Perfect learning and power law in learning from stochastic examples by Ising perceptrons: analysis under one-step replica symmetry breaking ansatz

1996 J. Phys. A: Math. Gen. 29 L439

(http://iopscience.iop.org/0305-4470/29/17/004)

J. Phys. A: Math. Gen. **29** (1996) L439–L445. Printed in the UK

LETTER TO THE EDITOR

Perfect learning and power law in learning from stochastic examples by Ising perceptrons: analysis under one-step replica symmetry breaking ansatz

Tatsuya Uezu† and Yoshiyuki Kabashima‡

Department of Physics, Nara Women's University, Nara 630, Japan

Received 19 April 1996, in final form 24 May 1996

Abstract. We study learning from stochastic examples by perceptrons with Ising weights in the framework of statistical mechanics. Under the one-step replica symmetry breaking ansatz, the behaviour of learning curves are classified according to some local property of the rules by which examples are drawn. The conditions for the existence of perfect learning together with other behaviours of the learning curves are given precisely. The results agree with those obtained by Seung (1995) using a refined annealled approximation.

In recent years, the problem of learning from examples by feed forward networks has attracted many researchers. In order to investigate how good a generalization ability can be acquired through learning, learning curves of the generalization error ϵ_g, which is the probability of false prediction on a novel example, have been calculated for various types of networks [1]. These studies revealed the following feature of learning. When the number of examples p is small relative to the number of synaptic weights N, the learning curves exhibit a rich behaviour depending on the details of the networks. In contrast, only a few types of behaviour are observed when $\alpha = p/N$ is large [2–7]. For example, learning curves of networks with continuous weights all exhibit power laws [2–8],

$$(\epsilon_g - \epsilon_{min}) \propto \alpha^{-\gamma}.$$

On the other hand, the learning behaviour for the case of discrete weights is quite different from those for the continuous cases [9, 10]. The most drastic difference is the existence of perfect learning for the Ising networks in which the values of synaptic weights are constrained to $+1$ or -1. That is, for the deterministic and realizable cases, learners' weight vectors coincide with the teacher's weight vector at a finite α. Then, natural questions arise about the existence of perfect learning when the weights are Ising and the rule to be learnt is stochastic.

Recently, Seung answered these questions by using a refined annealled approximation [11]. The target rule he considered is a (stochastic) relation between the N-dimensional input vector $\boldsymbol{x}$ and binary output $r \in \{1, -1\}$. He classified the learning behaviour of

† E-mail address: uezu@cc.nara-wu.ac.jp

‡ Present address: Department of Computational Intelligence and Systems Science, Interdisciplinary Graduate School of Science and Engineering, Tokyo Institute of Technology, Yokohama 226, Japan. E-mail address: kaba@dis.titech.ac.jp

0305-4470/96/070439+07$19.50

Ising networks by introducing the following two exponents y and z. The first exponent y is associated with $\rho(\epsilon_g)$ which is the logarithm of the number of weight vectors whose generalization errors have a value ϵ_g. He assumed that $\rho(\epsilon_g)$ increases as

$$\rho(\epsilon_g) \sim \mathcal{O}((\Delta\epsilon_g)^y)$$

when $\Delta\epsilon_g = \epsilon_g - \epsilon_{min}$ is small, where ϵ_{min} is the minimum value of the generalization error obtained by the unique optimal weight vector $\boldsymbol{w}^o$. The second exponent z is introduced to characterize $e_d(\boldsymbol{w}, \boldsymbol{w}^o)$ which is the probability that the output for the weight vector $\boldsymbol{w}$ differs from that for the optimal weight vector $\boldsymbol{w}^o$. He also assumed that $e_d(\boldsymbol{w}, \boldsymbol{w}^o)$ is scaled as follows:

$$e_d(\boldsymbol{w}, \boldsymbol{w}^o) \sim \mathcal{O}((\Delta\epsilon_g)^z).$$

The minimum-error algorithm, which minimizes the number of false predictions on the presented examples, is a natural learning strategy. He classified the behaviour of learning with this algorithm according to the values of $v \equiv y + z$. His results are as follows: (a) there is a first-order transition for $v > 2$; (b) $\Delta\epsilon_g$ decays as a power law as $\sim \alpha^{-\frac{1}{2-v}}$ for $v < 2$; and (c) there is a second-order transition or $\Delta\epsilon_g$ decays exponentially for $v = 2$. It should be remarked that these results concern the upper bounds of the generalization errors, which might differ from the typical or average behaviour. Further, although his method is a refined one, the application of annealled approximations to learning unrealizable rules could yield very wrong behaviour [2, 11]. Therefore, it is necessary to investigate the problem by other methods and judge the validity of the above results.

