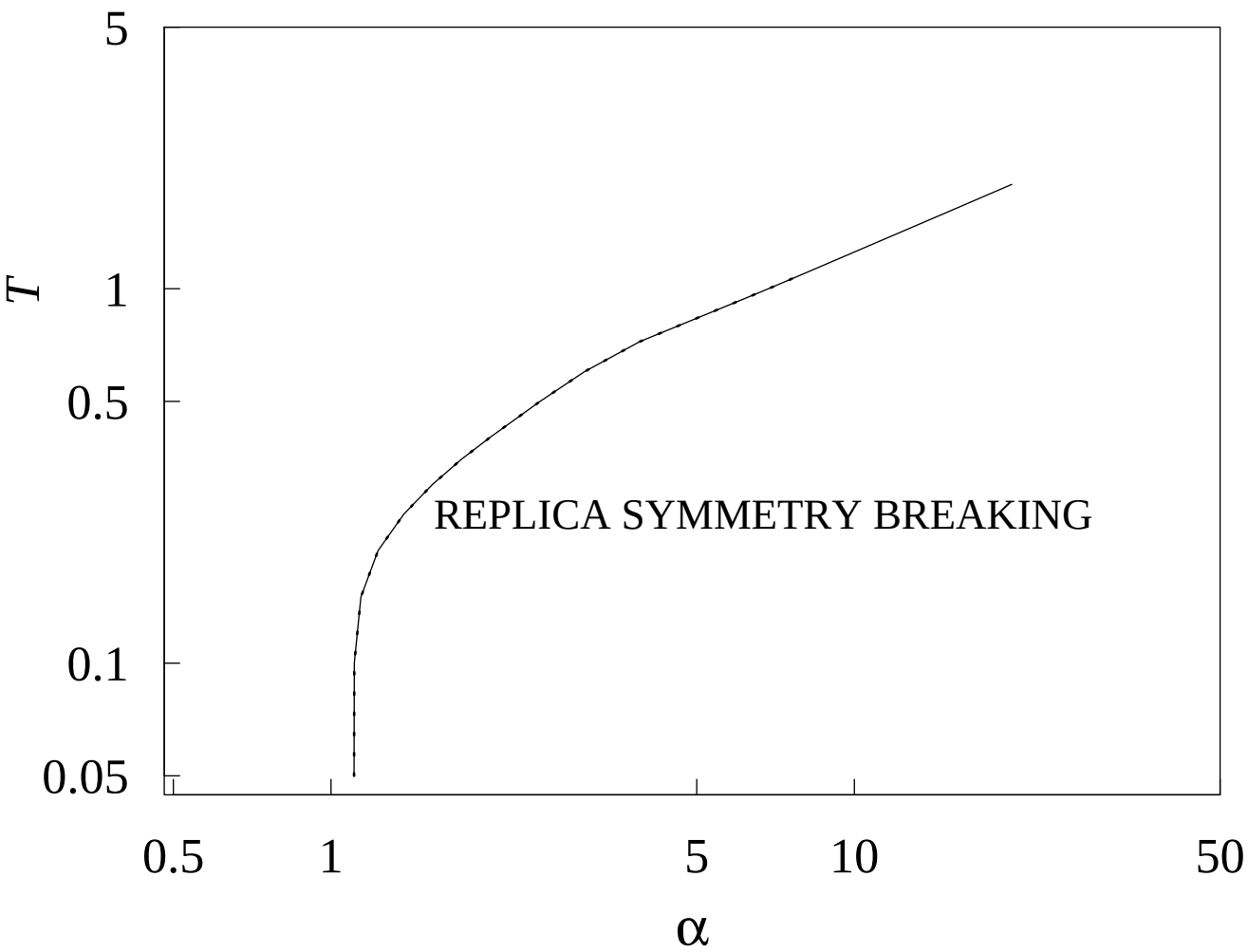
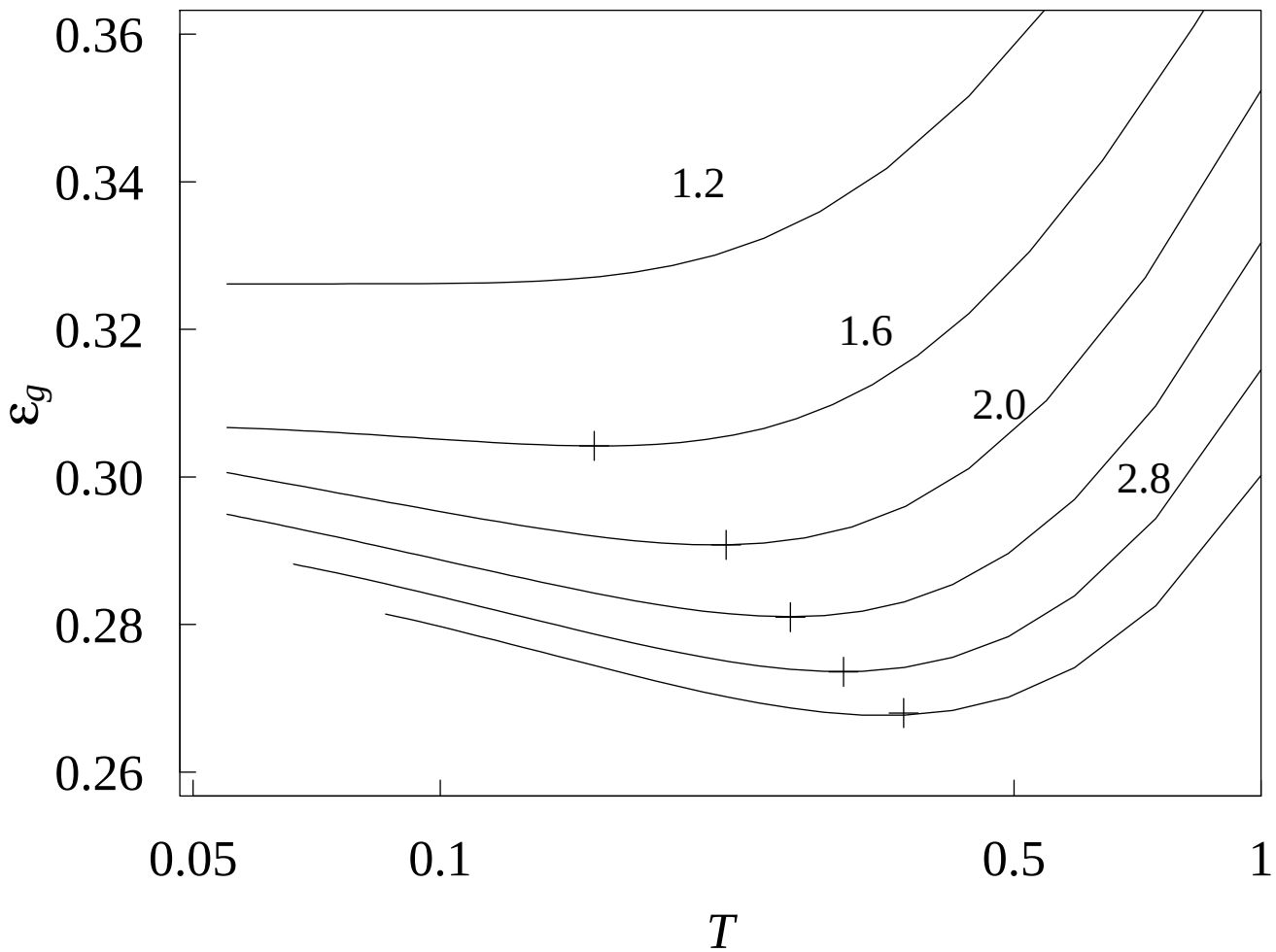
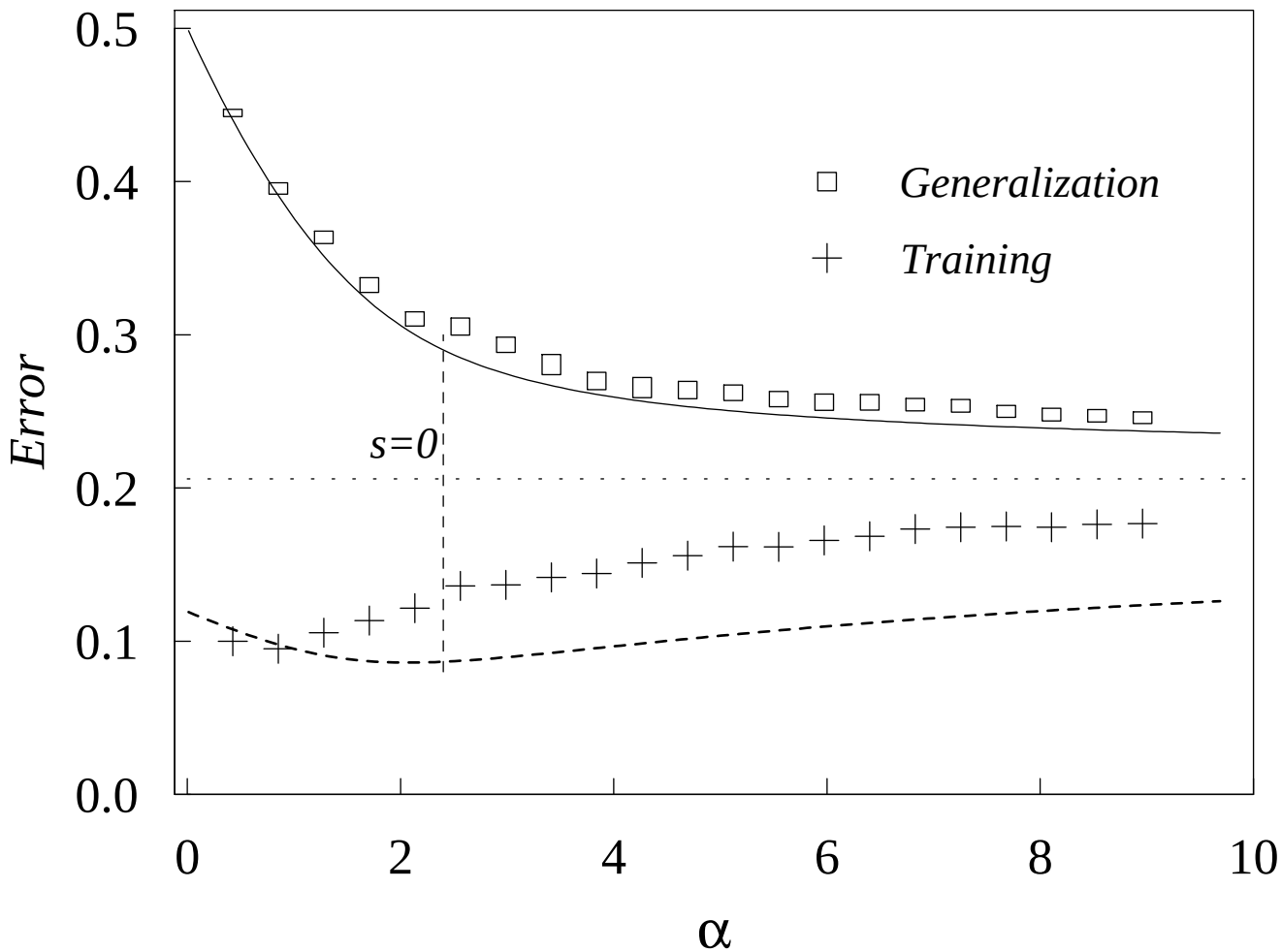
Part II: Figure 12, Zero Entropy



