1)\Pi(\chi)\left(\int_{-\infty}^{\chi} \Pi(u)\,\mathrm{d}u\right)\left(\int_{\chi}^{\infty} \Pi(u)\,\mathrm{d}u\right)^{N-2}, \tag{4}$$

which for large N is equal to $P(\varphi) = (\mathrm{d}G/\mathrm{d}\varphi)\, G(\varphi) \exp[-G(\varphi)]$, with

$$G(\varphi) = \int_0^\infty \mathrm{d}u \frac{u^{r+1}}{(r+1)!} P(u - \varphi). \tag{5}$$

From this we can deduce the closed integral equation for G, the order parameter function of the TSP:

$$G(x) = \int_{-x}^{\infty} \frac{(x+y)^r}{r!} \{1 + G(y)\} \exp[-G(y)]\,\mathrm{d}y, \tag{6}$$

G can be computed precisely by iteration. Its relation with the order parameter function defined in the replica approach is complicated.

From $G(x)$, we now compute the length of the optimal tour. Let us define the distribution of the (rescaled) links, $\mathscr{P}(l)$ as

$$\mathscr{P}(l) = \frac{1}{N} \overline{\sum_{1 \leqslant i < j \leqslant N} \delta(l - l_{ij} N^\delta n_{ij})} = \frac{1}{2} \overline{\sum_i \delta(l - l_{0i} N^\delta n_{0i})}, \tag{7}$$

where n_{ij} is the thermal average of the occupation number of link ij. At zero temperature, the cavity equation for n_{0l}, written down in [6], tells us that $n_{01} = n_{02} = 1$, and $n_{0i} = 0\,(i \geq 3)$. (The points are ordered as introduced after eq. (1).) Using this result and eq. (7), we find after some work the following expression for $\mathscr{P}$ at zero temperature:

$$\mathscr{P}(l) = \frac{1}{2}\frac{l^r}{r!}\left(-\frac{\partial}{\partial l}\right)\int \mathrm{d}x\,(1+G(x))\exp[-G(x)](1+G(l-x))\exp[-G(l-x)], \tag{8}$$

where G is the solution of eq. (6). The length of the optimal tour (as $N \to \infty$) is then $L \sim N^{1-1/(r+1)}\hat{L}_r$ with

$$\hat{L}_r = \int \mathscr{P}(l)\,\mathrm{d}l = \frac{r+1}{2}\int \mathrm{d}x\,G(x)\,\{1+G(x)\}\exp[-G(x)]. \tag{9}$$

We summarize our theoretical results: for any r we can solve for $G(x)$ in eq. (6) and then get from eqs. (8), (9) the length of the optimal tour and the distribution of lengths of occupied links in this tour.

In the case $r = 0$ (flat distances) we find $L = \hat{L}_{r=0} = 2.0415\ldots$ and for $r = 1$ we have $\hat{L}_{r=1} = 1.8175\ldots$. This random link case with $r = 1$ is an approximation for the Euclidean TSP in 2 dimensions. Following [9], we estimate the length of a TSP for N points uniformely distributed in the unit square as $\hat{L}_{r=1}/\sqrt{2\pi} = 0.7251$ which is close to the known bounds [10]. A commonly accepted numerical result is 0.749.

We now turn to the numerical checks of these results (restricted to the case of flat distances $r = 0$, *i.e.*, $l_{ij} = l_{ji}$ are uniform random numbers on [0, 1]). The TSP being *NP*-complete, there are no tractable algorithms for solving large instances of it. The long-standing interest into the TSP has, however, led to the following situation: there do exist (very involved) algorithms using linear programming [11] or branch and bound strategies [12], the best of which are presently capable of solving instances with several hundreds of points.

On the other hand, a number of methods are available which provide good sub-optimal tours and thus upper bounds on the optimal tour length. Furthermore, algorithms are known which solve «relaxed» TSP problems, leading to lower bounds on the length of the shortest tour. It has been known for some time [12, 13] that these bounds might be quite tight, even for the random link TSP.

We used two well-known heuristics to determine upper and lower bounds for the TSP, following [13]. For the lower bound we used the Lagrangian one-tree relaxation of Held and Karp [14] (our implementation follows ref. [12]). This relaxation solves the problem

$$L_{\text{one-tree}} = \max_{\lambda}\left\{\left(\min_{\substack{n_{ij}=0,1 \\ n_{ij}\in \text{one-tree}}} \sum_{ij}(l_{ij}+\lambda_i+\lambda_j)\,n_{ij}\right) - 2\sum_j \lambda_j\right\}, \tag{10}$$

where the min (for fixed λ) is taken on certain spanning graphs containing exactly one cycle (cf. [12]). The upper bound on the optimal TSP-tour was determined with the Lin-Kernighan algorithm, which was implemented using the original ref. [15]. Both methods are quite fast: for an instance of 400 points we can calculate both the upper and the lower bound in less than 1 minute on a Convex C1 computer.