The purpose of this paper is to investigate conditions for the existence of perfect learning in the problem of learning from stochastic examples by perceptrons with Ising weights in the framework of statistical mechanics. In order to perform a more precise analysis than the annealled approximation, we investigate a stochastic learning model with Ising weights by the replica method with the one-step replica symmetry breaking (RSB) ansatz.

The problem considered in this paper is as follows. We consider a stochastic target relation between the N-dimensional input vector $\boldsymbol{x}$ and binary output $r \in \{1, -1\}$ which is represented by a conditional probability $p_r(r|\boldsymbol{x})$. It is assumed that the input vector $\boldsymbol{x}$ is normalized as $|\boldsymbol{x}| = \sqrt{N}$ and $p_r(r|\boldsymbol{x})$ is a function of the inner product between the input $\boldsymbol{x}$ and the optimal Ising weight $\boldsymbol{w}^o$ as

$$p_r(+1|\boldsymbol{x}) = \mathcal{P}(u^0) = \frac{1 + P(u^0)}{2} \tag{1}$$
$$u^0 \equiv (\boldsymbol{x} \cdot \boldsymbol{w}^o)/\sqrt{N}.$$

We further assume that the function $P(u)$ is increasing with respect to u and behaves as

$$P(u) \simeq a\,\mathrm{sgn}(u)|u|^\delta \qquad (\delta \geqslant 0) \tag{2}$$

near $u = 0$. Further, $P(-u) = -P(u)$ is assumed for brevity.

In a learning model studied by Opper and Haussler [4], the target rule is a perceptron, whose sign of output is reversed to the opposite by noise with a probability λ (output noise model). Such a rule is represented by the conditional probability $p_r(+1|\boldsymbol{x}) = 1 - p_r(-1|\boldsymbol{x}) = \lambda + (1 - 2\lambda)\Theta[(\boldsymbol{w}^o \cdot \boldsymbol{x})/\sqrt{N}]$, which corresponds to the case of $\delta = 0$ in our model. In the above expression, $\Theta(x)$ is the Heaviside function. Another typical noise is the input noise studied by Györgyi and Tishby [3] (input noise model). In their model, the target rule is a perceptron whose input is corrupted by uncorrelated Gaussian noise with mean zero and $r = \mathrm{sgn}[\boldsymbol{w}^o \cdot (\boldsymbol{x} + \boldsymbol{\eta})]$ is finally returned. The conditional probability in this case is represented as

$$p_r(+1|\boldsymbol{x}) = 1 - p_r(-1|\boldsymbol{x}) = H[-(\boldsymbol{w}^o \cdot \boldsymbol{x})/\sqrt{\langle(\boldsymbol{w}^o \cdot \boldsymbol{\eta})^2\rangle}]$$

where

$$H(u) = \int_u^\infty \mathrm{d}x \ \exp[-x^2/2]/\sqrt{2\pi}$$

and $\langle\cdots\rangle$ represents the average over $\boldsymbol{\eta}$. This corresponds to the case of $\delta = 1$ in our model. Namely, although assumption (2) is seemingly artificial and weights in the above two models are assumed not discrete but continuous, our model includes the above typical models of learning disturbed by noise with specific values of the parameter δ.

The results obtained in this paper are summarized as follows. It turns out that the replica symmetry (RS) solution becomes inadequate for the case of large α and the RSB solution should be considered instead. The behaviour of ϵ_{g} is classified into the following three categories according to the value of δ.

(1) If $\delta < \frac{1}{2}$, when α increases the RSB solution disappears at a finite α and a first-order phase transition from the RSB solution to perfect learning takes place.

(2) If $\delta > \frac{1}{2}$, perfect learning does not exist and ϵ_{g} decays as a power law with a logarithmic correction, $\Delta\epsilon_{\mathrm{g}} \sim (\ln\alpha/\alpha)^{(1+\delta)/(2\delta-1)}$.