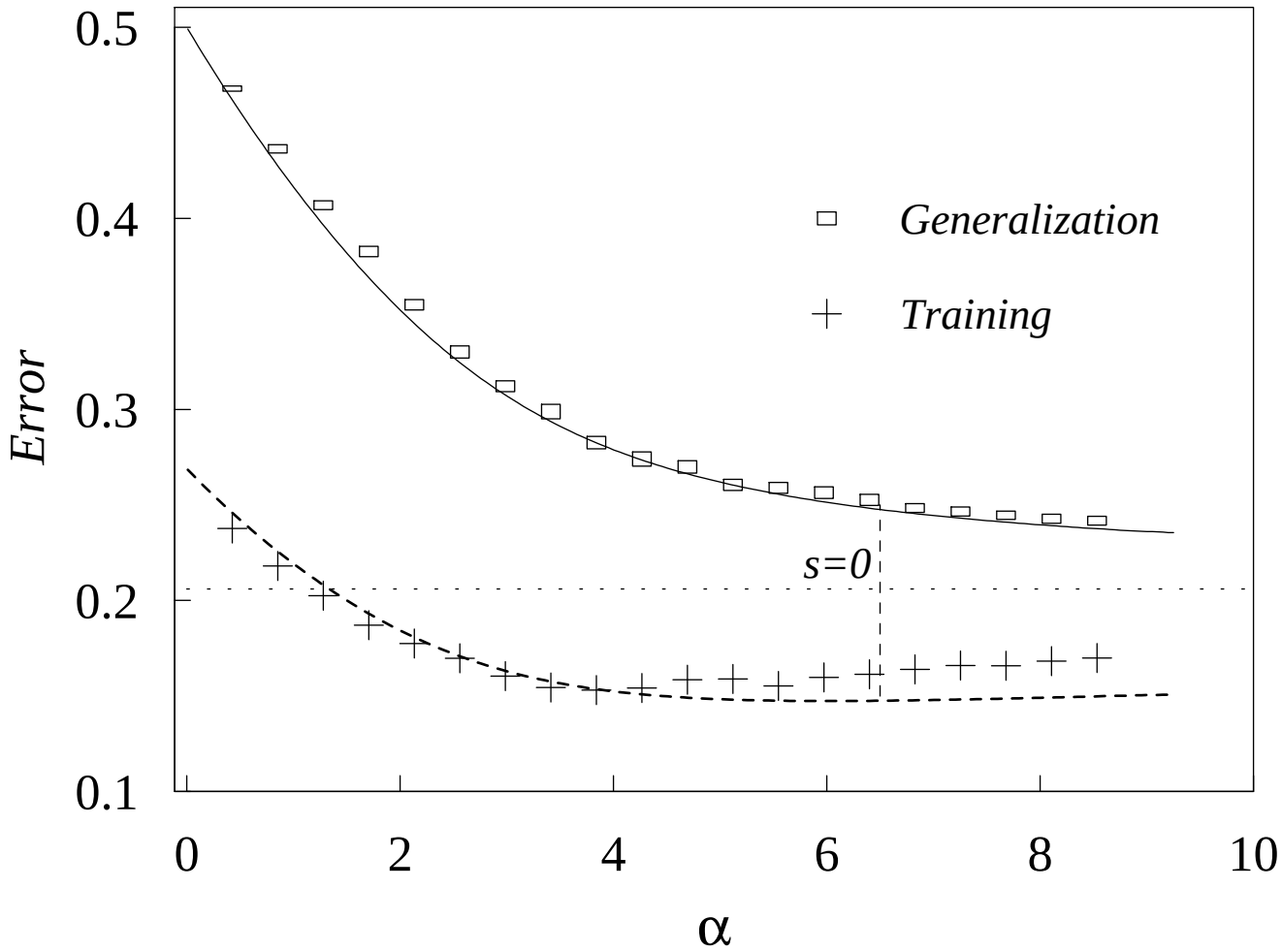
Part II: Figure 11, Optimal T



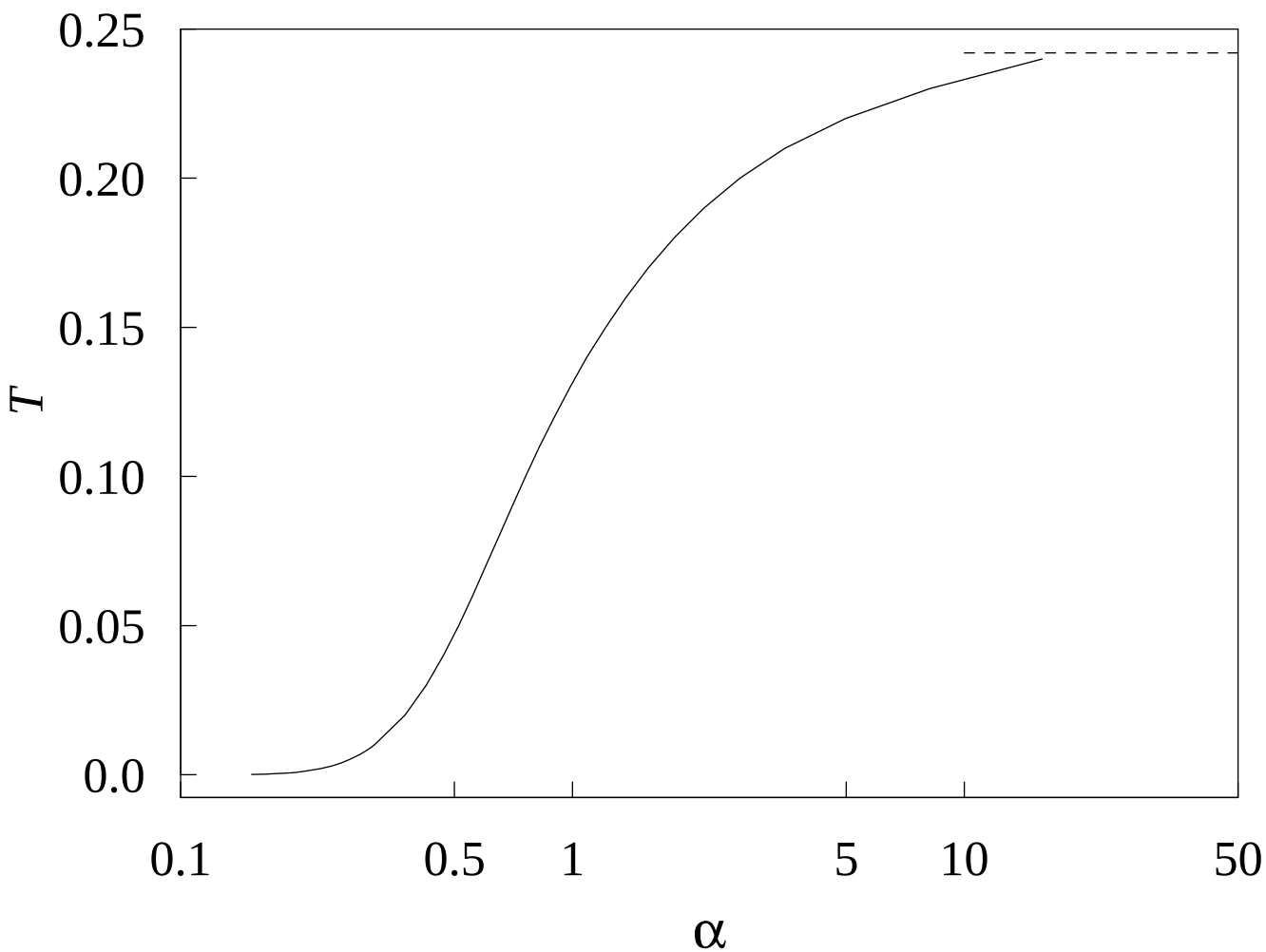
Part II: Figure 10b, T=0.5



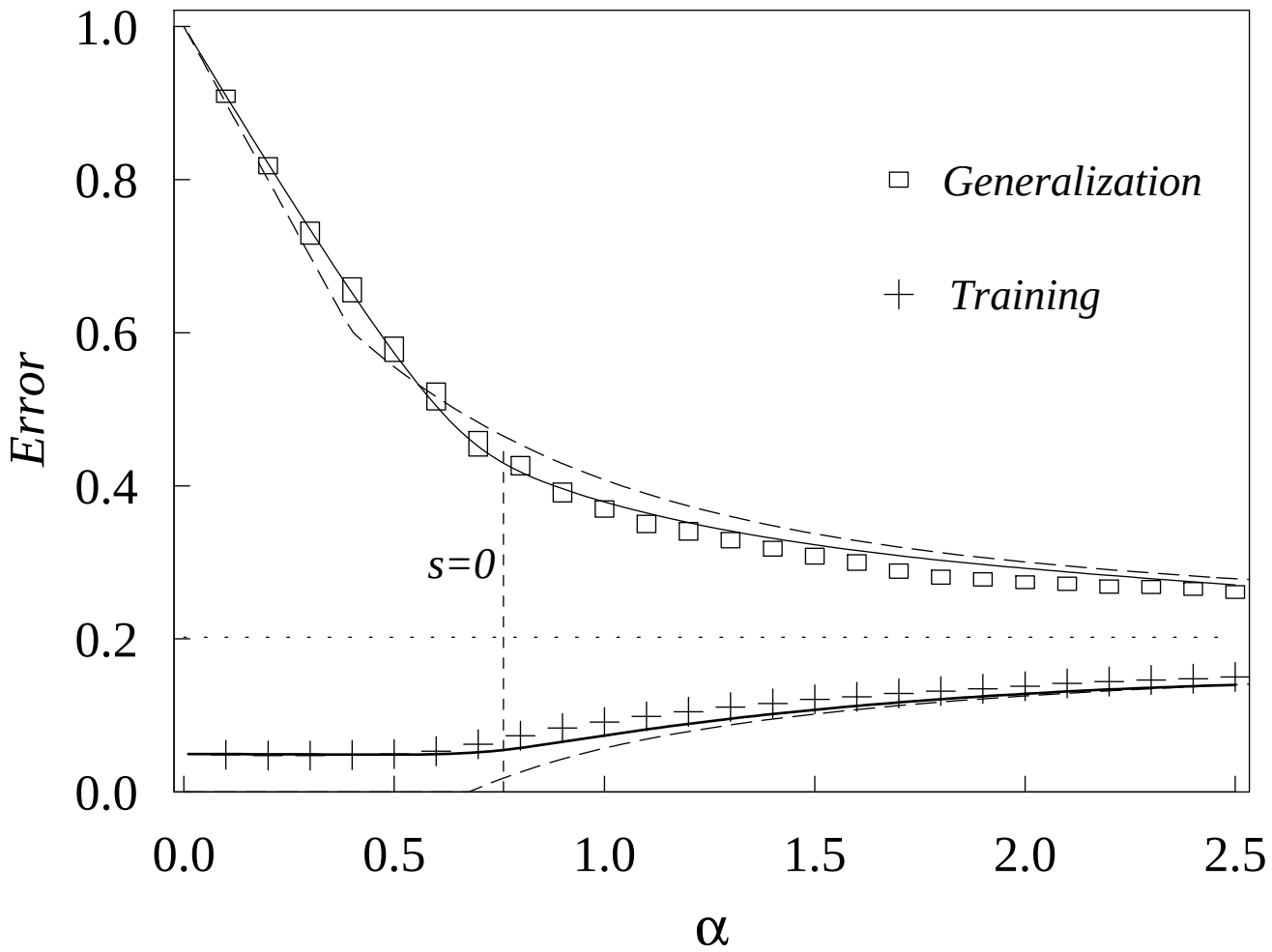
Part II: Figure 10a, T=1.0



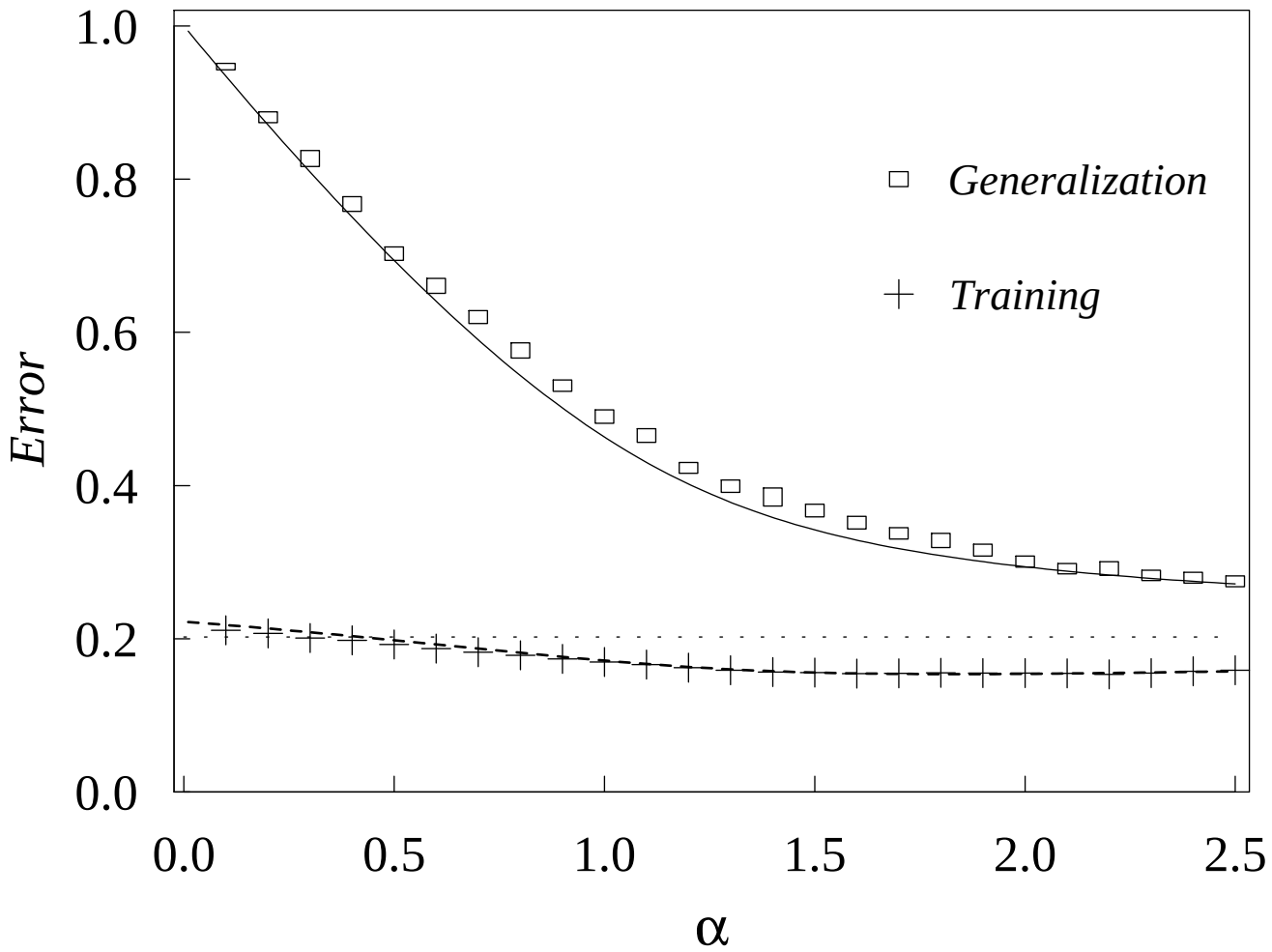
Part II: Figure 9, RS Zero Entropy



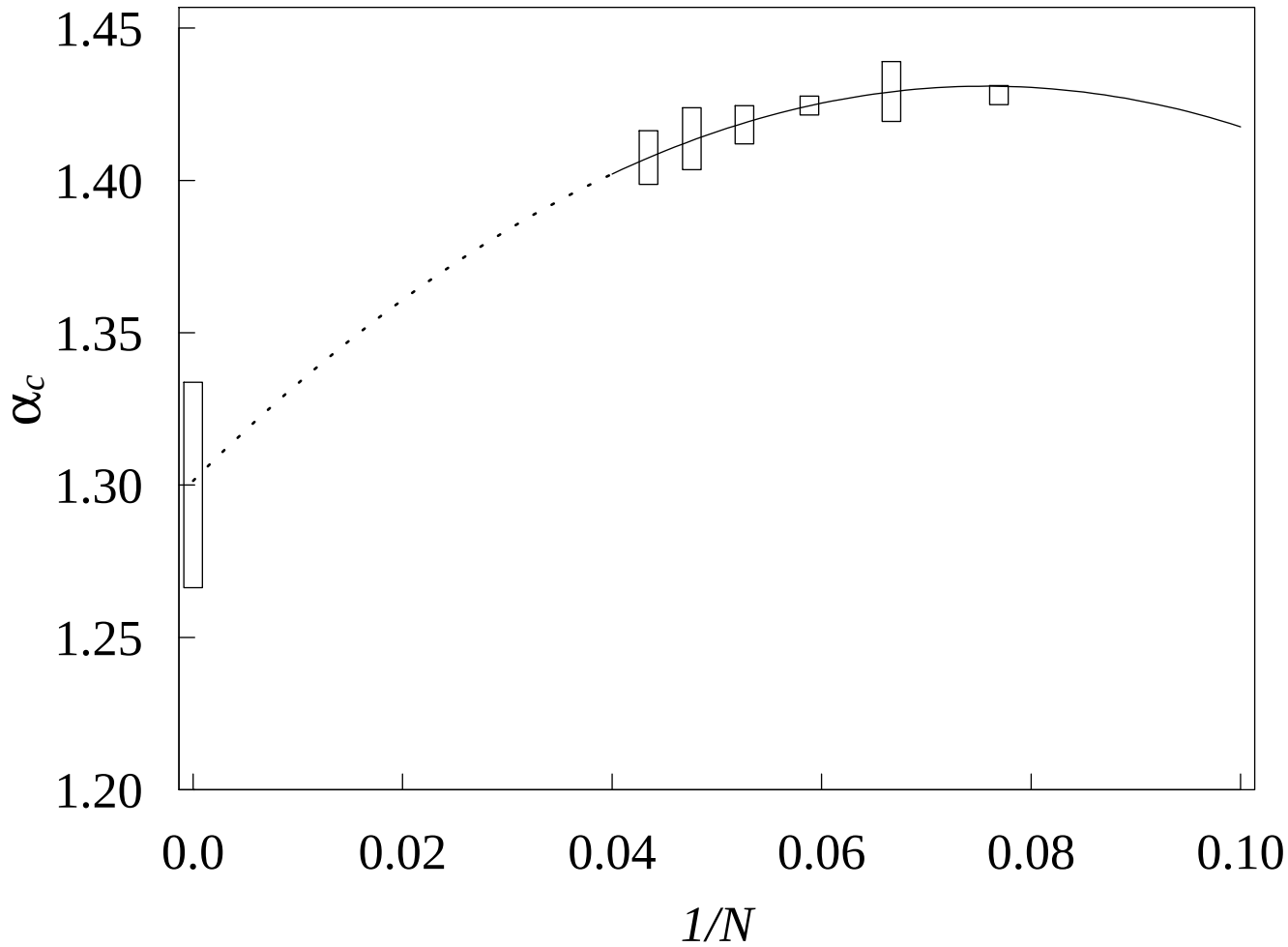
Part II: Figure 8b, T=0.1



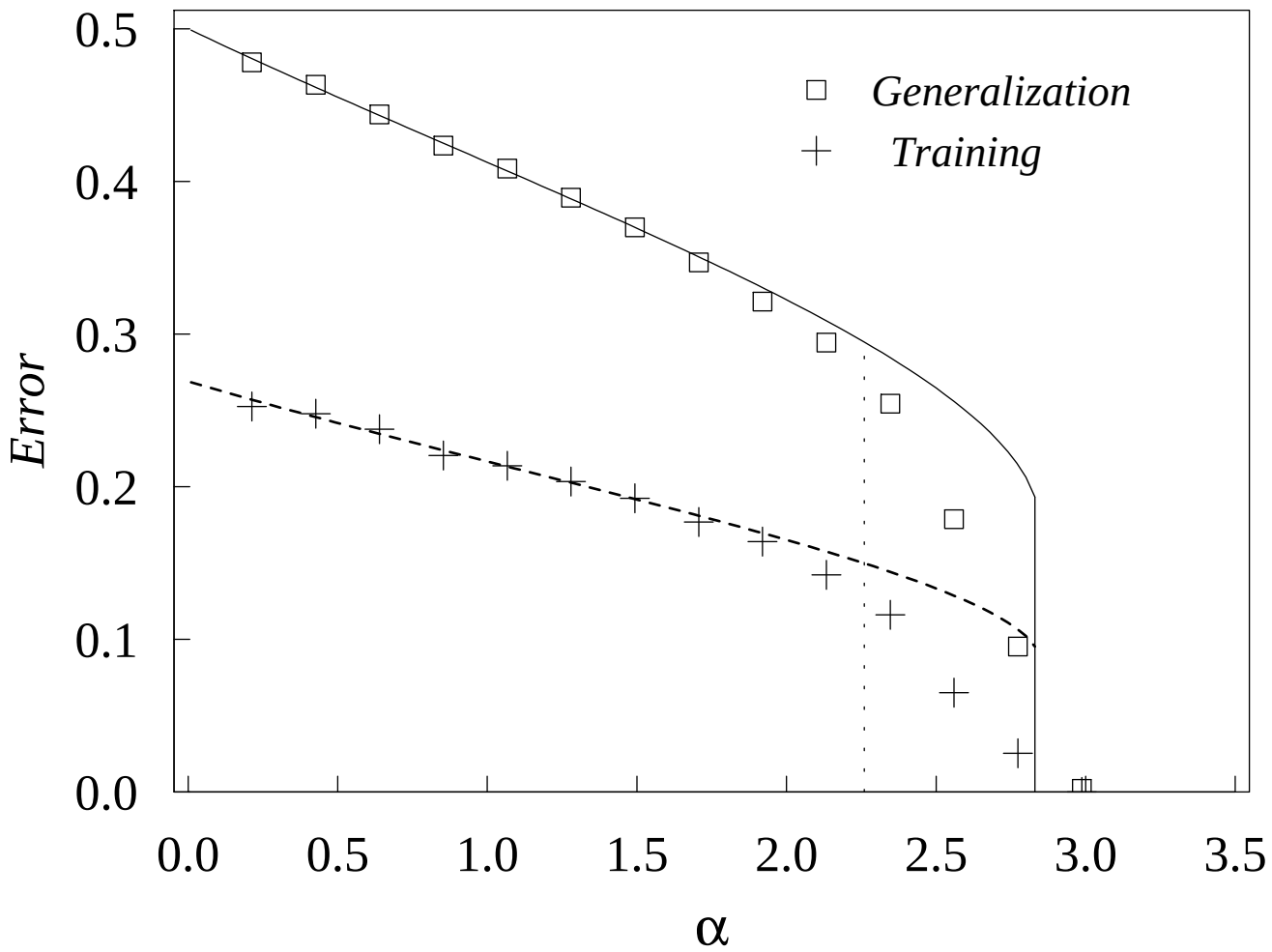
Part II: Figure 8a, T=0.5



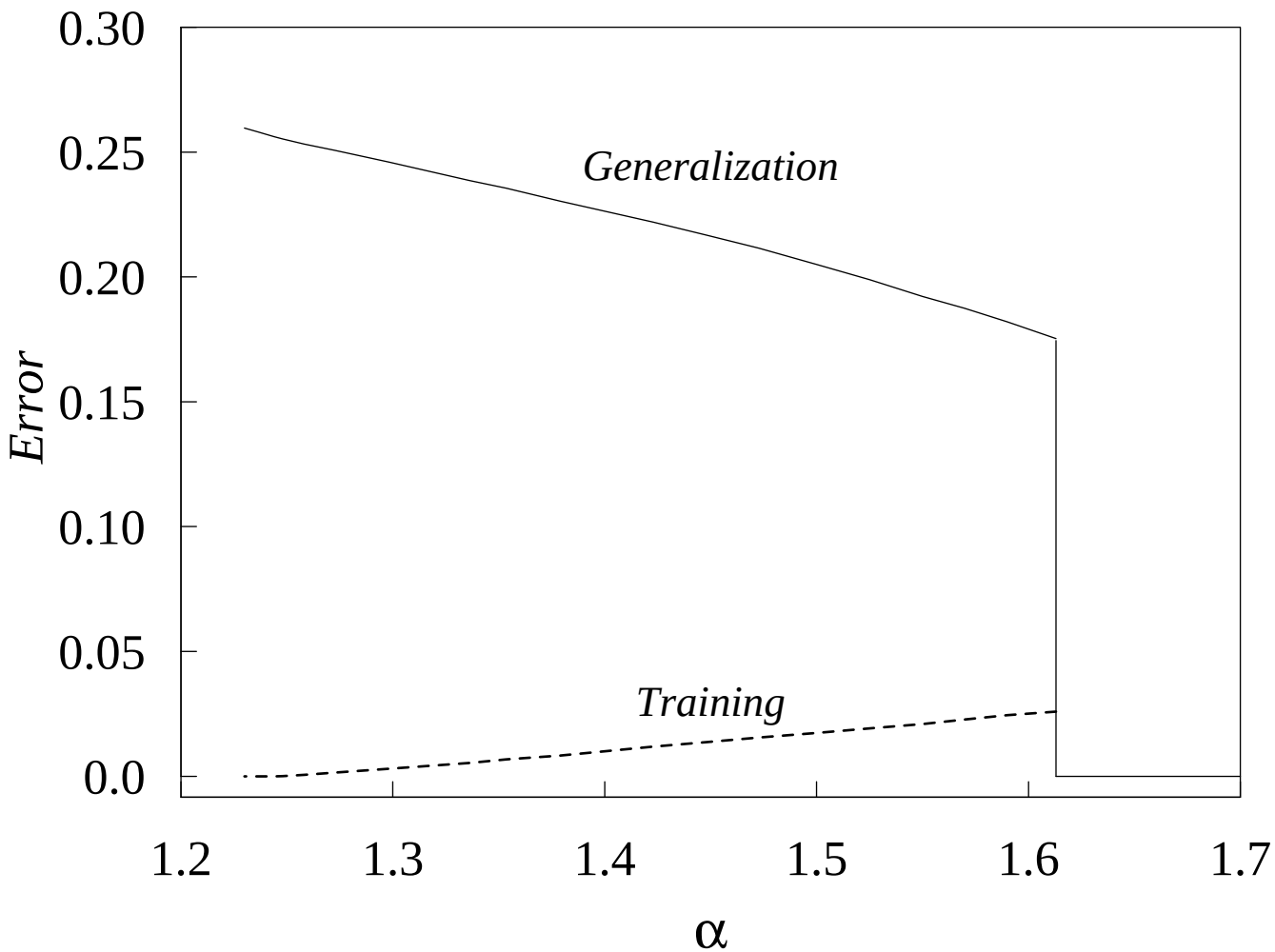
Part II: Figure 7, Exhaustive



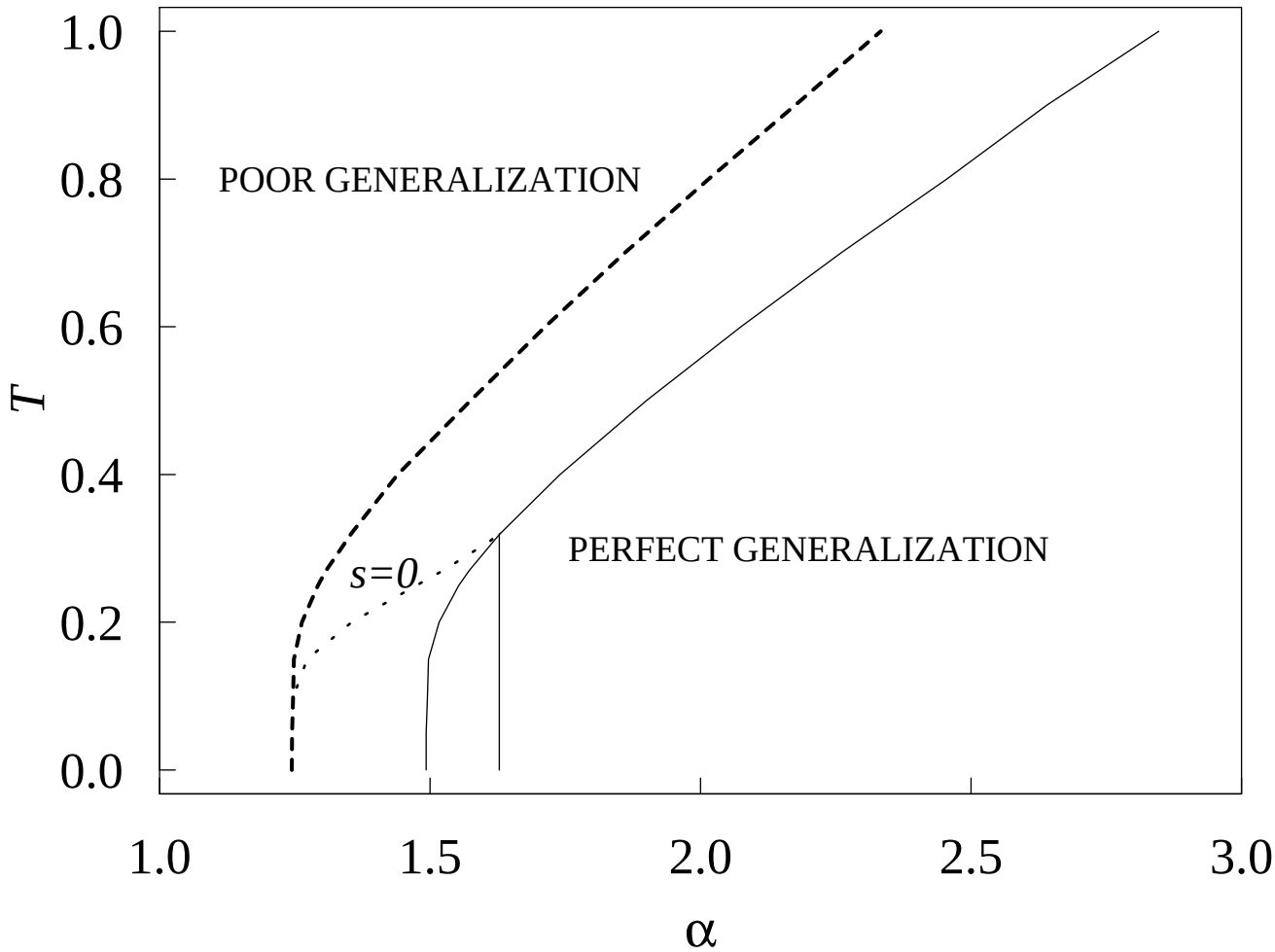
Part II: Figure 6, T=1.0, N=75



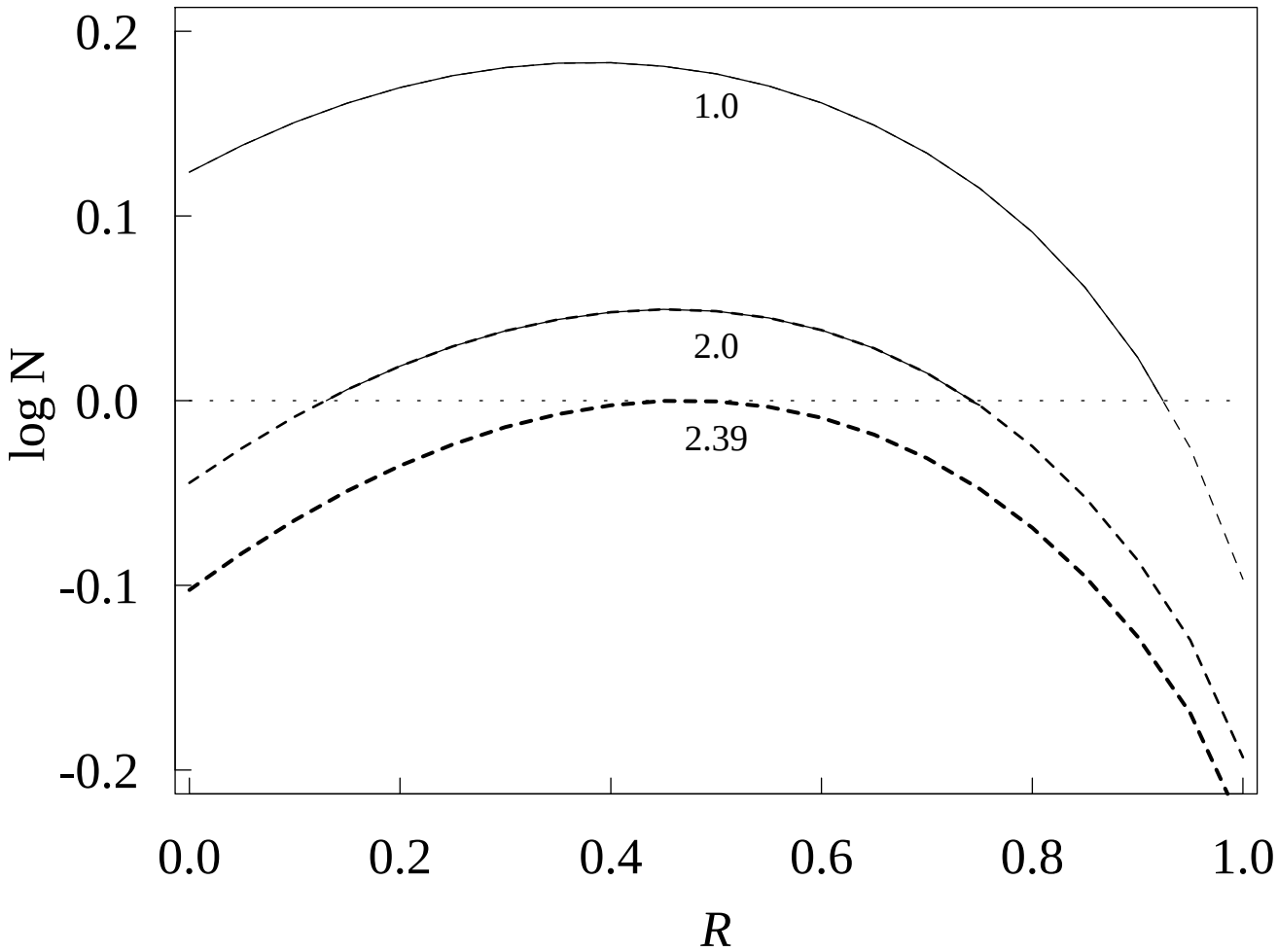