Our numerical results for the bounds are given in fig. 1. In both cases we find good agreement with a dependence linear in $1/N$. As $N \to \infty$ we get $2.039 \pm 3 \cdot 10^{-3} \leqslant$ $\leqslant L \leqslant 2.21 \pm 1.3 \cdot 10^{-2}$. The theoretical value $L = 2.0415\ldots$ is in agreement with the bounds,

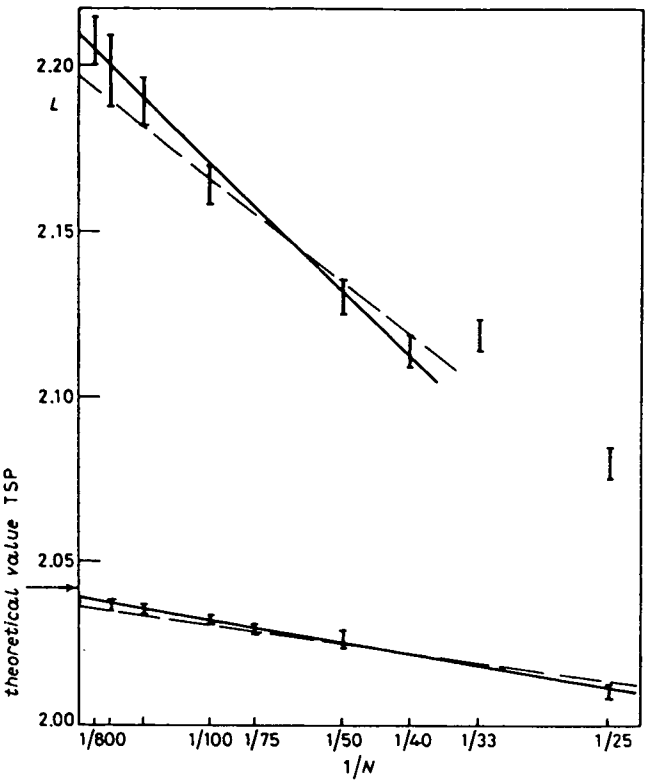


Fig. 1. – Upper and lower bounds for the optimal tour length of the random link TSP (flat distances $l_{ij} \in [0,1]$). The upper curve was produced using the Lin-Kernighan algorithm, averaged over, *e.g.*, 360 samples for $N = 800$, 5000 samples for $N = 25$. The lower curve results from the Langrangian one-tree relaxation with, *e.g.*, 3000 samples ($N = 400$), 20000 samples ($N = 25$).

in fact it coincides within our 2‰ level of precision with the results of the one-tree relaxation.

We compared also the probability distribution of links both of the Lin-Kernighan solution and of the one-tree relaxation with the theoretical prediction. To eliminate sample-to-sample fluctuations, we rescalded the lengths of occupied links, in each solution, by the average length of the occupied links. The distribution of these rescaled lengths, for $N = 800$, was then averaged over, *e.g.*, 100 Lin-Kernighan tours and compared to the theoretical prediction

$$I(x) = \int_0^{xL_{T=0}} \mathrm{d}l \; \mathscr{P}(l) .$$

The agreement is excellent, see fig. 2. We found equally good agreement with the one-tree relaxation. We do not know why a simple length rescaling makes agree the Lin-Kernighan solution (which we suspect to be 10% above optimal) the one-tree relaxation (which does not produce a tour) and the theoretical result for the optimal solution.

The simulations give the numerical value of the optimal TSP tour length, albeit on a rather crude level. They also illustrate the power of the Lin-Kernighan algorithm, compared to simpler heuristics which up to now have all given bounds on the optimal tour length diverging like $\log N$ for $N \to \infty$ [4]. Finally, the extremely close agreement between the one-

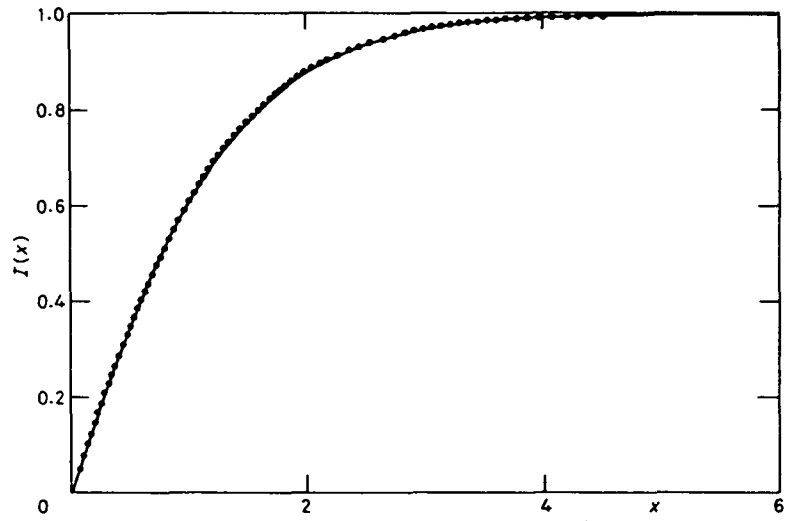


Fig. 2. – Integrated distribution function $I(x)$ for the occupied links in TSP solutions *vs.* reduced length variable x ($x = 1$ corresponds to the mean length of occupied links in each sample). The curve gives the theoretical result, the dotted line the numerical results for Lin-Kernighan solutions (100 samples at $N = 800$).