(3) If $\delta = \frac{1}{2}$, perfect learning does not exist and ϵ_{g} decays exponentially, $\Delta\epsilon_{\mathrm{g}} \sim \mathrm{e}^{-3F_0\alpha}$, where F_0 is a constant.

In our model, the exponents y and z are expressed as $y = 2/(1+\delta)$, $z = 1/(1+\delta) = y/2$ respectively, and then $v = 3/(1+\delta)$. Therefore, $\delta = (3-v)/v$ follows and it is found that our results on the typical learning behaviour are consistent with Seung's results which are the upper bounds of the learning curves.

Now, let us proceed to detailed calculations. We assume that a set of p examples $\xi_p = \{(\boldsymbol{x}_1, r_1), (\boldsymbol{x}_2, r_2), \ldots, (\boldsymbol{x}_p, r_p)\}$ is obtained as follows. $\boldsymbol{x}_i$ is independently and uniformly drawn from an N-dimensional sphere of radius $\sqrt{N}$ at the origin and r_i is obtained with the conditional probability $p_r(r_i|\boldsymbol{x}_i)$ for each $\boldsymbol{x}_i$. For the given realization of examples ξ_p, the number of false predictions is defined as

$$E[w, \xi_p] = \sum_{\mu=1}^{p} \Theta(-r_\mu u_\mu) \qquad u_\mu \equiv (\boldsymbol{x}_\mu \cdot \boldsymbol{w})/\sqrt{N}. \tag{3}$$

The performance of the learning is evaluated by the generalization error ϵ_{g}. This is expressed as

$$\begin{aligned} \epsilon_{\mathrm{g}} &= \langle\langle \mathcal{P}(u^0)(1-\Theta(u)) + (1-\mathcal{P}(u^0))\Theta(u)\rangle\rangle \\ &= \epsilon_{\min} + \int_0^\infty Dy P(y) H\left(\frac{Ry}{\sqrt{1-R^2}}\right) \\ \epsilon_{\min} &= \tfrac{1}{2} - \int_0^\infty Dy P(y) \end{aligned} \tag{4}$$

where $\langle\langle\cdots\rangle\rangle$ represents the average over a novel example and $\epsilon_{\min}$ is the minimum value of the generalization error obtained by the optimal weight $\boldsymbol{w}^{\mathrm{o}}$. R is the overlap between the optimal weight vector and a weight vector of a learner, and $Dy = \exp(-y^2/2)\,\mathrm{d}y/\sqrt{2\pi}$. In particular, when $\Delta R = 1 - R$ is small, we obtain the relation

$$(\epsilon_{\mathrm{g}} - \epsilon_{\min}) \propto (\Delta R)^{(1+\delta)/2}. \tag{5}$$

From the energy defined by equation (3) the partition function Z with inverse temperature β is given by

$$Z = \mathrm{Tr}_w\, \mathrm{e}^{-\beta E[w,\xi_p]} = \mathrm{Tr}_w \prod_{\mu=1}^{p} [\mathrm{e}^{-\beta} + (1-\mathrm{e}^{-\beta})\Theta(r_\mu u_\mu)].$$

The average free energy per weight f is calculated by the standard recipe

$$-\beta N f = \langle\langle \ln Z \rangle\rangle_{\xi_p} = \lim_{n\to 0} \frac{1}{n}(\langle\langle Z^n \rangle\rangle_{\xi_p} - 1)$$

where $\langle\langle \cdots \rangle\rangle_{\xi_p}$ denotes the average over quenched variables.

$\langle\langle Z^n \rangle\rangle_{\xi_p}$ becomes a function of several replica order parameters, namely the overlap between weight vectors of learners $q^{\alpha\beta} = (\boldsymbol{w}^\alpha \cdot \boldsymbol{w}^\beta)/N$, its conjugate $\hat{q}^{\alpha\beta}$, the overlap between the weight vector of a learner and the optimal weight vector $R^\alpha = (\boldsymbol{w}^{\mathrm{o}} \cdot \boldsymbol{w}^\alpha)/N$, and its conjugate $\hat{R}^\alpha$. First, we consider the RS solution which has the simplest symmetry. That is, we put $q^{\alpha\beta} = q$, $\hat{q}^{\alpha\beta} = \hat{q}$, $R^\alpha = R$ and $\hat{R}^\alpha = \hat{R}$. Then, the RS free energy f_{RS} is

$$-\beta f_{\mathrm{RS}}(q, \hat{q}, R, \hat{R}, \beta) = -\frac{\hat{q}}{2}(1-q) - R\hat{R} + \alpha \int Dy 2\mathcal{P}(y) \times \int Du \ln \tilde{H}\left(\frac{\sqrt{q-R^2}u - Ry}{\sqrt{1-q}}\right) + \int Dt \ln[2\cosh(\sqrt{\hat{q}}t + \hat{R})] \tag{6}$$

$\tilde{H}(t) \equiv \mathrm{e}^{-\beta} + (1-\mathrm{e}^{-\beta})H(t)$.