Part II: Figure 5, RSB errors



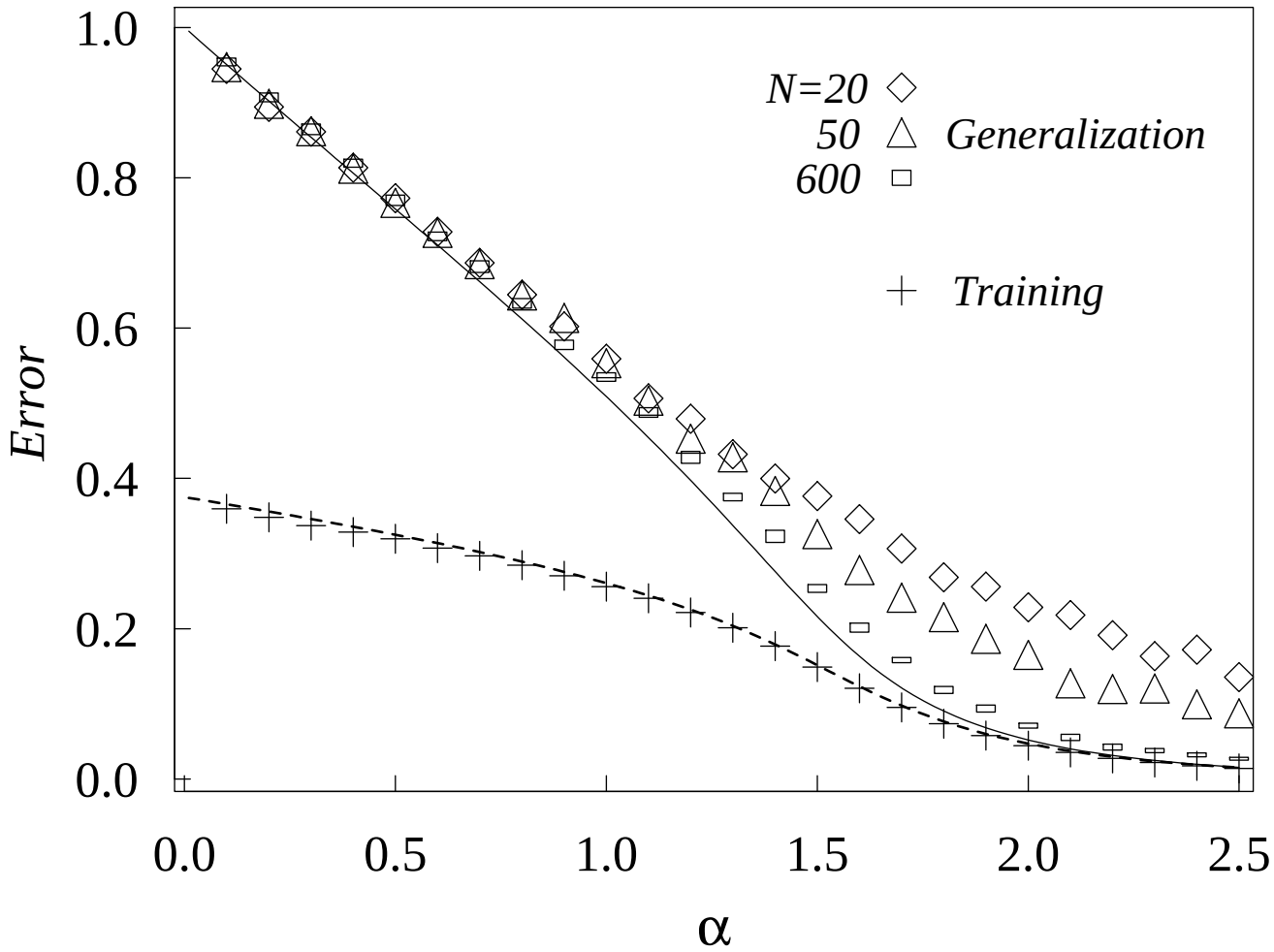
Part II: Figure 4



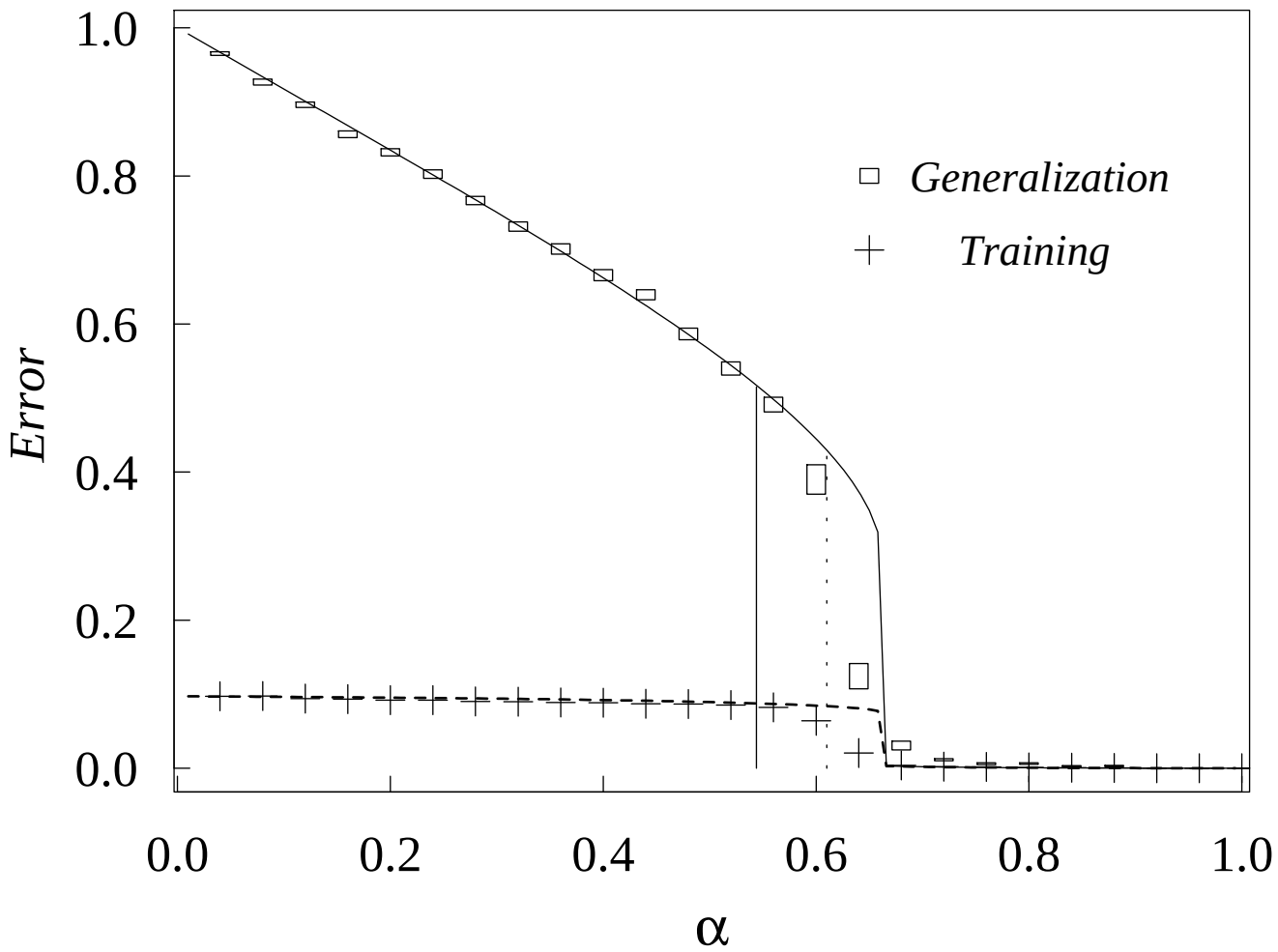
Part II: Figure 3



Part II: Figure 2b, T=1



Part II: Figure 2a, T=0.2



Part II: Figure 1

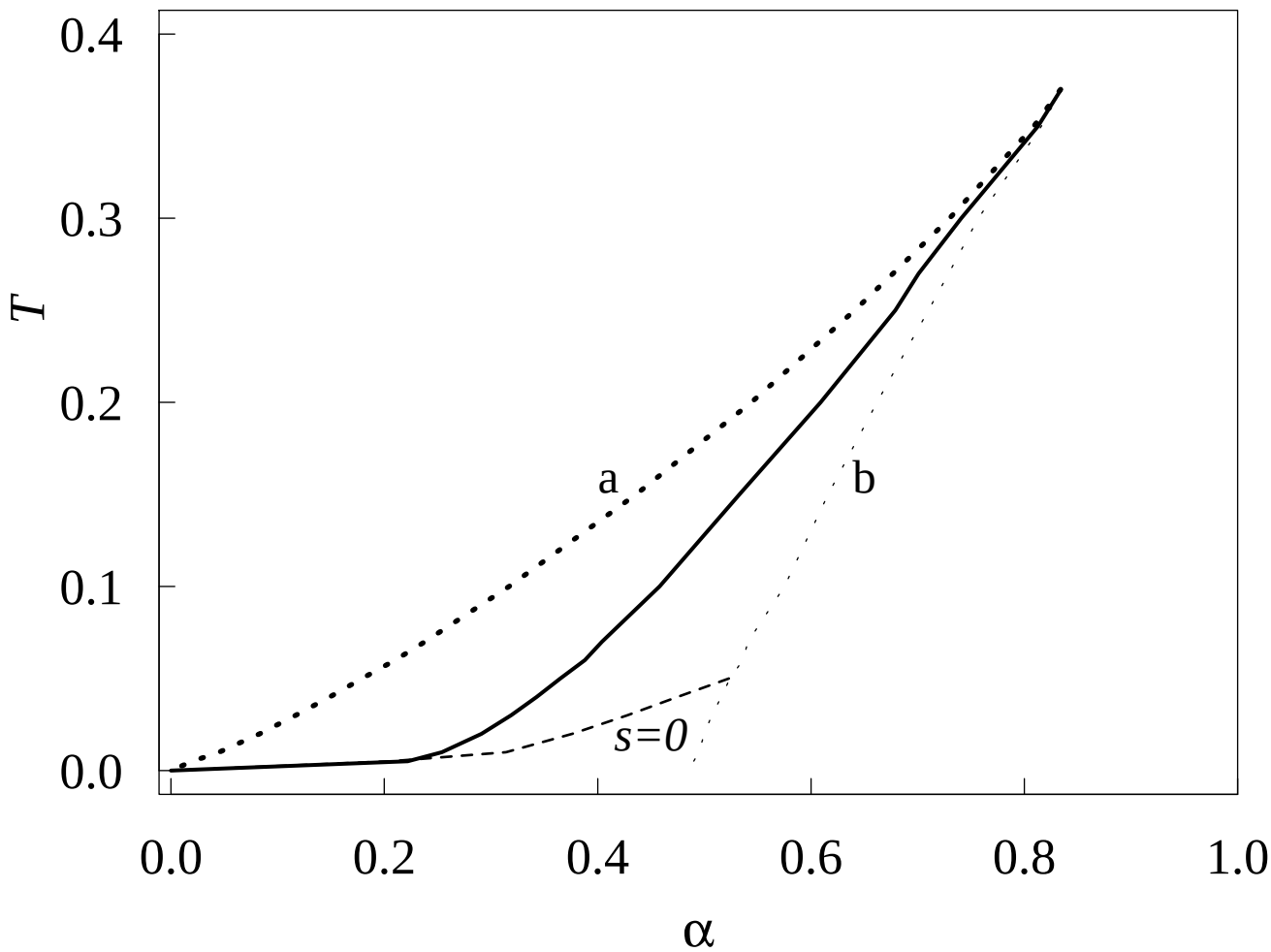


FIG. 6. Learning curves for the boolean/discrete model. Monte Carlo simulations at $T = 1.0$ and $N = 75, 32$ samples. As in the linear/discrete model, there are deviations from the theory between the two spinodals (solid vertical lines), because of finite size effects. The thermodynamic transition is marked by the dashed vertical line.