tree relaxation and the proposed value for the optimal tour suggests to us the exciting conjecture that the two values may indeed be identical. It would be interesting to apply the cavity method directly to the one-tree relaxation.

We found some support for the conjecture in additional simulations using the Lin-Kernighan heuristics repeatedly on each sample and keeping the lowest result. In this way we can lower the upper bound considerably for finite N, but not for $N \to \infty$. (Preliminary results with an exact algorithm, by Kirkpatrick [16], agree with an optimal TSP length around 2.04.)

In conclusion we remark that the recently developed cavity method of statistical physics in combination with rather traditional tools of Operations Research have led to a better understanding of the random link TSP.

* * *

Stimulating discussions with G. PARISI and with S. KIRKPATRICK are gratefully acknowledged. The simulations on the Convex C1 were supported by GRECO 70 «Expérimentation Numérique». WK acknowledges financial support by Studienstiftung des deutschen Volkes.

REFERENCES

[1] KIRKPATRICK S., GELATT C. D. and VECCHI M. P., *Science*, 220 (1983) 671.
[2] MÉZARD M., PARISI G. and VIRASORO M. A., *Spin Glass Theory and Beyond* (World Scientific, Singapore) 1987.
[3] MÉZARD M. and PARISI G., *J. Phys. (Paris)*, 47 (1986) 1285.
[4] KIRKPATRICK S. and TOULOUSE G., *J. Phys. (Paris)*, 46 (1985) 1277.
[5] SOURLAS N., *Europhys. Lett.*, 2 (1986) 919.
[6] MÉZARD M. and PARISI G., *Europhys. Lett.*, 2 (1986) 913.

[7] VANNIMENUS J. and MÉZARD M., *J. Phys. (Paris) Lett.*, 45 (1984) L-1145.
[8] BASKARAN G., FU Y. and ANDERSON P. W., *J. Stat. Phys.*, 45 (1986) 1.
[9] MÉZARD M. and PARISI G., LPTENS preprint 88/21, to appear in *J. Phys. (Paris)*.
[10] BEARDWOOD J., HALTON J. H. and HAMMERSLEY J. M., *Proc. Cambridge Philos. Soc.*, 55 (1959) 299.
[11] PADBERG M. W. and GRÖTSCHEL M., in *The Traveling Salesman Problem*, edited by E. L. LAWLER, J. K. LENSTRA, A. H. G. RINNOOY KAN and D. B. SHMOYS (Wiley, Chichester) 1985, p. 307.
[12] BALAS E. and TOTH P., in *The Traveling Salesman Problem*, op. cit. [11].
[13] JOHNSON D., *Nature (London)*, 330 (1987) 525.
[14] HELD R. M. and KARP R. M., *Oper. Res.*, 18 (1970) 1138.
[15] LIN S. and KERNIGHAN B. W., *Oper. Res.*, 21 (1973) 498.
[16] KIRKPATRICK S., *Talk at the Cargèse school «Common trends in particle and condensed matter Physics», May 24-June 3, 1988.*

TABLE DES MATIERES

RESUME

Dans la première partie nous étudions l'apprentissage et le rappel dans des réseaux de neurones à une couche (modèle de Hopfield). Nous proposons un algorithme d'apprentissage qui est capable d'optimiser la 'stabilité', un paramètre qui décrit la qualité de la représentation d'un pattern dans le réseau. Pour des patterns aléatoires, cet algorithme permet d'atteindre la borne théorique de Gardner. Nous étudions ensuite l'importance dynamique de la stabilité et d'un paramètre concernant la symétrie de la matrice de couplages. Puis, nous traitons le cas où les couplages ne peuvent prendre que deux valeurs (inhibiteur, excitateur). Pour ce modèle nous établissons les limites supérieures de la capacité par un calcul numérique, et nous proposons une solution analytique.

La deuxième partie de la thèse est consacrée à une étude détaillée - du point de vue de la physique statistique - du problème du voyageur de commerce. Nous étudions le cas spécial d'une matrice aléatoire de connexions. Nous exposons la théorie de ce problème (suivant la méthode des répliques) et la comparons aux résultats d'une étude numérique approfondie.