Since we adopt the minimum-error algorithm as the learning strategy, we have to choose weights with minimum errors, and on that account we take the limit $T \to +0$. However, it turns out that the entropy of the RS solution becomes negative as $T \to +0$ in the case of large α. Thus, we have to consider the breaking of the replica symmetry [12]. In the one-step RSB solution, the matrix $q^{\alpha\beta}$ is divided into $(n/m)^2$ small matrices with the dimension $m \times m$. The components of each off-diagonal matrix are all q_0 and the components of each diagonal matrix are q_1 except for diagonal components with the value zero. Likewise, $\hat{q}_0$ and $\hat{q}_1$ are defined for the matrix $\hat{q}^{\alpha\beta}$. Further, $R^\alpha = R$ and $\hat{R}^\alpha = \hat{R}$ are assumed. Then, the one-step RSB free energy, f_{RSB}, is

$$\begin{aligned}-\beta f_{\mathrm{RSB}}(q_0, \hat{q}_0, q_1, \hat{q}_1, R, \hat{R}, \beta, m) &= -\frac{\hat{q}_1}{2}(1-q_1) \\ &+ \frac{m}{2}(\hat{q}_0 q_0 - \hat{q}_1 q_1) - R\hat{R} + \frac{\alpha}{m}\int Dy 2\mathcal{P}(y) \int Dz_0 \ln \\ &\times \int Dz_1 \left\{\tilde{H}\left(\frac{\sqrt{q_0 - R^2}z_0 + \sqrt{q_1 - q_0}z_1 - Ry}{\sqrt{1-q_1}}\right)\right\}^m \\ &+ \frac{1}{m}\int Dz_0 \ln \int Dz_1 \left[2\cosh\left(\sqrt{\hat{q}_0}z_0 + \sqrt{\hat{q}_1 - \hat{q}_0}z_1 + \hat{R}\right)\right]^m .\end{aligned} \tag{7}$$

Further, according to Krauth–Mézard [13], we take the limits $q_1 \to 1$ and $\hat{q}_1 \to \infty$. Thus, we obtain

$$f_{\mathrm{RSB}}(q_0, \hat{q}_0, q_1 = 1, \hat{q}_1 = \infty, R, \hat{R}, \beta, m) = \frac{1}{m} f_{\mathrm{RS}}(q_0, m^2\hat{q}_0, R, m\hat{R}, \beta m). \tag{8}$$

From this relation, the equations for $q_0, \hat{q}_0, R, \hat{R}$ and m become the coupled equations of the saddle-point equations for the RS solution and the equation of $S_{\mathrm{RS}} = 0$, where S_{RS} is the entropy for the RS solution. Let us denote the solutions of these equations by $q = q_{\mathrm{c}}$, $\hat{q} = \hat{q}_{\mathrm{c}}$, $R = R_{\mathrm{c}}$, $\hat{R} = \hat{R}_{\mathrm{c}}$ and $\beta = \beta_{\mathrm{c}}$. Then, the one-step RSB solutions are expressed by $q_0 = q_{\mathrm{c}}$, $\hat{q}_0 = (\beta/\beta_{\mathrm{c}})^2\hat{q}_{\mathrm{c}}$, $R = R_{\mathrm{c}}$, $\hat{R} = (\beta/\beta_{\mathrm{c}})R_{\mathrm{c}}$ and $m = \beta_{\mathrm{c}}/\beta$. Thus, to obtain the $T \to +0$ limit we only have to know the solutions at $T = T_{\mathrm{c}}$.