FIG. 7. Graph of α_c versus $1/N$ at $T = 0$ for the boolean/discrete model from exhaustive search. Each data point is the average of roughly 10^5 or 10^6 random samples of examples. The error bars are the standard error of measurement over the sample distribution. A least squares fit to a quadratic function yields the extrapolation $\alpha_c(N = \infty) = 1.30 \pm 0.03$. The replica symmetric prediction $\alpha_c = 1.24$ is also marked.

FIG. 8. Learning curves for the linear/mismatched model: (a) Monte Carlo simulations at $T = 0.5$ with $N = 100$ and 64 samples, with lines from RS theory. (b) Simulations at $T = 0.1$. Note that the RS curves (solid lines) fit the data even in the RSB regime. The zero entropy point $\alpha = 0.76$ is marked. The dashed line is the $T = 0$ RS curve, shown for comparison.

FIG. 9. The RS zero entropy line for the linear/mismatched model, which is a lower bound for the temperature at which RSB occurs. The line approaches the limit $T = 0.242$ as $\alpha \to \infty$.

FIG. 10. Learning curves for the boolean/mismatched model: (a) Simulations at $T = 1.0$ with $N = 100$ and 32 samples. The fit with RS theory is good until $\alpha \approx 4$, where the training error starts to deviate. The RS entropy is zero at $\alpha = 6.4$. (b) At $T = 0.5$ the deviation between simulation and RS theory is more severe, although the generalization error is still not much affected.

FIG. 11. Generalization error for fixed α as a function of T for the boolean/mismatched model. The minimum of each of these curves gives the optimal generalization temperature for that α.

FIG. 12. RS zero entropy line for the boolean/mismatched model. Note that the line intersects the α axis at the finite value $\alpha = 1.1$.

FIGURES

FIG. 1. Phase diagram of the linear/discrete model from RS theory. In the region between the two dashed spinodal lines marked (a) and (b), there are two locally stable solutions to the RS saddle point equations. The solid line is the thermodynamic transition, where the free energies of the solutions are equal. The line marked $s = 0$ is the zero entropy line, below which the RS ansatz for the metastable state is incorrect.

FIG. 2. Learning curves for the linear/discrete model. (a) Monte Carlo simulations at $T = 0.2$ for $N = 100$, 64 samples. The simulations deviate from the RS theory between the two spinodals (solid vertical lines), where some mixing occurs because of finite size effects. The thermodynamic transition is marked by the dashed vertical line. (b) At $T = 1$ there is no transition. Deviations from the theory appear to be finite size effects, since they diminish with increasing N.

FIG. 3. Annealed upper bound for the logarithm of the density of local minima in the linear/discrete model as a function of R. For $\alpha = 1.0$ there is a small gap around $R = 1.0$. As α increases, the density of local minima decreases, until above $\alpha = 2.39$ there are no local minima at all, except the isolated minimum at $R = 1$.

FIG. 4. Phase diagram for the boolean/discrete model. The lines marked (a) and (b) are the transition lines predicted by replica symmetric theory. To the left of the thermodynamic transition line (a), the equilibrium state has $R < 1$ (poor generalization), while the perfect generalization state ($R = 1$) is metastable. Between lines (a) and (b), the $R = 1$ state is the equilibrium one, while the $R < 1$ state is metastable. To the right of the spinodal line (b), the $R < 1$ state vanishes. According to the one-step RSB theory, the metastable $R < 1$ state is a spin-glass phase at low T, with phase boundaries given by the zero entropy line $T_g(\alpha)$ and the line marked (c).

FIG. 5. Training and generalization errors of metastable spin glass phase in the boolean/discrete model, according to the one–step RSB ansatz. In this phase, the errors are independent of temperature, and are given by their RS values on the zero entropy line $T_g(\alpha)$ of Fig. 4.

If one proceeds to expand the A_i, one obtains a double expansion about $(1/T, T/\alpha) = (0,0)$

$$A(1/T, T/\alpha) = \sum_{a,b} A_{ab} \left(\frac{1}{T}\right)^a \left(\frac{T}{\alpha}\right)^b \quad . \tag{C2}$$

This high-T expansion is the proper tool for investigating power laws of the form $T \sim \alpha^r$. We must only assume that $0 < r < 1$, so that both $1/T$ and T/α approach zero as $\alpha \to \infty$.

Assuming now that we have an expression for ϵ_g of the form (C2), let us find the power law $T_{opt} \sim \alpha^r$ such that ϵ_g decreases at the fastest rate. In the series (C2) each term a, b scales like $1/\alpha^{ra+(1-r)b}$, so that the exponent is some weighted average of a and b. The power of the dominant term in the sum is $\min_i\{ra_i + (1-r)b_i\}$. This exponent must be maximized, to insure the fastest decreases of ϵ_g.

The problem thus reduces to linear programming. Given a set of pairs (a_i, b_i), find the r that maximizes $\min_i\{ra_i + (1-r)b_i\}$. The problem looks difficult because there are an infinite number of pairs (a_i, b_i) to consider, but in fact most of the pairs can be eliminated from consideration. We define a partial ordering of the set of pairs: $(a_i, b_i) \leq (a_j, b_j)$ means $a_i \leq a_j$ and $b_i \leq b_j$. A term (a, b) is *minimal* if $(a, b) \geq (c, d)$ implies $(a, b) = (c, d)$ for all (c, d). Only the subset of minimal pairs (which cannot be ordered) are relevant, because $(a_i, b_i) \leq (a_j, b_j)$ implies $ra_i + (1-r)b_i \leq ra_j + (1-r)b_j$. The linear programming problem thus only includes the finite minimal subset of pairs (a_i, b_i).

For the boolean/mismatched model, we calculated the asymptotics of the generalization error as a series in $\alpha^{-\frac{1}{2}}$ at fixed T,

$$\epsilon_g(T, \alpha) - \epsilon_{\min} = \sum_{j=1}^{\infty} e_{j/2}(\beta)\alpha^{-j/2} \quad . \tag{C3}$$

To convert this to the form (C2) of the high-T expansion, we expand the e_j in powers of β. Only the leading term from each e_j need be retained, since the higher order terms are irrelevant. From the terms of order $\alpha^{-1/2}$ through α^{-2} we obtain the $(1, 1/2)$, $(1, 1)$, $(1, 3/2)$, and $(0, 2)$ terms in the high-T expansion (C2). From these four terms, we finally extract the minimal subset $(1, 1/2)$ and $(0, 2)$. Terms in (C3) of order $\alpha^{-5/2}$ and higher need not be considered, since they are all bounded below by $(0, 2)$ and are thus irrelevant.

The maximin problem, which involves only $(1, 1/2)$ and $(0, 2)$, has $r = 3/5$ as its solution, justifying the result $T_{opt} \sim \alpha^{3/5}$ quoted in Eq. (6.50). The calculation of the zero entropy line is similar, except that the minimal terms must add up to zero.

$$m(T,\alpha) = T/T_g(\alpha) \ , \tag{B13}$$

where $T_g(\alpha)$ is the $s = 0$ line of the RS solution.

The physical meaning of the above solution is that for α in the regime

$$\begin{aligned} 1.24 < \alpha \leq 1.49 \ , \\ 0 \leq T \leq T_g(\alpha) \ , \end{aligned} \tag{B14}$$

there is a metastable SG phase in which thermal fluctuations are completely frozen and hence entropy is zero. For fixed α, the energy, i.e. the training error $\epsilon_t(T,\alpha)$, is independent of T. Throughout the SG phase, it retains its value on the boundary $T_g(\alpha)$. Similarly $R(T,\alpha)$ and $q(T,\alpha)$ are frozen at their values on $T_g(\alpha)$. Thus the values $\epsilon_t(T_g(\alpha))$, $q_0(T_g(\alpha),\alpha)$, and $R(T_g(\alpha),\alpha)$ can be calculated using the RS theory, since it is correct on this boundary. The results are shown in Fig. 5.

It should be emphasized that the above one-step RSB solution exists only for $1.24 < \alpha$ and is metastable (i.e. $\epsilon_t > 0$). Thus it predicts that the critical α beyond which there is no solution with zero training error other than $\mathbf{W} = \mathbf{W}_0$ is

$$\alpha_c = 1.24 \tag{B15}$$

which is the point where the entropy of the RS solution (at $T = 0$) vanishes. The complete phase diagram is shown in Fig. 4. We have not attempted to search for other RSB solutions or to check the stability of this solution. However, this solution is probably exact both here and in the random perceptron problem of Krauth and Mézard. Similar completely frozen SG phases are known to exist in the REM, the "simplest spin glass"[27], and the large-P Potts glass. Recently, they have also been found in learning of random mappings by two-layer networks[30, 31].

APPENDIX C: POWER COUNTING IN THE HIGH TEMPERATURE EXPANSION

For the boolean/mismatched model, the optimal generalization and zero entropy lines both follow power laws of the form $T \sim \alpha^r$, with $0 < r < 1$. The determination of this exponent requires the balancing of the two dominant terms in the asymptotic expansion. This is somewhat tricky, because determining which are the two dominant terms in turn depends on the exponent.

Since both T and α are diverging, it would seem natural to perform a double expansion around $(1/T, 1/\alpha) = (0,0)$. Such an expansion is in fact ill-defined, since the existence of a nontrivial high-temperature limit shows that the $T,\alpha \to \infty$ limit depends on the ratio T/α.

As discussed in Appendix B of Part I, the proper high temperature expansion for a quantity A that is finite in the high-T limit is

$$A(\beta, T/\alpha) = \sum_{i=0}^{\infty} \beta^i A_i(T/\alpha) \ . \tag{C1}$$

$$u \equiv t_0 \sqrt{\frac{q_0 - R^2}{1 - q_1}} + t_1 \sqrt{\frac{q_1 - q_0}{1 - q_1}} - \frac{y R}{\sqrt{1 - q_1}} \ . \tag{B4}$$

The free energy has to be minimized with respect to q_0, q_1, $\hat{q}_0$, $\hat{q}_1$, R, $\hat{R}$, and m. Note that after the $n \to 0$ limit has been taken the allowed range of m is $0 \leq m \leq 1$. Also, the physical meaning of q_0, q_1, and m is explained in Sec. V.D.2 (see Eqs. 5.52–5.54). Upon substituting $m = 0$, the free energy f reduces to the RS result, Eq. (5.45).

We now search for a solution to the saddle point equations at finite temperature with the property

$$q_1 = 1, \qquad \hat{q}_1 = \infty \ . \tag{B5}$$

According to Eq. (5.52), for such a solution each pure phase consists of a single (or a few almost identical) microscopic state. Thus the system collapses at finite temperatures to phases with zero entropy. This is not unlike the collapse of the system to the ground state at higher values of α (i.e. the perfect generalization state). However, it occurs in the metastable, spin glass regime.

To find a solution with the property (B5) we take the limit $q_1 \to 1$, $\hat{q}_1 \to \infty$ of Eqs. (B3)–(B4) while keeping β finite. We obtain

$$G_0 = \frac{1}{m}[\hat{q}_0(q_0 - 1) - R\hat{R}] + \frac{1}{m}\int_{-\infty}^{\infty} Dz \ln 2 \cosh\left(z\sqrt{\hat{q}_0} + \hat{R}\right) \ , \tag{B6}$$

$$G_r = -\frac{2}{m}\int_0^{\infty} Dy \int_{-\infty}^{\infty} Dt \ln\left[e^{-\beta m} + (1 - e^{-\beta m}) H(u)\right] \ , \tag{B7}$$

$$u \equiv t\sqrt{\frac{q_0 - R^2}{1 - q_0}} - \frac{y R}{\sqrt{1 - q_0}} \ , \tag{B8}$$

where $\hat{q}_0$ and $\hat{R}_0$ are the original order parameters multiplied by m^2 and m respectively. Comparing Eqs. (B6)–(B8) with Eq. (5.45) one obtains

$$f_{\rm RSB}(q_0, \hat{q}_0, R, \hat{R}, m, \beta) = \frac{1}{m} f_{\rm RS}(q_0, \hat{q}_0, R, \hat{R}, \beta, m) \ . \tag{B9}$$

This structure is similar to the result of Krauth and Mézard[26] for the case of training a perceptron with random input-output mappings.

Stationarity with respect to q, $\hat{q}$, R, $\hat{R}$ results in

$$q_0(T, m, \alpha) = q_{\rm RS}(T/m, \alpha) \ , \tag{B10}$$

$$R(T, \alpha, m) = R_{\rm RS}(T/m, \alpha) \ . \tag{B11}$$

Finally, stationarity with respect to m yields

$$s_{\rm RS}(T/m, \alpha) = 0 \ , \tag{B12}$$

which implies

APPENDIX B: ONE STEP RSB SOLUTION OF THE BOOLEAN/DISCRETE MODEL

In Section V.D we have found that the replica symmetric (RS) solution of the boolean/discrete perceptron model yields an unphysical negative entropy in a restricted range of α at low T. The line in the T-α plane where the RS "poor generalization" state has zero entropy is marked $T_g(\alpha)$ in Fig. 4. The entropy of this state is negative for all T and α between this line and the spinodal line, where the state vanishes. It should be noted that the "$s = 0$" line is to the right of the thermodynamic transition line and coincides with it only at $T = 0$ (at $\alpha = 1.24$). This implies that the RS theory is invalid for this metastable phase only. To find the correct metastable state (with poor generalization) at low T we must search for saddle point solutions to the replica theory that break the replica symmetry.

Following the Parisi theory of RSB[19] we make the following "one-step" ansatz for the form of the order parameter matrix $Q_{\mu\nu}$,

$$Q_{\mu\nu} = \left(\begin{array}{ccc|ccc|c} 1 & q_1 & \cdots & q_0 & q_0 & \cdots & \\ q_1 & 1 & \cdots & q_0 & q_0 & \cdots & \cdots \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \\ \hline q_0 & q_0 & \cdots & 1 & q_1 & \cdots & \\ q_0 & q_0 & \cdots & q_1 & 1 & \cdots & \ddots \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \\ \hline & & & & & & \\ \vdots & & & \ddots & & & \ddots \\ & & & & & & \end{array}\right) \quad . \tag{B1}$$

Each block in this structure is an $m \times m$ matrix. All off-diagonal blocks consist of $Q_{\mu\nu} = q_0$. The diagonal block has $Q_{11} = 1$ and $Q_{\mu\nu} = q$ for $\mu \neq \nu$. The conjugate matrix $\hat{Q}_{\mu\nu}$ has a similar block structure. The order parameters R_μ and $\hat{R}_\mu$ are symmetric at the saddle point, i.e. $R_\mu = R$ and $\hat{R}_\mu = \hat{R}$, as before.

Upon replacing Eqs. (A7)–(A10) by the RSB ansatz and taking the appropriate $n \to 0$ limit, the free energy (A6) is given by

$$-\beta f = G_0(q_0, q_1, \hat{q}_0, \hat{q}_1, R, \hat{R}, m) - \alpha G_r(q_0, q_1, R, m) \ , \tag{B2}$$

where

$$\begin{aligned} G_0 = & \frac{1}{2}\left(m q_0 \hat{q}_0 + \left[(1-m)q_1 - 1\right]\hat{q}_1\right) - R\hat{R} \\ & + \frac{1}{m}\int_{-\infty}^{\infty} Dz_0 \ln\left(\int_{-\infty}^{\infty} Dz_1 \left[2\cosh\left(z_0\sqrt{\hat{q}_0} + z_1\sqrt{\hat{q}_1 - \hat{q}_0}\right)\right]^m\right) \\ G_r = & -\frac{2}{m}\int_0^{\infty} Dy \int_{-\infty}^{\infty} Dt_0 \ln\left(\int_{-\infty}^{\infty} Dt_1 \left[e^{-\beta} + (1 - e^{-\beta})H(u)\right]^m\right) \end{aligned} \tag{B3}$$

According to the RS ansatz, the saddle point takes the form

$$Q_{\sigma\rho} = \delta_{\sigma\rho} + (1-\delta_{\sigma\rho})q \ , \tag{A7}$$

$$R_\sigma = R \ , \tag{A8}$$

$$\hat{Q}_{\sigma\rho} = \delta_{\sigma\rho} + (1-\delta_{\sigma\rho})\hat{q} \ , \tag{A9}$$

$$\hat{R}_\sigma = \hat{R} \ . \tag{A10}$$

With this substitutions, the free energy takes the form

$$-\beta f = \mathrm{extr}_{q,R} \left\{ G_0(q,R,\hat{q},\hat{R}) - \alpha G_r(q,R) \right\} \ , \tag{A11}$$

where we have defined

$$G_r \equiv \lim_{n\to 0} \frac{\mathcal{G}_r}{n} \ , \tag{A12}$$

$$G_0 \equiv \lim_{n\to 0} \frac{\mathcal{G}_0}{n} \ . \tag{A13}$$

To calculate G_r, we substitute the RS ansatz into Eq. (A2) and perform the integral over $\hat{y}$, leaving

$$\begin{aligned} e^{-\mathcal{G}_r} = \int Dy \int \prod_\sigma \frac{dx_\sigma d\hat{x}_\sigma}{2\pi} \exp\left(-\frac{1}{2}\beta \sum_\sigma [g(x_\sigma) - g(y)]^2 \right) \\ \times \exp\left(-\frac{1}{2}(1-q)\sum_\sigma \hat{x}_\sigma^2 + i\sum_\sigma \hat{x}_\sigma (x_\sigma - Ry) \right) \\ \times \int Dt \exp\left(it\sqrt{q-R^2} \sum_\sigma x_\sigma \right) \ . \end{aligned} \tag{A14}$$

The auxiliary variable t has been introduced via the identity $\int Dt e^{bt} = \exp(\frac{1}{2}b^2)$. Performing the integrals over the $\hat{x}_\sigma$, and shifting and rescaling the x_σ integrals, we finally obtain

$$\begin{aligned} e^{-\mathcal{G}_r} = \int Dt \int Dy \\ \left[\int Dx \exp\left(-\frac{1}{2}\beta \left[g\left(x\sqrt{1-q} + yR - t\sqrt{q-R^2} \right) - g_0(y) \right]^2 \right) \right]^n \ . \end{aligned} \tag{A15}$$

Taking the limit $n \to 0$ gives Eq. (4.17) for G_r.

The RS result for $\mathcal{G}_0$ is

$$\begin{aligned} \mathcal{G}_0 = n \left(-R\hat{R} - \frac{1}{2}n(n-1)q\hat{q} - \frac{1}{2}\hat{q} \right) \\ \times \frac{1}{N} \ln \int D\mathbf{z} \left(\int d\mu(\mathbf{W}) \exp[\mathbf{W}\cdot(\mathbf{z}\sqrt{\hat{q}} + \mathbf{W}^0\hat{R})] \right)^n \ . \end{aligned} \tag{A16}$$

The $n \to 0$ limit of this expression yields Eq. (4.16) for G_0.

$$\times \exp\left(-\frac{1}{2}\beta\sum_{\sigma}[g(x_\sigma)-g(y)]^2+i\sum_{\sigma}x_\sigma\hat{x}_\sigma+iy\hat{y}\right)$$
$$\times\int D\mathbf{S}\exp\left[-iN^{-\frac{1}{2}}\left(\sum_{\sigma}\mathbf{W}^\sigma\hat{x}_\sigma+\mathbf{W}^0\hat{y}\right)\cdot\mathbf{S}\right]\ .$$

The average over $\mathbf{S}$ is now a simple Gaussian integral, and yields

$$e^{-\mathcal{G}_r[Q_{\sigma\rho},R_\sigma]}=\int\prod_{\sigma}\frac{dx_\sigma\,d\hat{x}_\sigma}{2\pi}\int\frac{dy\,d\hat{y}}{2\pi}$$
$$\times\exp\left(-\frac{1}{2}\beta\sum_{\sigma}[g(x_\sigma)-g(y)]^2+i\sum_{\sigma}x_\sigma\hat{x}_\sigma+iy\hat{y}\right) \tag{A2}$$
$$\times\exp\left(-\frac{1}{2}\sum_{\sigma,\rho}\hat{x}_\sigma\hat{x}_\rho Q_{\sigma\rho}-\hat{y}\sum_{\sigma}\hat{x}_\sigma R_\sigma-\frac{1}{2}\hat{y}^2\right)\ .$$

Since $\mathcal{G}_r$ depends on the weights only through the order parameters $Q_{\sigma\rho}$ and R_σ that were defined in Eqs. (4.9)–(4.10), the replicated partition function can be written as an integral over these order parameters

$$\langle\!\langle\, Z^n\,\rangle\!\rangle=\int\prod_{\sigma<\rho}dQ_{\sigma\rho}\int\prod_{\sigma}dR_\sigma\exp(-N\alpha\mathcal{G}_r[Q_{\sigma\rho},R_\sigma]) \tag{A3}$$
$$\times\int\prod_{\sigma}d\mu(\mathbf{W}^\sigma)\prod_{\sigma<\rho}\delta\left(Q_{\sigma\rho}-N^{-1}\mathbf{W}^\sigma\cdot\mathbf{W}^\rho\right)\prod_{\sigma}\delta\left(R_\sigma-N^{-1}\mathbf{W}^\sigma\cdot\mathbf{W}^0\right)$$
$$=\int\prod_{\sigma<\rho}\frac{dQ_{\sigma\rho}d\hat{Q}_{\sigma\rho}}{2\pi i}\int\prod_{\sigma}\frac{dR_\sigma d\hat{R}_\sigma}{2\pi i}\exp N(\mathcal{G}_0[Q_{\sigma\rho},R_\sigma,\hat{Q}_{\sigma\rho},\hat{R}_\sigma]-\alpha\mathcal{G}_r[Q_{\sigma\rho},R_\sigma])\ , \tag{A4}$$

where

$$\mathcal{G}_0=-\sum_{\sigma}\hat{R}_\sigma R_\sigma-\sum_{\sigma<\rho}\hat{Q}_{\sigma\rho}Q_{\sigma\rho} \tag{A5}$$
$$\times\frac{1}{N}\ln\int\prod_{\sigma}d\mu(\mathbf{W}^\sigma)\exp\left(\sum_{\sigma}\hat{R}_\sigma\mathbf{W}^\sigma\cdot\mathbf{W}^0+\sum_{\sigma<\rho}\hat{Q}_{\sigma\rho}\mathbf{W}^\sigma\cdot\mathbf{W}^\rho\right)$$

is the logarithm of the density of replicated networks with the overlaps $Q_{\sigma\rho}$ and R_σ.

In the thermodynamic limit, the integral (A3) over the order parameters is dominated by the saddle point in $Q_{\sigma\rho}$ and R_σ. The free energy is obtained by analytically continuing this saddle point to $n=0$,

$$-\beta f=\lim_{n\to0}\frac{1}{nN}\ln\langle\!\langle\, Z^n\,\rangle\!\rangle$$
$$=\mathrm{extr}_{Q_{\sigma\rho},R_\sigma}\left\{\mathcal{G}_0[Q_{\sigma\rho},R_\sigma,\hat{Q}_{\sigma\rho},\hat{R}_\sigma]-\alpha\mathcal{G}_r[Q_{\sigma\rho},R_\sigma]\right\}\ . \tag{A6}$$

Scientific, Singapore, 1987.
[20] S.F. Edwards and P.W. Anderson. Theory of spin glasses. *J. Phys.*, F5:965–974, 1975.
[21] D. Sherrington and S. Kirkpatrick. Solvable model of a spin glass. *Phys. Rev. Lett.*, 35:1792–1796, 1975.
[22] S. Kirkpatrick and D. Sherrington. Infinite-ranged models of spin-glasses. *Phys. Rev.*, B17:4384–4403, 1978.
[23] K. Binder and A.P. Young. Spin glasses: Experimental facts, theoretical concepts, and open questions. *Rev. Mod. Phys.*, 58:801–976, 1986.
[24] J.R.L. de Almeida and D.J. Thouless. Stability of the sherrington-kirkpatrick solution of a spin glass model. *J. Phys.*, A11:129–136, 1978.
[25] E. Gardner. Structure of metastable states in the hopfield model. *J. Phys.*, A19:L1047–L1052, 1986.
[26] W. Krauth and M. Mézard. Storage capacity of memory networks with binary couplings. *J. Phys. France*, 50:3057–3066, 1989.
[27] B. Derrida. Random-energy model: An exactly solvable model of disordered systems. *Phys. Rev.*, B24:2613–2626, 1981.
[28] J.F. Fontanari and R. Köberle. Landscape statistics of the binary perceptron. *J. Phys. France*, 51:1403–1413, 1990.
[29] D. Hansel and H. Sompolinsky. unpublished.
[30] E. Barkai, D. Hansel, and I. Kanter. Statistical mechanics of a multilayered neural network. *Phys. Rev. Lett.*, 65:2312–2315, 1990.
[31] E. Barkai and I. Kanter. Storage capacity of a multilayer neural network with binary weights. *Europhys. Lett.*, 14:107–112, 1991.

APPENDIX A: REPLICA THEORY OF PERCEPTRON LEARNING

The starting point of the replica calculations is the derivation of the replicated Hamiltonian (2.15), which resembles the derivation of the annealed G_{an} described in Appendix C of Part I. We introduce auxiliary variables x_σ, $\hat{x}_\sigma$, y, and $\hat{y}$ in order to simplify the average over $\mathbf{S}$,

$$\begin{aligned} e^{-\mathcal{G}_r[\mathbf{W}^\sigma]} &= \int D\mathbf{S} \exp\left(-\beta \sum_{\sigma=1}^{n} \epsilon(\mathbf{W}^\sigma;\mathbf{S})\right) \\ &= \int \prod_\sigma dx_\sigma \int dy \exp\left(-\frac{1}{2}\beta \sum_\sigma [g(x_\sigma)-g(y)]^2\right) \\ &\qquad \times \int D\mathbf{S} \prod_{\sigma=1}^{n} \delta\left(x_\sigma - N^{-\frac{1}{2}}\mathbf{W}^\sigma\cdot\mathbf{S}\right)\delta\left(y - N^{-\frac{1}{2}}\mathbf{W}^0\cdot\mathbf{S}\right) \\ &= \int \prod_\sigma \frac{dx_\sigma\, d\hat{x}_\sigma}{2\pi} \int \frac{dy\, d\hat{y}}{2\pi} \end{aligned} \tag{A1}$$

REFERENCES

[1] P. Carnevali and S. Patarnello. Exhaustive thermodynamical analysis of boolean learning networks. *Europhys. Lett.*, 4:1199–1204, 1987.

[2] J. Denker, D. Schwartz, B. Wittner, S. Solla, R. Howard, L. Jackel, and J. Hopfield. Automatic learning, rule extraction and generalization. *Complex Systems*, 1:877–922, 1987.

[3] M. Mézard W. Krauth and J-P. Nadal. *Complex Systems*, 1988.

[4] E. Baum and D. Haussler. What size net gives valid generalization? *Neural Computation*, 1:151–160, 1989.

[5] N. Tishby, E. Levin, and S. Solla. Consistent inference of probabilities in layered networks: Predictions and generalization. In *IJCNN International Joint Conference on Neural Networks*, volume 2, pages 403–409. IEEE, 1989.

[6] D. Haussler. Decision theoretic generalizations of the pac model for neural net and other learning applications. Technical report, UCSC, 1990.

[7] J.A. Hertz. Statistical dynamics of learning. *preprint: Nordita 90/34 S*, 1990.

[8] E. Gardner and B. Derrida. Three unfinished works on the optimal storage capacity of networks. *J. Phys.*, A22:1983–1994, 1989.

[9] P. Del Giudice, S. Franz, and M.A. Virasoro. Perceptron beyond the limit of capacity. *J. Phys. France*, 50:121–134, 1989.

[10] D. Hansel and H. Sompolinsky. Learning from examples in a single layer neural network. *Europhys. Lett.*, 11:687–692, 1990.

[11] G. Györgyi and N. Tishby. Statistical theory of learning a rule. In W.K. Theumann and R. Köberle, editors, *Neural Networks and Spin Glasses*, pages 3–36, 1990.

[12] G. Györgyi. Inference of a rule by a neural network with thermal noise. *Phys. Rev. Lett.*, 64:2957–2960, 1990.

[13] W. Kinzel and M. Opper. Dynamics of learning. In *Physics of Neural Networks*. Springer Verlag, Berlin, 1989.

[14] H. Sompolinsky N. Tishby and H.S. Seung. Learning from examples in large neural networks. *Phys. Rev. Lett.*, 65(13):1683–1686, 1990.

[15] G. Györgyi. First-order transition to perfect generalization in a neural network with binary synapses. *Phys. Rev.*, A41:7097–7100, 1990.

[16] M. Opper and D. Haussler. Generalization performance of bayes optimal classification algorithm for learning a perceptron. Submitted to PRL, 1991.

[17] A. Krogh and J. A. Hertz. Dynamics of generalization in linear perceptrons. In *Advances in Neural Information Processing Systems 3*, 1991.

[18] B. Derrida, R.B. Griffiths, and A. Prügel-Bennett. Finite size effects and bounds for perceptron models. *Saclay preprint: SPhT/91/035*, 1991.

[19] M. Mézard, G. Parisi, and M.A. Virasoro. *Spin Glass Theory and Beyond.* World

$$T_{opt}(\alpha) \to T_o \quad , \quad \alpha \to \infty \ . \tag{7.4}$$

For the boolean/mismatched perceptron (see Eq. (6.50) of Sec. VI.C) it diverges with α as

$$T_{opt}(\alpha) \sim \alpha^{3/5} \quad , \quad \alpha \to \infty \ . \tag{7.5}$$

These two models also differ in the effect of training at T_{opt}. Whereas in the linear perceptron the leading asymptotics remains the same as in Eq. (7.2), in the boolean perceptron the power law is modified to

$$\epsilon_g(T_{opt}(\alpha), \alpha) - \epsilon_{\min} \sim \alpha^{-4/5} \tag{7.6}$$

(see Eq. (6.52)), which is a significant improvement in the generalization. A nonzero optimal temperature has been also found in other unrealizable perceptron rules, although the shapes of $T_{opt}(\alpha)$ have not been studied yet.

3. Spin-Glass Phases

Another feature of unrealizable rules is the prevalence of spin-glass phases even at large α. In the present work we have evaluated the T at which the entropy vanishes. This provides a lower bound for the onset of the spin-glass phase. Here again the linear and boolean perceptrons differ. In the linear case this temperature levels off at a finite value as α increases, as shown in Fig. 9. In the boolean case it grows with the same power law as T_{opt}, see Eq. (6.54) and Fig. 12. These results suggest that spin-glass effects are strong in unrealizable cases.

The fact that the spin-glass phase exists also for large α suggests that the fluctuations in the training energy do not necessarily shrink as α increases. Nevertheless we believe that their *relative* scale does indeed vanish. This is because the energy itself grows linearly with α. This conjecture is supported by the observation that the temperature that marks the onset of spin-glass phenomena grows only sublinearly with α (see Eq. 6.54). The effect of increasing the number of examples on the roughness of the training energy surface (particularly in unrealizable rules) is an important issue which deserves further study.

Finally, it is interesting to note that in both cases studied here the zero entropy line is below T_{opt}. This suggests that by choosing cleverly the level of T one might be able to avoid the spin-glass regime while still obtaining good generalization performance.

ACKNOWLEDGMENTS

We would like to thank D. Hansel, D. Huse, E. Baum, and D. Haussler for helpful discussions. We also thank the Aspen Center for Physics, where part of this research was conducted. H.S. Seung acknowledges support from Harvard University and AT&T Bell Laboratories.

dynamical slowing may be caused by nonthermodynamic local minima, as studied in the linear/discrete model. These and other important dynamical phenomena deserve further investigation.

C. Unrealizable Rules

1. Shapes of Learning Curves

For unrealizable rules the optimal weight vector $\mathbf{W}^*$ is reached only in the limit $\alpha \to \infty$. This is true at all temperatures. As $T \to 0$ the generic shape of the learning curves is roughly that shown in Fig.8b. There is a critical value α_c which marks the loading capacity. As $\alpha \to \alpha_c$ the volume of solutions with zero training error shrinks to zero. Above α_c the training error increases, and approaches the same limit as the generalization error, i.e. approaches $\epsilon_{\rm min}$ from below as $\alpha \to \infty$.

Although this α_c does not mark a transition to perfect learning it does give a reasonable scale for the decrease of ϵ_g. It is reasonable to expect that most of the decrease in ϵ_g is achieved for α less than, say, $2\alpha_c$. Also, in some unrealizable models (not encountered in this work but see Refs. [10, 11]) there is a local maximum in ϵ_g near α_c.

The asymptotic learning curves of unrealizable models are very different from the high-T and AA predictions at any fixed T. For example, for the models with weight mismatch, these approximations predict α^{-2} tails for fixed T, but the RS theory yields (see Eqs. (6.25) and (6.42))

$$\epsilon_g(T,\alpha) - \epsilon_{\rm min} \sim \alpha^{-1} \tag{7.2}$$

for the linear output, and

$$\epsilon_g(T,\alpha) - \epsilon_{\rm min} \sim \alpha^{-\frac{1}{2}} \tag{7.3}$$

for the boolean one. It should be noted that an $\alpha^{-\frac{1}{2}}$ tail has also been found in a boolean perceptron with continuous weights. In this model the rule is unrealizable due to corruption of the examples by noise[11] and for the case of a random rule[10].

2. Optimal Temperature

Learning of unrealizable rules is generically characterized by nonmonotonic dependence of ϵ_g on T, implying that in general one should not anneal the training energy to $T = 0$ but rather train at finite T. This is one example of several behaviors that are all referred to as *overtraining*, describing a situation in which optimizing the training error leads to suboptimal generalization. Other examples include optimizing with respect to training time[29] and with respect to network size.

In general the optimal T varies with the number of examples. We have found two different behaviors of T_{opt}. In the case of linear output with mismatched weights (see Eq. (6.33) of Sec. VI.B)

In nonsmooth networks the generalization performance depends on the nature of the task as well as the network architecture, as the results of Secs. V and VI demonstrate. These results also indicate that there is a qualitative difference between the learning of realizable and unrealizable rules, as discussed below.

B. Realizable Rules

In all the realizable models studied in this work, the quenched RS behavior at finite T is qualitatively similar to that given by the AA. Furthermore the asymptotic shapes of the RS learning curves in the different models agree with those of the AA as presented in Part I. This suggests that for realizable rules the effect of disorder is minor for large α.

The main qualitatively new result of the quenched theory is the appearance of spin-glass phases, as shown in Figs. 1 and 4. These phases result from the randomness and frustration associated with optimizing for a particular realization of examples, and cannot be predicted by the high-T or the annealed approximations. Two special features distinguish the SG phases of a realizable models. First, it exists only at low T and in a restricted regime of α, typically $\alpha \lesssim 1$. Second, it is only metastable, i.e. its free energy (as well as training energy) is higher than that of the optimal state, which is therefore the true equilibrium state in that regime of α.

The absence of SG phases at large α indicates that as α increases, the relative scale of fluctuations in the training energy becomes smaller, i.e. the training energy surface gets smoother as the number of examples increases. This is indeed demonstrated by the analytical bounds of Sec. V.B on the number of local minima for the linear perceptron with discrete weights, Fig. 3. These bounds exhibit two important, and possibly general, features. First, the energy surface in the neighborhood of the optimal state $R = 1$ is smooth. Second, the number of local minima is exponentially large for small α but decreases monotonically with α. In the linear/discrete model there are no local minima above a critical value of α. Fontanari and Köberle[28] have studied numerically the local minima of the boolean/discrete model for small system sizes. The number of local minima was found to scale exponentially with N and to decrease monotonically with α. No critical value of α for the disappearance of local minima in the boolean model was found. However, a definitive conclusion would require the investigation of larger sizes.