Next, we study the asymptotic behaviour. In order to investigate the advance of learning we take the limits $\alpha \to \infty$, $q_{\mathrm{c}} \to 1$ and $R_{\mathrm{c}} \to 1$. As suggested by numerical results, we

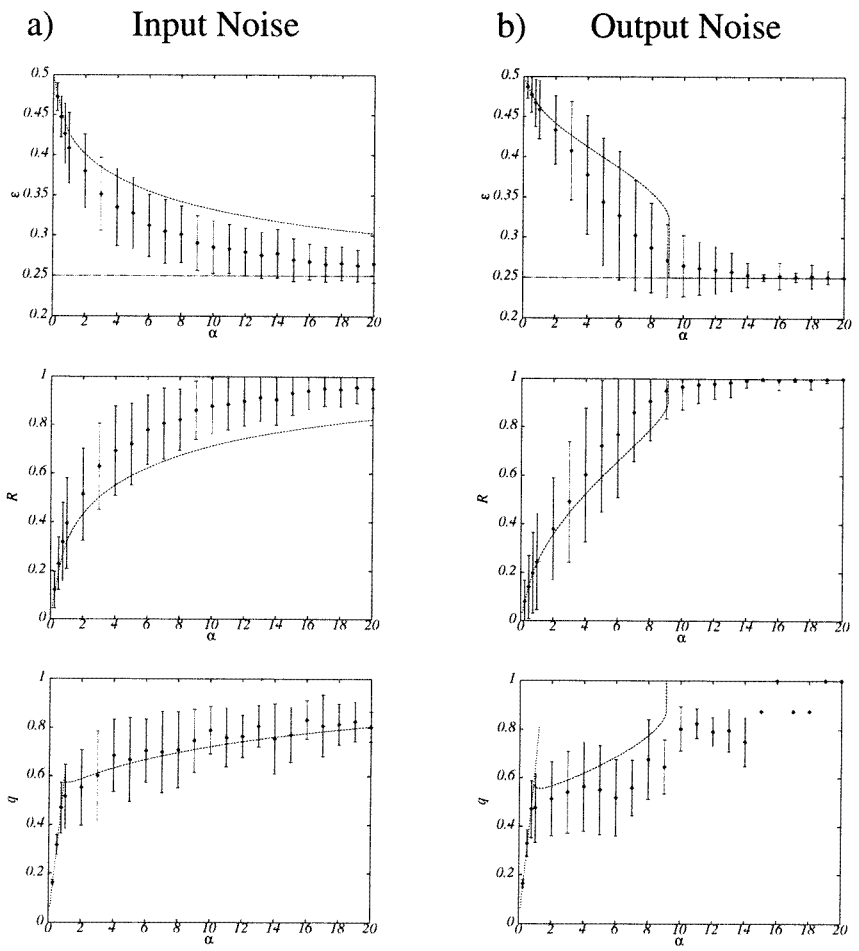

Figure 1. Numerical simulations were performed both for (a) the input noise model and (b) the output noise model with the system size $N = 16$. In both systems, noise levels were set such that $\epsilon_{\min} = \frac{1}{4}$. Abscissas are α and ordinates are ϵ_g, R, and q for RS or q_0 for RSB. Markers and bars represent the averages and the standard deviations obtained from 100 training sets, respectively. Broken curves represent the theoretical prediction (RS or RSB) obtained by the replica method.

consider the limit $\beta_c \ll 1$. Then, f_{RS} and S_{RS} become

$$-\beta_c f_{RS} \simeq -\frac{\hat{q}_c}{2}\Delta q + \hat{R}_c \Delta R + e^{-2(\hat{R}_c - \hat{q}_c)} - \alpha\beta_c \left\{ \epsilon_{\min} + \frac{1}{(1+\delta)\pi} 2^{(1+2\delta)/2} \Gamma\left(\frac{\delta}{2}\right) (\Delta R)^{(1+\delta)/2} - \frac{\beta_c}{2\pi\sqrt{2}}\sqrt{\Delta q} \right\} \tag{9}$$

$$S_{RS} \simeq -\frac{\hat{q}_c}{2}\Delta q + \hat{R}_c \Delta R - \frac{\alpha\beta_c^2}{2\pi\sqrt{2}}\sqrt{\Delta q} + e^{-2(\hat{R}_c - \hat{q}_c)} \tag{10}$$

where $\Delta q = 1 - q_c$ and $\Delta R = 1 - R_c$. Solving the saddle-point equations using the above expressions, we find that the condition $\delta \geqslant \frac{1}{2}$ is necessary for α to increase to ∞. The obtained asymptotic solutions are classified as follows:

- In the case of $\delta > \frac{1}{2}$,

$$\Delta R \propto \left(\frac{\ln\alpha}{\alpha}\right)^{2/(2\delta-1)} \qquad \Delta q \propto \Delta R \qquad \hat{R}_c \propto \ln\left(\frac{\alpha}{\ln\alpha}\right) \qquad \hat{q}_c \propto \hat{R}_c$$
$$\beta_c \propto \left(\frac{\ln\alpha}{\alpha}\right)^{\delta/(2\delta-1)}. \tag{11}$$

- In the case of $\delta = \frac{1}{2}$,

$$\Delta R \propto e^{-4F_0\alpha} \qquad \Delta q \propto \Delta R \qquad \hat{R}_c \sim F_0\alpha \qquad \hat{q}_c \propto \hat{R}_c \qquad \beta_c \propto e^{-F_0\alpha} \tag{12}$$

where F_0 is a constant.

- In the case of $\delta < \frac{1}{2}$, perfect learning takes place, i.e.

$$\Delta R = 0 \qquad \Delta q = 0 \tag{13}$$

at finite α.

Putting together all the results obtained above and using relation (5), we come to the statements (1)–(3). From our results, we note the following. For the input noise model studied by Györgyi and Tishby [3] ($\delta = 1$), perfect learning does not exist and the asymptotic behaviour is $\Delta\epsilon_g \sim \alpha^{-2}$. For the output noise model studied by Opper and Haussler [4] ($\delta = 0$), a first-order transition from the RSB solution to perfect learning takes place. To check these results we performed numerical simulations for $\delta = 0$ and $\delta = 1$. We adopted the exhaustive method, considering all the 2^N weight vectors. In the simulations of networks with Ising weights, usually there exist finite size effects and it is difficult to find a proper finite size scaling [14]. The situation is the same in our model. However, as shown in figure 1, qualitatively the numerical results agree with the theoretical results.

In conclusion, the statements (a)–(c) obtained by Seung using a refined annealled approximation agree with our statements (1)–(3) obtained by the replica method under the one-step RSB ansatz. The conclusions support the validity of both a refined annealled approximation and the one-step RSB ansatz.

The authors are grateful to P Davis for valuable discussions. One of the authors (YK) was partially supported by the Japanese Grant-in-Aid for Science Research Fund from the Ministry of Education, Science and Culture No 06260102 and No 06740325.

References

[1] See, for example, Watkins T L H, Rau A and Biehl M 1993 *Rev. Mod. Phys.* **65** 499
[2] Seung H S, Sompolinsky H and Tishby N 1992 *Phys. Rev.* A **45** 6056
[3] Györgyi G and Tishby N 1990 *Neural Networks and Spin Glasses* ed W K Theumann and R Köberle (Singapore: World Scientific) p 3
[4] Opper M and Haussler D 1991 *Phys. Rev. Lett.* **20** 2677
[5] Kabashima Y and Shinomoto S 1992 *Neural Comput.* **4** 712
[6] Kim J and Pollard D 1990 *Ann. Stat.* **18** 191
[7] Uezu T, Kabashima Y, Nokura K and Nakamura N 1995 *Stat. Phys.* **19** (Abstracts) 62
Uezu T and Kabashima Y 1996 *J. Phys. A: Math. Gen.* **29** L55–L60
Uezu T, Kabashima Y, Nokura K and Nakamura N 1996 *J. Phys. Soc. Japan* to appear
[8] Amari S, Fujita N and Shinomoto S 1992 *Neural Comput.* **4** 605
[9] Györgyi G 1990 *Phys. Rev.* A **41** 7097
[10] Sompolinsky H, Tishby N and Seung H S 1990 *Phys. Rev. Lett.* **65** 1683
[11] Seung H S 1995 *Proc. of the CTP-PBSRI Joint Workshop on Theoretical Physics* ed Jong-Hoon Oh *et al* (Singapore: World Scientific) p 32

[12] Parisi G 1980 *J. Phys. A: Math. Gen.* **13** 1101; 1980 *J. Phys. A: Math. Gen.* **13** L115; 1980 *J. Phys. A: Math. Gen.* **13** 1887
[13] Krauth W and Mézard M 1989 *J. Physique* **50** 3057
[14] Derrida B, Griffiths R B and Bennett A P 1991 *J. Phys. A: Math. Gen.* **24** 4907