Finally, in all our realizable models the generalization error decreases monotonically with T, so that the optimal temperature for learning is

$$T_{opt} = 0 \ . \tag{7.1}$$

Of course this holds only provided that equilibrium is achieved. Dynamically the situation may be quite different, as the time to equilibrate (or to reach some desired value of ϵ_g) may increase considerably with decreasing T. For example, the metastable spin glass phases in the linear/discrete and boolean/discrete models could cause severe slowing of learning dynamics. And even in the absence of spin glass effects,

so that the two dominant terms in the entropy cancel.

According to the RS theory, the generalization curve (6.42) has a $1/\sqrt{\alpha}$ tail. This result may not be correct, because the $\alpha \to \infty$ limit at fixed temperature always enters the region of RSB. We do, however, expect the RS result for the optimal generalization curve $\epsilon_{opt}(\alpha)$ to be correct, since the optimal generalization temperature $T_{opt}(\alpha)$ runs above the zero entropy line.

For this model, the spin glass effects are quite strong. The spin glass phase (shaded area in Fig. 12) persists for arbitrarily high temperatures. Furthermore, for fixed T, the RS entropy goes to $-\infty$ as $\alpha \to \infty$, indicating that replica symmetry is violated more and more severely.

At $T = 0$, the zero entropy line intersects the α axis at

$$\alpha_c = 1.106 \ , \tag{6.55}$$

which provides a more accurate estimate of α_c than the RS result (6.46). The true α_c derived from a RSB calculation might be even lower.

VII. DISCUSSION

In Part II we have studied the problem of learning from examples using the replica formalism. We have derived some general results and studied specific perceptron models. Our main focus in this part has been models of learning of unrealizable rules, for which the simple high-T and annealed approximations have been shown to be inadequate.

A. Inverse Power Law for Smooth Networks

The most important general result is the inverse power law for the asymptotic behavior of both the generalization and training errors, Eqs. (3.12) and (3.14) of Section III. These results also provide a simple relationship between the two errors, namely Eq. (3.15). This power law is consistent with the general bound obtained within PAC learning theory. However, our theory is not distribution-free and holds only for smooth networks. On the other hand it holds for general unrealizable rules whereas the PAC bounds are essentially for realizable rules.

The results regarding smooth networks were derived using a perturbative approach, i.e., assuming that essentially all the components of $\mathbf{W}$ deviate only slightly from $\mathbf{W}^*$ for sufficiently large α. Of course in reality there can also be contributions coming from $\mathbf{W}$ far away from $\mathbf{W}^*$, leading to nonperturbative terms to ϵ_g. However we expect that for sufficiently large α the nonperturbative terms will be negligible (e.g., exponentially dependent on α) relative to the power law contribution of the smooth fluctuations. On the other hand, these localized nonperturbative errors may be important for the dynamics of the training, since their relaxation may be extremely slow compared with the continuous fluctuations.

from zero and eventually diverges as $\alpha \to \infty$. This is similar to the temperature dependence found for the boolean/continuous model with noisy examples, as described in Ref. [11, 12]. Here however the rule is deterministic but unrealizable.

To investigate the limit $T \to \infty, \alpha \to \infty$, a double expansion in the variables β and T/α is needed In Appendix C, we discuss the expansion, and how to locate the dominant terms for evaluating $T_{opt}(\alpha)$ through power-counting. They are

$$\epsilon_g = \epsilon_{\min} + \frac{1}{4\pi}\left(\frac{\pi}{2}-1\right)^{1/4}\frac{1}{T}\sqrt{\frac{T}{\alpha}} + \frac{\pi^2}{12}\left(\frac{\pi}{2}-1\right)^{1/2}\frac{T^2}{\alpha^2} \ . \tag{6.49}$$

The $1/\sqrt{\alpha}$ term comes from expanding the coefficient of the quenched result (6.42) in β. Thus it represents the result of first taking the $T/\alpha \to 0$ limit and then the $\beta \to 0$ limit. The $1/\alpha^2$ term is just the leading term ¿from the high-T limit, which amounts to taking first the $\beta \to 0$ limit and then the $T/\alpha \to 0$ limit. The optimal T at large α is obtained by minimizing with respect to T, yielding

$$T_{opt} \sim 0.416\alpha^{3/5} \ , \tag{6.50}$$

which balances the two terms above.

We also define the *optimal generalization curve* ϵ_{opt}

$$\epsilon_{opt} = \epsilon_g(T_{opt}(\alpha), \alpha) \ . \tag{6.51}$$

Here it follows the nontrivial power law

$$\epsilon_{opt} - \epsilon_{\min} \sim 0.107\alpha^{-4/5} \ . \tag{6.52}$$

This power law decrease is faster than the zero T result $1/\sqrt{\alpha}$, Eq. (6.44). So we see that scaling the temperature with the number of examples can produce considerably better generalization performance than simple zero temperature training. However, the improvement still falls short of the $1/\alpha$ tails that we have seen in other models.

4. Replica Symmetry Breaking

The RS entropy of this model vanishes at a line $T_{s=0}(\alpha)$ shown in Fig. 12. Here again the $s = 0$ line provides a lower bound for the spin-glass transition temperature. Comparing with Fig. 9 two differences should be noted. First, the line intersects the α-axis at a finite value, suggesting that this model has a nonzero α_c. Second, the line diverges with α. This fact can be deduced ¿from the high temperature expansion for the entropy, which reads

$$s = -\frac{1}{4\pi}\left(\frac{\pi}{2}-1\right)^{1/4}\frac{1}{T}\sqrt{\frac{\alpha}{T}} + \frac{\pi^2}{6}\left(\frac{\pi}{2}-1\right)^{1/2}\frac{T}{\alpha} \ , \tag{6.53}$$

again keeping only the dominant terms. The zero entropy line follows the power law

$$T_{s=0} \sim 0.315\alpha^{3/5}, \qquad \alpha \to \infty \ , \tag{6.54}$$

$$\epsilon_g(T=0,\alpha) = \epsilon_{\min} + \frac{1}{3}\sqrt{\frac{2}{\pi}}\left(\frac{\pi}{2}-1\right)^{1/4}\frac{1}{\sqrt{\alpha}} + \mathcal{O}(\alpha^{-1}) \ . \tag{6.44}$$

For any fixed temperature, the generalization curve possesses a $1/\sqrt{\alpha}$ tail. This is the slowest asymptotics observed of any learning model in this work.

At $T = 0$, one solution of the RS equations is obtained by taking the limit $\beta \to \infty$ with $1-q$ finite. We can find when this solution breaks down by looking at the approach $q \to 1$. This yields the equations

$$\frac{R_c}{\sqrt{1-R_c^2}} = \frac{\alpha_c}{\pi} = \cot\frac{\pi}{\alpha_c} \ , \tag{6.45}$$

which have the solutions

$$\alpha_c = 1.99, \qquad R_c = 0.534 \ . \tag{6.46}$$

Above α_c, the limit $T \to 0$ must be taken with $\beta(1-q)$ finite, yielding

$$\frac{R}{\sqrt{1-R^2}} = \frac{\alpha}{\pi}\left[1 - e^{-\beta(1-q)/(1-R^2)}\right] \ , \tag{6.47}$$

$$\frac{2}{\pi} - R^2 = 2\alpha\int_0^{\sqrt{2\beta(1-q)}} Duu^2 H\left(\frac{Ru}{\sqrt{1-R^2}}\right) \ . \tag{6.48}$$

The learning curves and MC simulations for this model are shown in Fig. 10.

2. Numerical Simulations

Figure 10 shows Monte Carlo simulations at $T = 0.5$, $T = 1.0$, and $T = 2.0$ with $N = 75$, averaged over 64 samples. At all temperatures the generalization curves approach the slow $1/\sqrt{\alpha}$ asymptotics predicted by the RS theory, rather than the $1/\alpha^2$ of the annealed approximation. Good agreement of the training curve with the RS theory can be seen only in the higher temperature simulations. Below $T = 1$ substantial deviations in the training error occur beyond the $s = 0$ line of the RS theory. These deviations for $T = 0.5$ are significantly larger than those observed for the linear/mismatched model, indicating the importance of RSB for this model. However, the RS generalization curves appear to be good approximations even beyond the RSB transition.

3. Optimal Temperature

In the present model the dependence of the generalization error on temperature can be quite strong. Figure 11 shows the generalization and training errors plotted as functions of T for various values of α. The optimal temperature for generalization $T_{opt}(\alpha)$ is zero below a threshold value of $\alpha_{th} = 1.27$. Above this threshold, it rises

C. Boolean Output with Weight Mismatch

As in the previous model, the teacher weights are again drawn from a continuous Gaussian distribution, whereas the student weights are constrained to ± 1. However, here we consider the case where the perceptron transfer functions of both the teacher and the student are boolean. The maximal overlap is still $R_\infty = \sqrt{2/\pi}$, but the optimal generalization error is now

$$\epsilon_{\min} = \frac{1}{\pi} \cos^{-1} R_\infty = 0.206 \ . \tag{6.37}$$

As in the previous model, both the high-T and the annealed approximations predict $\epsilon_g(\alpha) - \epsilon_{\min} \propto \alpha^{-2}$, $\alpha \to \infty$. We shall see below that the true asymptotics are quite different.

1. Replica Symmetric Theory

For this model, G_0 is the same as that of the previous section, and G_r is that of the previous boolean models Eq. (5.37). Hence the RS free energy is given by

$$\begin{aligned} -\beta f = &-\frac{1}{2}(1-q)\hat{q} - \frac{1}{2}\frac{R^2}{1-q} + \int_{-\infty}^{\infty} Dz \ln 2\cosh(\sqrt{\hat{q}}z) \\ &+2\alpha \int_0^\infty Dy \int_{-\infty}^{\infty} Dt \ln\left[e^{-\beta} + (1-e^{-\beta})H(u)\right] \ . \end{aligned} \tag{6.38}$$

The saddle point equations are

$$\frac{R}{\sqrt{1-q}} = \frac{\alpha}{\pi} \int_{-\infty}^{\infty} Dt \, \frac{e^{-\frac{1}{2}v^2}}{(e^\beta - 1)^{-1} + H(v)} \ , \tag{6.39}$$

$$\hat{q} = \frac{R^2}{(1-q)^2} + \int_0^\infty Dy \int_{-\infty}^{\infty} Dt \, \frac{e^{-u^2}}{[(e^\beta - 1)^{-1} + H(u)]^2} \ , \tag{6.40}$$

$$q = \int Dz \tanh^2(\sqrt{\hat{q}}z) \ . \tag{6.41}$$

For large α, the generalization curve is given by an expansion in $1/\sqrt{\alpha}$

$$\epsilon_g = \epsilon_{\min} + \frac{1}{2\sqrt{\pi}} \left(\frac{\pi}{2} - 1\right)^{1/4} \frac{I(\beta)}{\beta^{3/2}} \frac{1}{\sqrt{\alpha}} + \mathcal{O}(\alpha^{-1}) \ , \tag{6.42}$$

where we have defined the integral

$$I(\beta) = \int Du \frac{u}{(e^\beta - 1)^{-1} + H(u)} \ . \tag{6.43}$$

As $\beta \to \infty$, $I(\beta) \approx (2\beta)^{3/2}/3$, so that

3. Replica Symmetry Breaking

The above RS theory is not exact at low T. First, the prediction of a finite value of α_c is probably incorrect. As in the corresponding realizable discrete model of Section V.B, we expect that

$$\alpha_c = 0 \ . \tag{6.34}$$

i.e., there is no vector of discrete weights that satisfies order N real, random linear equations. Second, the entropy of the RS solution becomes negative at low T, as shown in Fig. 9. The asymptotic form of the zero entropy line can be calculated by expanding the entropy in powers of α^{-1},

$$s = \frac{\pi}{6R_\infty}\left(T - \frac{3R_\infty}{\pi^2}\right)\frac{1}{\alpha} + \mathcal{O}(\alpha^{-2}) \ . \tag{6.35}$$

The $s = 0$ temperature approaches the finite limit

$$T_{s=0}(\alpha) \to \frac{3R_\infty}{\pi^2} = 0.242 \tag{6.36}$$

as $\alpha \to \infty$. The full line is drawn in Fig. 9 for all α. This temperature gives a lower bound on the temperature at which replica symmetry is broken. Below this temperature, a spin-glass phase with replica symmetry breaking occurs. As $T \to 0$, the zero entropy line approaches the origin, i.e. the RS entropy is negative for all α at $T = 0$. This means that the RS solution is incorrect at $T = 0$ for any α, and in particular that the RS prediction (6.30) for α_c is incorrect, which is consistent with our prediction, Eq. (6.34).

Below $T = 0.242$, the system never escapes from the spin glass phase even as $\alpha \to \infty$ (see Fig. 9). Note however that the entropy (6.35) approaches zero from below. This suggests that the effects of RSB become less severe as $\alpha \to \infty$.

4. Numerical Simulations

Figure 8 exhibits Monte Carlo simulations at $T = 0.5$ and $T = 0.1$ with $N = 100$ and 64 samples. The asymptotic behavior of $\epsilon_g - \epsilon_{\min}$ approaches the $1/\alpha$ prediction of the quenched theory rather than the $1/\alpha^2$ result of the annealed approximation. Even at $T = 0.1$ the fit to the RS theory is very good, even though for $\alpha > 0.76$ the curves are in the RSB region. This suggests that the effects of RSB in this system are weak. In Fig. 8 the RS $T = 0$ learning curves are plotted for comparison. Note that above $\alpha \gtrsim 0.6$, the generalization error for $T = 0.1$ is less than that at $T = 0$. At least in this range, we may thus conclude that the optimal generalization temperature $T_{opt} > 0$, assuming that the RS theory is a good approximation for the true behavior in the RSB region.

$$\epsilon_t = \epsilon_{\min} - \frac{\epsilon_{\min} R_\infty}{\alpha} + R_\infty \left[\frac{1}{2} + \frac{1}{\pi} - R_\infty + \frac{\pi^2}{24} T^2 \right] \frac{1}{\alpha^2} + \mathcal{O}(\alpha^{-3}) \ . \tag{6.26}$$

Thus for any fixed T the leading behavior of the errors is an α^{-1} power law. This power law is not a consequence of the general results (3.12) and (3.14), since the present network is not smooth. Note that in the high-T limit, only the term $T^2\alpha^{-2}$ survives. Thus for this unrealizable rule the behavior predicted by the high-T limit (as well as by the annealed approximation) does not reflect the correct behavior at large α for any fixed T.

In the $T \to 0$ limit with $1-q$ finite the saddle point equations reduce to

$$R = \alpha \ , \tag{6.27}$$

$$\hat{q}(1-q)^2 = -\alpha^2 + \alpha(1+q) \ , \tag{6.28}$$

$$1 - q = \int Dz \, \mathrm{sech}^2(\sqrt{\hat{q}} z) \ . \tag{6.29}$$

Criticality is reached when $\hat{q} \to \infty$, and $1-q \to 0$, which happens at

$$\alpha_c = 1 - \sqrt{1 - \frac{2}{\pi}} = 0.397 \ . \tag{6.30}$$

Above α_c, the limit $\beta \to \infty$ must be taken with $\beta(1-q)$ constant, yielding

$$0 = 2R^3 - (2+\alpha)R^2 + \frac{2\alpha}{\pi} \ , \tag{6.31}$$

$$\beta(1-q) = \frac{R}{\alpha - R} \ . \tag{6.32}$$

At α_c, $\beta(1-q)$ diverges, indicating that the finite $1-q$ solution takes over.

2. Optimal Temperature

We define the *optimal temperature*, $T_{opt}(\alpha)$ as the temperature that minimizes $\epsilon_g(T, \alpha)$. For our models of realizable rules, the optimal generalization error was at $T = 0$ for all values of α. But in general, T_{opt} may be greater than zero although the convexity of the free energy guarantees that the *training* error is a non-decreasing function of T. This is the case in the present model, as can be seen from Eq. (6.25), which implies that

$$T_{opt}(\alpha) \to \frac{6R_\infty}{\pi^2} = 0.485 \tag{6.33}$$

as $\alpha \to \infty$. Note however that the leading term of Eq. (6.25) is independent of T, implying that in the present model the effect on ϵ_g of optimizing with respect to T is relatively small.

which corresponds to the maximal overlap

$$R_\infty = \frac{1}{N}\sum_i |W_i^0| \\ = \int dW^0\, P(W^0)\, |W^0| = \sqrt{\frac{2}{\pi}}\ . \tag{6.19}$$

The second equality holds in the thermodynamic limit. The minimal generalization error, achieved in the limit $\alpha \to \infty$, is

$$\epsilon_{\min} = 1 - R_\infty = 0.202\ . \tag{6.20}$$

Before we present the replica solution we note that both the high-T and the annealed approximations predict for this model $\epsilon_g(\alpha) - \epsilon_{min} \propto \alpha^{-2},\ \alpha \to \infty$. As we shall see below, this is not the correct asymptotic behavior.

1. Replica Symmetric Theory

Here the replicated Hamiltonian G_r is that of the previous linear models, Eq. (5.6). From Eq. (4.16) G_0 may be calculated for the case of mismatched weight spaces. Making the change of variables $\hat{q} + \hat{R}^2 \to \hat{q}$, and then eliminating $\hat{R}$ altogether, one obtains the free energy

$$-\beta f = -\frac{1}{2}(1-q)\hat{q} - \frac{1}{2}\frac{R^2}{1-q} + \int Dz \ln 2\cosh(\sqrt{\hat{q}}z) \\ -\frac{\alpha}{2}\ln[1+\beta(1-q)] - \frac{\alpha}{2}\frac{\beta(q-2R+1)}{1+\beta(1-q)}\ . \tag{6.21}$$

The saddle point equations are

$$\hat{q} = \frac{R^2}{(1-q)^2} + \alpha\beta^2\frac{q-2R+1}{[1+\beta(1-q)]^2}\ , \tag{6.22}$$

$$R = \frac{\alpha\beta(1-q)}{1+\beta(1-q)}\ , \tag{6.23}$$

$$q = \int Dz \tanh^2(\sqrt{\hat{q}}z)\ . \tag{6.24}$$

To obtain the asymptotic form of the learning curves we have expanded the solutions in powers of α^{-1} keeping T fixed. We find

$$\epsilon_g = \epsilon_{\min} + \frac{\epsilon_{\min}R_\infty}{\alpha} \\ +R_\infty\left[-\frac{3}{2} - \frac{3}{\pi^3} - \frac{5}{\pi} + 4R_\infty + \frac{\pi^2}{24}\left(T - \frac{6R_\infty}{\pi^2}\right)^2\right]\frac{1}{\alpha^2} + \mathcal{O}(\alpha^{-3})\ , \tag{6.25}$$

$$\alpha_c = \frac{2 + \theta^2 - \theta\sqrt{2+\theta^2}}{2} \ . \tag{6.12}$$

For $\alpha < \alpha_c$, q is less than 1 and the values of the order parameters are given by the saddle point equations

$$R = \alpha \ , \tag{6.13}$$

$$q = \alpha + \frac{\alpha\theta^2}{1-\alpha} \ . \tag{6.14}$$

In this regime the minimum of ϵ_t is highly degenerate with an extensive zero temperature entropy. Note that for any threshold θ, the critical α satisfies the bounds $1/2 < \alpha_c < 1$. For $\alpha > \alpha_c$, $q \to 1$ as $T \to 0$, with $\beta(1-q)$ approaching a finite value. Thus at $T = 0$ the order parameters are given above α_c by

$$2R^3 = (2 + \alpha + \theta^2)R^2 - \alpha \ , \tag{6.15}$$

$$\beta(1-q) = \frac{R}{\alpha - R} \ . \tag{6.16}$$

The zero temperature α_c of this and other unrealizable models is similar to the α_c defined for the problem of loading random patterns onto a network. In both cases α_c is the limit of storage capacity above which no weight vectors can achieve zero training error. This is in contrast to the α_c that we defined for realizable models, above which there is exactly one weight vector which can achieve zero training error.

As in the zero threshold linear/continuous model, we expect that in this linear model the RS theory is exact for all T and α. Further, the learning dynamics should be quick, since there are no spurious local minima in the training energy.

B. Linear Perceptron with Weight Mismatch

In this model, the unrealizability is due to a mismatch in weight space between teacher and student. We assume that the weights of the teacher network, $\mathbf{W}^0$ are real valued whereas the trained network is restricted to $W_i = \pm 1$. For simplicity we consider here the case where the individual teacher weights W_i^0 are drawn from a continuous Gaussian distribution $P(W_i^0)$,

$$P(W_i^0) = (2\pi)^{-1/2} e^{-\frac{1}{2}(W_i^0)^2} \ , \tag{6.17}$$

As in the previous linear perceptron models the error function is given by Eq. (5.1), and the generalization error for a given network by Eq. (5.2). In the present model the optimal weights for the restricted architecture of the trained network are

$$W_i^* = \mathrm{sign}(W_i^0) \ , \tag{6.18}$$

$$\epsilon(\mathbf{W}) = 1 - R + \frac{1}{2}\theta^2 \ . \tag{6.4}$$

At $R = 1$, the generalization error takes on its minimum value, $\epsilon_{\rm min} = \theta^2/2$. Recall that according to Eq. (2.24), the average training and generalization errors both approach $\epsilon_{\rm min}$ as $\alpha \to \infty$, at all T.

As in the linear/continuous model, the $T = 0$ behavior can be understood from geometric arguments. Weight vectors with zero training error must satisfy P linear equations

$$(\mathbf{W} - \mathbf{W}^0) \cdot \mathbf{S}^l = \sqrt{N}\theta \ , \quad l = 1 \ldots P \ , \tag{6.5}$$

and the spherical constraint $\mathbf{W} \cdot \mathbf{W} = N$. There exists an $\alpha_c < 1$ such that for $\alpha \leq \alpha_c$, these equations have solutions. For $\alpha > \alpha_c$, the equations have no solution, so that the training error rises from zero, and asymptotically approaches $\epsilon_{\rm min}$ as $\alpha \to \infty$.

The RS free energy is

$$-\beta f = -\beta f_{\theta=0} - \frac{1}{2}\alpha \frac{\beta\theta^2}{1 + \beta(1-q)} \ . \tag{6.6}$$

where $f_{\theta=0}$ is the RS free energy of the realizable case, see Eqs. (5.5)-(5.6). The saddle point equations are the same as (5.7)–(5.10), except that Eq. (5.10) for $\hat{q}$ must be modified to

$$\hat{q} = \alpha\beta^2 \frac{q - 2R + 1 + \theta^2}{[1 + \beta(1-q)]^2} \ . \tag{6.7}$$

For fixed temperature and $\alpha \to \infty$, we find the asymptotic learning curves

$$\epsilon_g = \epsilon_{\rm min} + \frac{T + \theta^2}{2\alpha} + \mathcal{O}(\alpha^{-2}) \tag{6.8}$$

$$\epsilon_t = \epsilon_{\rm min} + \frac{T - \theta^2}{2\alpha} + \mathcal{O}(\alpha^{-2}) \tag{6.9}$$

To compare with the general results Eqs. (3.12) and (3.14) for smooth networks, we note that in the present case $\partial_i \epsilon(\mathbf{W}^0, \mathbf{S}) = -N^{-\frac{1}{2}}\theta S_i$, and $\partial_i \partial_j \epsilon(\mathbf{W}^0, \mathbf{S}) = N^{-1} S_i S_j$, so that the matrices defined in Eqs. (3.4)–(3.5) are

$$U_{ij} = N^{-1}\delta_{ij} \ , \tag{6.10}$$

$$V_{ij} = N^{-1}\theta^2\delta_{ij} \ . \tag{6.11}$$

Hence the coefficient $N^{-1}{\rm Tr}VU^{-1}$ in Eqs. (3.12) and (3.14) equals θ^2, in agreement with Eqs. (6.8) above.

In the $T \to 0$ limit there is a critical α below which the examples can be loaded with zero training error. This value is

account for the remaining disagreement between the numerics and the theory. See Ref. [18] for more simulations and an attempt at addressing the question of finite size scaling.

VI. PERCEPTRON LEARNING OF UNREALIZABLE RULES

So far we have studied examples of perceptron learning of realizable rules. Models of unrealizable rules possess some qualitatively different features. The first, already noted in Part I, is that the annealed approximation is in general a poor approximation to the performance of the system. This was illustrated in Part I using the linear/continuous model with unrealizable threshold. Here we solve this model exactly using the replica method. Second, the task of training for an unrealizable rule is more complex. In general there will be a critical value of α above which the examples cannot be trained at $T = 0$ with zero training error. In addition, spin-glass effects may be quite common, due to the frustration caused by the mismatch between the rule and the trained system. Third, the problem of overtraining is more prominent in unrealizable rules. In learning unrealizable rules the strategy of minimizing the training energy is in general not the optimal strategy; training at finite T may generate lower generalization errors. These points will be demonstrated in the last two examples below.

We study three rules that are unrealizable because of architectural mismatch. In the first model, the student and teacher perceptrons are mismatched due to their different transfer functions. In the second and third models, the teacher's weight vector is not included in the weight space of the student.

A. Linear/Continuous Perceptron with Unrealizable Threshold

In this model, the perceptrons of the student and teacher both have linear outputs and continuous weights, as in the model of Section V.A. However, in the present case the transfer function of the teacher has a threshold, i.e. the rule is given by Eq. (4.4) with a transfer function $g(x) = g_0(x)$, where

$$g_0(x) = x + \theta \ . \tag{6.1}$$

The output of the trained network is given by Eq. (4.1) with

$$g(x) = x \ . \tag{6.2}$$

The realizable linear/continuous model of Section V.A corresponds to the case $\theta = 0$.

The average of the error function

$$\epsilon(\mathbf{W};\mathbf{S}) = \frac{1}{2N}[(\mathbf{W} - \mathbf{W}_0)\cdot\mathbf{S} - \sqrt{N}\theta]^2 \tag{6.3}$$

over the input distribution, Eq. (4.3), yields the generalization error

vary with temperature within the SG phase; it is frozen at its value on the phase boundary $T_g(\alpha)$. The same holds true for ϵ_t and R, i.e.

$$q_0(T,\alpha) = q_0(T_g(\alpha),\alpha) \ , \tag{5.55}$$

$$\epsilon_t(T,\alpha) = \epsilon_t(T_g(\alpha),\alpha) \ , \tag{5.56}$$

$$R(T,\alpha) = R(T_g(\alpha),\alpha) \ , \tag{5.57}$$

everywhere in the SG phase. The value of $\epsilon_t(\alpha)$ in this phase is shown in Fig. 5. The only interesting temperature dependence occurs in m, which takes the linear form $m = T/T_g(\alpha)$.

A similar frozen SG phase exists in the perceptron model of Krauth and Mézard[26], and was first discovered by Derrida in the random-energy model[27]. However, unlike these other examples, the SG phase in the present model is only a metastable phase, as shown in Fig. 5. Hence the one-step RSB theory does not alter the RS prediction that the zero temperature thermodynamic transition is at $\alpha = 1.245$. However, it raises to

$$\alpha_o = 1.628 \tag{5.58}$$

the spinodal line for the vanishing of the metastable SG phase, marked (c) in Fig. 4.

3. Numerical Simulations

Figure 6 shows learning curves for Monte Carlo simulations with $N = 75$, averaged over 32 samples. At $T = 1$, there is quite a good fit if the transition is taken to occur at the spinodal line. At lower temperatures, the Metropolis algorithm tends to become trapped in local minima, making equilibration difficult. Hence we cannot use it to check the $T = 0$ predictions of the quenched theory.

The Metropolis algorithm produces a random walk through weight space that samples the Gibbs distribution. For small system sizes, we do not have to sample the weight space; we can explore it exhaustively. Gardner and Derrida[8] applied this idea to compute α_c at $T = 0$. Starting with all 2^N possible student vectors, an example is chosen at random, and all student vectors that "disagree" with the teacher are eliminated. Eventually some number of examples P is reached such that the addition of one more example produces perfect learning. Then $\alpha_c(N) = P/N$ for this sample. The procedure is then repeated with a different set of examples, so that α_c can be sample averaged.

We have extended the results of Gardner and Derrida up to the size $N = 23$. With the discrete inputs $S_i = \pm 1$, only odd N were used, so as to avoid the situation $\mathbf{W}\cdot\mathbf{S} = 0$. Figure 7 exhibits α_c as a function of $1/N$, with each data point an average of from 10^5 (large N) to 10^6 (small N) samples. Fitting to a quadratic function and extrapolating to $N = \infty$ yields the estimate $\alpha_c = 1.30 \pm 0.03$, which is fairly close to the prediction 1.245 of the quenched theory. It is possible that the quadratic function assumed here may be a poor approximation to the true finite size scaling. This would

2. Replica Symmetry Breaking and a Spin Glass Phase

To check the validity of the RS solution we have evaluated its entropy. In Fig. 4 we show the line $T_g(\alpha)$ on which the entropy of the metastable $R < 1$ state is zero. Since this system is discrete the entropy cannot be negative. Hence at least in the regime below $T_g(\alpha)$ the RS theory is incorrect for the metastable state.

To gain a better understanding of this metastable state, we have studied a solution to the replica mean–field theory with a one–step replica symmetry breaking ansatz. Our study is based on the work of Krauth and Mézard[26] concerning the problem of loading *random* dichotomies on a perceptron with discrete weights. The formal derivation of the mean field equations is presented in Appendix B. Here we discuss the physical meaning of this solution.

Recall that in the RS theory the $n \times n$ order parameter matrix $Q_{\mu\nu}$, defined in Eq. (4.9), takes a very simple form: its diagonal elements are all 1, and its off diagonal elements are q. In contrast, in the one–step ansatz, $Q_{\mu\nu}$ acquires two off-diagonal values q_0 and q_1 arranged in a block structure of $m \times m$ submatrices (see Appendix B). This block structure reflects the existence of many, almost degenerate, spin-glass (SG) states. These states are valleys in the free energy surface that are separated by barriers that diverge with the system size. The parameter q_1 represents the overlap of each state with itself, i.e. it is the order parameter q, Eq. (4.14), measured within a single valley a,

$$q_1 = N^{-1} \langle\!\langle \, \langle \mathbf{W}_a \rangle_T^2 \, \rangle\!\rangle \ . \tag{5.52}$$

The parameter q_0 represents the average overlap between a pair of two different states a and b, i.e.

$$q_0 = N^{-1} \langle\!\langle \, \langle \mathbf{W}_a \rangle_T \cdot \langle \mathbf{W}_b \rangle_T \, \rangle\!\rangle \ . \tag{5.53}$$

When the $n \to 0$ limit is taken, the size m of the step in $q(x)$ (see Appendix B) must be determined variationally, like q_0 and q_1. Denoting the Gibbs probabilities of the different states by $P_a = \exp(-\beta F_a)$, it can be shown that

$$m = 1 - \sum_a P_a^2 \ . \tag{5.54}$$

Hence m is the probability of finding two copies of the system in two different states. Near the transition temperature T_g, the degeneracy is very severe so that $m \approx 1$. At $T = 0$ the degeneracy is broken, and the Gibbs weight concentrates in the SG state with minimal energy, resulting in $m = 0$.

In the present model, a one–step RSB solution exists in the regime marked SG in Fig. 4, below the line $T_g(\alpha)$. This SG phase is special in that $q_1 = 1$, independent of both T and α. From Eq. (5.52) it follows that each of the different valleys is completely frozen, i.e. there are no fluctuations in $\mathbf{W}$. Indeed, the entropy of this phase is zero (to linear order in N). For fixed α, the order parameter q_0 does not

The RS free energy is given by combining G_0 of the perceptron with discrete weights, Eq. (5.16), with G_r for a Boolean output, Eq. (5.37), yielding

$$-\beta f = -\frac{1}{2}(1-q)\hat{q} - R\hat{R} + \int Dz \ln 2\cosh(\sqrt{\hat{q}}z + \hat{R}) \\ +2\alpha \int_0^\infty Dy \int_{-\infty}^\infty Dt \ln\left[e^{-\beta} + (1-e^{-\beta})H(u)\right] \tag{5.45}$$

The saddle point equations are

$$R = \int Dz \tanh(\sqrt{\hat{q}}z + \hat{R}) \ , \tag{5.46}$$

$$q = \int Dz \tanh^2(\sqrt{\hat{q}}z + \hat{R}) \ , \tag{5.47}$$

$$\hat{R} = \frac{\alpha}{\pi\sqrt{1-q}} \int_{-\infty}^\infty Dt \, \frac{e^{-\frac{1}{2}v^2}}{(e^\beta - 1)^{-1} + H(v)} \ , \tag{5.48}$$

$$\hat{q} = \frac{\alpha}{\pi(1-q)} \int_0^\infty Dy \int_{-\infty}^\infty Dt \, \frac{e^{-u^2}}{[(e^\beta - 1)^{-1} + H(u)]^2} \ , \tag{5.49}$$

where u and v are given by Eqs. (5.38) and (5.43) above.

At $T = 0$, the equations simplify somewhat, since $q = R$ and $\hat{q} = \hat{R}$. For α less than

$$\alpha_c = 1.245 \ , \tag{5.50}$$

there are two solutions, one with $R = 1$ (perfect generalization), and one with $R < 1$ (poor generalization). The $R < 1$ saddle point has the lower free energy, and is therefore the equilibrium phase. Upon crossing this critical α, the balance of free energy shifts, and there is a first order transition to the $R = 1$ state. Hence at $\alpha_c = 1.245$ there is a discontinuity in the generalization curve, a sudden transition to perfect learning. However, the $R < 1$ state still remains as a metastable phase until the spinodal point

$$\alpha_o = 1.492 \ . \tag{5.51}$$

With increasing α at any fixed T, there is the same sequence of thermodynamic transition followed by spinodal transition. The full RS phase diagram in the T-α plane is shown in Fig. 4, with the thermodynamic transition line marked as (a) and spinodal line as (b). The loci of the lines are very similar to that derived in Part I within the annealed approximation. To the left of the thermodynamic transition line (a), the state of poor generalization $R < 1$ is the equilibrium state, and the state of perfect generalization $R = 1$ is metastable. Between the lines (a) and (b), the situation reverses, with $R = 1$ becoming the equilibrium state and $R < 1$ the metastable state. To the right of the spinodal line (b), $R = 1$ is still the equilibrium state, but now the $R < 1$ state vanishes.

$$u \equiv \frac{t\sqrt{q-R^2} - yR}{\sqrt{1-q}} \ , \tag{5.38}$$

and $H(x)$ is defined as in (5.35). The saddle point equations are

$$R = \hat{R}(1-q) \ , \tag{5.39}$$

$$q = (\hat{q} + \hat{R}^2)(1-q)^2 \ , \tag{5.40}$$

$$\hat{R} = \frac{\alpha}{\pi\sqrt{1-q}} \int_{-\infty}^{\infty} Dt \, \frac{e^{-\frac{1}{2}v^2}}{(e^{\beta}-1)^{-1} + H(v)} \ , \tag{5.41}$$

$$\hat{q} = \frac{\alpha}{\pi(1-q)} \int_0^{\infty} Dy \int_{-\infty}^{\infty} Dt \, \frac{e^{-u^2}}{[(e^{\beta}-1)^{-1} + H(u)]^2} \ , \tag{5.42}$$

where u is defined in Eq. (5.38) above, and

$$v \equiv t\sqrt{\frac{q-R^2}{1-q}} \ . \tag{5.43}$$

The solution of these equations leads to a learning curve with a $1/\alpha$ tail for all T. Note that this power law is not a consequence of the general $1/\alpha$ law derived in Section III, since a network with a Boolean output is not a smooth network. At $T = 0$, the asymptotic learning curve is

$$\begin{aligned} \epsilon_g &= \sqrt{2}\left(\int Dt \frac{e^{-t^2/2}}{H(t)}\right)^{-1} \frac{1}{\alpha} + \mathcal{O}(\alpha^{-2}) \\ &= \frac{0.625}{\alpha} + \mathcal{O}(\alpha^{-2}) \ . \end{aligned} \tag{5.44}$$

Unlike the previous models, there is no transition at finite α to perfect learning at $T = 0$. In fact, there is no phase transition at any T and α.

Finally we note that in this model the replica symmetric theory is stable and most probably is the exact solution for all T and α. It has been recently shown using the replica method that the above $1/\alpha$ law can be improved by at most a factor of $\sqrt{2}$, using a Bayes optimal learning algorithm for the perceptron[16].

D. Boolean Output with Discrete Weights

1. Replica Symmetric Theory

We have shown in Part I that this model exhibits a discontinuous transition at any T from a state of poor generalization to perfect learning. Because of this unusual feature we will analyze the quenched theory of this model in some detail. This model was also studied independently, for zero temperature, by Györgyi[15]. We first study the replica symmetric theory and then investigate the replica symmetry breaking in this system.

which yields an upper bound for the typical number of such states, $\langle\langle \ln \mathcal{N}(N,R) \rangle\rangle$. A saddle point expansion yields

$$\begin{aligned} F(R,\alpha) = & \frac{1+R}{2} \ln H(-y) + \frac{1-R}{2} \left[\ln H(x) + \frac{1}{2}(x+y)^2 \right] \\ & - \frac{\alpha}{2} \left[\frac{x}{y} + \ln \frac{(1-R)y^2}{\alpha} + \ln 2 \right] \\ & - \frac{1-R}{2} \ln \frac{1-R}{2} - \frac{1+R}{2} \ln \frac{1+R}{2} \ , \end{aligned} \tag{5.34}$$

where we have defined

$$H(x) \equiv \int_x^\infty Dt \ , \tag{5.35}$$

and F has to be extremized with respect to x and y.

Figure 3 shows graphs of $F(R,\alpha)$ as a function of R for various values of α, obtained by solving the saddle point equations for x and y numerically. Wherever F is negative, then there are no local minima in the thermodynamic limit, since the annealed average is an upper bound. As seen in the Figure, F is negative near $R = 1$ for all values of α. In fact it can be shown that as $R \to 1$, $F \to -\alpha(\frac{1}{2}\ln 2 - \frac{1}{4}) = -0.0966\alpha$. This implies that there are no local minima in the neighborhood of the optimal $\mathbf{W}$, and the energy surface there is smooth. Note that at $R = 1$, $F = 0$, i.e. $\mathcal{N} = 1$ as expected.

Figure 3 also shows that as α increases the size of the hole in the number of local minima increases, until $\alpha = 2.39$, above which F is everywhere negative, so that there are no local minima at all. The implication for learning dynamics is that there exists some $\alpha_c \leq 2.39$ above which learning is fast even for $T = 0$. This prediction has been confirmed by numerical simulations of the model at $T = 0$. We have found that the system converges rapidly to $R = 1$ from almost all initial conditions for

$$\alpha \gtrsim 1.0 \ . \tag{5.36}$$

C. Boolean Output with Continuous Weights

The boolean perceptron with continuous weights has been previously studied in detail[11]. We present below a few of the results for completeness.

Since the a priori measure $d\mu(\mathbf{W})$ is the same as in the linear/continuous model of Section V.C above, G_0 is again given by (5.5). For a boolean output, G_r equals

$$G_r = -2 \int_0^\infty Dy \int_{-\infty}^\infty Dt \ln \left[e^{-\beta} + (1 - e^{-\beta}) H(u) \right] \ , \tag{5.37}$$

where

in particular the value of α_o, may not be large. The main qualitative change is the appearance of a line within the low R metastable state signaling a transition to a spin-glass metastable phase.

4. Local minima

The above finite-T statistical mechanical results account for the equilibrium state, as well as for metastable states that are separated by barriers that diverge as $N \to \infty$. However, the system may in addition possess states that are local minima of the energy (2.1), but are separated by barriers that remain finite in the $N \to \infty$ limit. Although these states are washed away by thermal fluctuations at any finite temperature, they may dominate the dynamics at $T = 0$. Even at finite low T these barriers may be high enough to prevent equilibration in a reasonable amount of time.

Following Gardner[25], we calculate an upper bound for the number of local minima as a function of the overlap R. Defining the outer product matrix of the examples by

$$T_{ij} = \frac{1}{N} \sum_{l=1}^{P} S_i^l S_j^l \ , \tag{5.28}$$

the energy can be written as

$$E = \frac{1}{2} \sum_{i \neq j=1}^{N} T_{ij} W_i W_j \ - \ \sum_{ij=1}^{N} T_{ij} W_i W_j^0 \ + \ \frac{1}{2} N \alpha \ , \tag{5.29}$$

with $W_i = \pm 1$. This has the standard form of an Ising Hamiltonian. The condition for a local minimum is

$$h_i W_i > 0 \ , \tag{5.30}$$

for all i, where the local fields h_i are defined by

$$h_i = \sum_{j=1}^{N} T_{ij} (W_j^0 - W_j) + \alpha W_i \ . \tag{5.31}$$

The number of local minima is then

$$\mathcal{N} = \mathrm{Tr}_{\mathbf{W}} \prod_i \Theta(h_i W_i) \ . \tag{5.32}$$

For a typical sample of examples we expect the logarithm of the number of local minima $\ln \mathcal{N}(N, R)$ to be extensive and hence self-averaging. We instead perform the simpler annealed average

$$N F(R, \alpha) \equiv \ln \langle\langle \, \mathcal{N}(N, R) \, \rangle\rangle \ , \tag{5.33}$$

these lines are very close to those of the approximate annealed phase diagram, shown in Fig. 2 of Part I.

At $T = 0$ the transition lines intersect the α–axis at

$$\alpha_c = 0 \ , \tag{5.26}$$

$$\alpha_o = 0.48 \tag{5.27}$$

The result $\alpha_c = 0$ implies that for any $\alpha > 0$, the training energy possesses a unique global minimum, $R = 1$. However, the training energy may still possess low-R metastable states. These states vanish above the spinodal point $\alpha_o = 0.48$.

2. Numerical Simulations

The results of the RS theory are in very good quantitative agreement with Monte Carlo simulations of the model at least for $T \geq 0.2$. In Fig. 2 we present the numerical results as well as the theoretical predictions for the training and generalization errors as a function of α for $T = 1$ and $T = 0.2$. At the higher T the prominent feature is the smooth decrease of ϵ_g to zero with an exponential tail. At the lower T the prominent feature is the rapid transition to $R \approx 1$ near $\alpha \approx 0.65$, in agreement with the theoretical prediction for the spinodal point for this temperature $\alpha_o = 0.66$. The roundness of the transition in the simulations is consistent with the expected smearing of the discontinuity in a finite system. The finite size effects are much more noticeable for the simulations at the higher temperature, as can be seen in Fig. 2b. We found numerically that the finite size corrections to the asymptotic theoretical curve for $T = 1$ scale approximatly like $N^{-0.2}$.

3. Spin-glass Phase

The above theoretical results were based on the replica symmetric (RS) ansatz described by Eqs. (4.13)–(4.14). Although we have not conducted a stability analysis of the RS theory, we expect it to be correct for the equilibrium phase. On the other hand the RS predictions regarding the metastable states are not exact, at least for sufficiently low T and α. We know this because the RS entropy of the metastable state with low R becomes negative below the dotted line in Fig. 1. Thus we expect that at least in this region the metastable state must be described by a theory with replica symmetry breaking (RSB). The interpretation of the RSB is that for small α the energy surface far away from the optimal overlap $R = 1$ is rough and contains many, almost degenerate, metastable states separated by high barriers. This leads to characteristic spin-glass behavior. On the other hand the energy surface in the neighborhood of the optimal state is rather smooth. Hence the equilibrium phase which has $R \approx 1$ does not exhibit spin-glass properties. This is substantiated by our calculation of the number of local minima in the next paragraph.

To summarize our scenario for the full RSB theory, we expect the location of the spinodal line at low T and α to differ from the RS results, but this difference, and

This model is the same as the preceding one, except that the weights are assumed to satisfy the Ising constraint $W_i = \pm 1$. The form of G_0 reflecting these a priori constraints on $\mathbf{W}$ is now

$$G_0 = -\frac{1}{2}(1-q)\hat{q} - R\hat{R} + \int Dz \ln 2\cosh(\sqrt{\hat{q}}z + \hat{R}) \ . \tag{5.16}$$

The replicated Hamiltonian G_r, which depends on the error function but not on the weight constraints, remains the same as in the previous case, Eq. (5.6). The resulting saddle-point equations are

$$R = \int Dz \tanh(\sqrt{\hat{q}}z + \hat{R}) \ , \tag{5.17}$$

$$q = \int Dz \tanh^2(\sqrt{\hat{q}}z + \hat{R}) \ , \tag{5.18}$$

$$\hat{R} = \frac{\alpha\beta}{1+\beta(1-q)} \ , \tag{5.19}$$

$$\hat{q} = \alpha\beta^2 \frac{q-2R+1}{[1+\beta(1-q)]^2} \ . \tag{5.20}$$

For any fixed temperature, $R \to 1$ and $q \to 1$ as $\alpha \to \infty$. To investigate this approach to the optimum, we note that for large α

$$\hat{q} \sim \alpha\beta^2[2(1-R) - (1-q)] \ , \tag{5.21}$$

$$\hat{R} \sim \alpha\beta \ . \tag{5.22}$$

Clearly there is a divergence of $\hat{R}$, the strength of the ordering term in the local field of Eq. (4.20). At the same time, $\sqrt{\hat{q}}/\hat{R} \to 0$, so that the relative strength of the disordering term is going to zero. This means that the saddle point equations for q and R, Eqs. (5.17)–(5.18), behave like

$$R \sim \tanh \hat{R} \approx \tanh(\alpha\beta) \ , \tag{5.23}$$

$$q \sim \tanh^2 \hat{R} \approx \tanh^2(\alpha\beta) \ . \tag{5.24}$$

Hence the generalization curve has an exponential tail

$$\epsilon_g = 1 - R \sim 2e^{-2\alpha\beta} \ . \tag{5.25}$$

This decay is much faster than the inverse power law of the linear/continuous model (5.14), reflecting the severe constraints on the weight space.

The RS phase diagram is shown in Fig. 1. Between the dashed spinodal lines marked (a) and (b), there are two locally stable solutions of the saddle point equations. On the solid line the two solutions have equal free energy. Upon crossing this line the equilibrium state changes discontinuously in a first order transition. The loci of

$$q = R = \begin{cases} \alpha, & \alpha \leq 1, \\ 1, & \alpha \geq 1. \end{cases} \tag{5.12}$$

For $\alpha < 1$, the fact that $q < 1$ reflects the degeneracy of the ground states, according to the definition Eq. (4.14). When α reaches the critical value

$$\alpha_c = 1 \ , \tag{5.13}$$

the degeneracy is finally broken ($q = 1$), and the training energy possesses a unique minimum $\mathbf{W} = \mathbf{W}_0$ ($R = 1$). Thus there is a continuous transition to perfect learning at $\alpha = 1$.

This model illustrates the generic behavior of smooth networks learning realizable rules at zero temperature. Below $\alpha_c \sim 1$, there are many degenerate global minima, with some nonzero entropy. At α_c, the entropy goes to zero (and $q \to 1$), signaling the emergence of a unique global minimum. Hence α_c is the location of a continuous transition to perfect learning.

However, this transition does not exist at any finite temperature, because of thermal fluctuations about the global minimum. From the saddle point equations, one can calculate that the asymptotic generalization curve is given by

$$\epsilon_g = \frac{T}{2\alpha} + \mathcal{O}(\alpha^{-2}) \ , \tag{5.14}$$

in accord with the general result (3.19) for smooth networks learning realizable rules. This means that at any finite temperature, perfect generalization is attained only in the limit of infinite α.

In this simple model, the replica symmetric theory is exact for all T and α. In other words, the replica symmetric ansatz is valid everywhere. However, the replica results add little to the analysis in Part I based on the annealed approximation (AA). The AA predicts correctly the asymptotic behavior (5.14), and the locus of the zero temperature transition, $\alpha_c = 1$. The only deficiency of the AA is that it modifies the shapes of the learning curves somewhat at low temperatures. This distortion is most severe at $T = 0$, where

$$\epsilon_g(T = 0, \alpha) = \begin{cases} 1 - \alpha, & \text{replica symmetric}, \\ \frac{2-2\alpha}{2-\alpha}, & \text{annealed}. \end{cases} \tag{5.15}$$

The full dynamics of this model has been studied by Krogh and Hertz[17]. Since E is quadratic there are no metastable states, and the convergence of the learning process is exponentially fast, except at the critical point $\alpha_c = 1$.

B. Linear Perceptron with Discrete Weights

$$\epsilon(\mathbf{W}) = \frac{1}{\pi}\cos^{-1} R \tag{5.4}$$

is simply proportional to the angle between the student and teacher weight vectors.

A. Linear Perceptron with Continuous Weights

The case of a perceptron with the quadratic error function Eq. (5.1) defined on a continuous weight space is particularly simple. Krogh and Hertz[17] have done a complete analysis of the training dynamics for this model; here we present only a simple calculation of the equilibrium properties.

Applying the replica symmetric theory, Eqs. (4.15)–(4.17), yields $-\beta f = G_0 - \alpha G_r$, with

$$\begin{aligned} G_0 = & \frac{1}{2}\lambda + \frac{1}{2}q\hat{q} - R\hat{R} - \frac{1}{2}\ln(\lambda + \hat{q}) \\ & + \frac{1}{2}\frac{\hat{R}^2 + \hat{q}}{\lambda + \hat{q}} - \frac{1}{2} \ , \end{aligned} \tag{5.5}$$

$$G_r = \frac{1}{2}\ln[1 + \beta(1-q)] + \frac{1}{2}\frac{\beta(q - 2R + 1)}{1 + \beta(1-q)} \ . \tag{5.6}$$

The additional order parameter λ is the Lagrange multiplier associated with the spherical constraint. Extremizing f with respect to the order parameters and eliminating λ yields

$$R = \hat{R}(1-q) \ , \tag{5.7}$$

$$q = (\hat{q} + \hat{R}^2)(1-q)^2 \ , \tag{5.8}$$

$$\hat{R} = \frac{\alpha\beta}{1 + \beta(1-q)} \ , \tag{5.9}$$

$$\hat{q} = \alpha\beta^2 \frac{q - 2R + 1}{[1 + \beta(1-q)]^2} \ . \tag{5.10}$$

Let us first consider the simple case of zero temperature. Only those weight vectors with zero training energy are allowed, i.e. those that satisfy

$$(\mathbf{W} - \mathbf{W}^0) \cdot \mathbf{S}^l = 0 \ , \quad l = 1, \ldots, P \ . \tag{5.11}$$

For $P \leq N$ these homogeneous linear equations determine only the projection of $\mathbf{W} - \mathbf{W}^0$ on the subspace spanned by the P random examples $\mathbf{S}^l$. This implies that the subspace of ground states of E has a huge degeneracy; it is $N - P$-dimensional. As $P \to N$ this degeneracy shrinks and for $P \geq N$ there is a unique solution to Eq. (5.11), $\mathbf{W} = \mathbf{W}^0$, for almost every random choice of examples.

The replica theory at $T = 0$ can be understood using this simple picture. The saddle point equations reduce to

competition between the two terms of the local field: the strength of the ordering term diverges ($\hat{R} \to \infty$) and the relative strength of the disorder goes to zero ($\sqrt{\hat{q}}/\hat{R} \to 0$).

One criterion for the validity of the RS Ansatz is the local stability of the RS saddle point. Often one finds that the RS solution becomes locally unstable at low temperatures, and hence invalid. In the phase diagram, the line at which the instability appears is known as the *Almeida-Thouless line*[24]. To find the true solution beyond this line, one must break replica symmetry.

For systems with discrete-valued degrees of freedom, a simpler diagnostic for replica symmetry breaking (RSB) is available, based on the fact that such systems must have nonnegative entropy. Below the *zero entropy line*, the entropy is negative and hence the RS ansatz must be incorrect. Hence the zero entropy line provides a lower bound for the temperature at which RSB first occurs. Since the zero entropy line is easier to calculate than the Almeida-Thouless line, we will rely on it to estimate the location of the RSB region.

In general it is extremely difficult to find the correct RSB solution. Although the model of Section V.D possesses a relatively simple RSB solution to be discussed later, for the rest of the models studied in this paper, we will consider only the RS solutions. We expect that the RS solution will still serve as a good approximation in the RSB region.

V. PERCEPTRON LEARNING OF REALIZABLE RULES

We now apply the formalism of the previous section to four specific models, which were previously studied within the annealed approximation in Part I. We consider two types of output unit, linear and boolean, and two types of weight spaces, spherical and Ising. Only realizable rules are considered, so that the target rule is a perceptron of identical architecture, transfer function, and weight space. Unrealizable rules will be studied in Section VI.

For the *linear perceptron*, the transfer function is $g(x) = x$. The error function, Eq. (4.5), is

$$\epsilon(\mathbf{W};\mathbf{S}) = \frac{1}{2N}[(\mathbf{W} - \mathbf{W}^0)\cdot\mathbf{S}]^2 \ . \tag{5.1}$$

Averaging this function over the whole input space, we find

$$\epsilon(\mathbf{W}) = 1 - R, \tag{5.2}$$

in accordance with Eq. (4.6) with a linear g.

A second output function is the *Boolean* output $g(x) = \mathrm{sign}(x)$, The error function, Eq. (4.5), is (up to a factor of 2)

$$\epsilon(\mathbf{W};\mathbf{S}) = \theta[-(\mathbf{W}\cdot\mathbf{S})(\mathbf{W}^0\cdot\mathbf{S})] \ , \tag{5.3}$$

which is 1 when the student and teacher agree, and 0 otherwise. The generalization error

The order parameter q is equal to

$$q = \frac{1}{N} \langle\!\langle \, \langle \mathbf{W} \rangle_T{}^2 \, \rangle\!\rangle \ . \tag{4.14}$$

The RS free energy of perceptron learning is (see Appendix A)

$$-\beta f = \frac{1}{N} \langle\!\langle \ln Z \, \rangle\!\rangle = G_0(q, R, \hat{q}, \hat{R}) - \alpha G_r(q, R) \ , \tag{4.15}$$

where

$$\begin{aligned} G_0 = & -\frac{1}{2}(1-q)\hat{q} - R\hat{R} \\ & + \frac{1}{N} \int D\mathbf{z} \ln \int d\mu(\mathbf{W}) \exp[\mathbf{W} \cdot (\mathbf{z}\sqrt{\hat{q}} + \mathbf{W}^0 \hat{R})] \ , \end{aligned} \tag{4.16}$$

$$\begin{aligned} G_r = & -\int Dt \int Dy \ln \int Dx \\ & \times \exp\left(-\frac{1}{2}\beta \left[g\left(x\sqrt{1-q} + yR - t\sqrt{q-R^2} \right) - g_0(y) \right]^2 \right) \ , \end{aligned} \tag{4.17}$$

and $\mathbf{z}$ is a vector of Gaussian variables $\{z_i\}_{i=1}^N$ with $D\mathbf{z} \equiv \prod_i Dz_i$. The free energy has to be extremized with respect to the order parameters q and R, and their "conjugate" counterparts $\hat{q}$ and $\hat{R}$. Differentiating with respect to $\hat{q}$ and $\hat{R}$ yields the saddle point equations

$$q = \frac{1}{N} \int D\mathbf{z} \, \langle \mathbf{W} \rangle_z \cdot \langle \mathbf{W} \rangle_z \ , \tag{4.18}$$

$$R = \frac{1}{N} \int D\mathbf{z} \, \langle \mathbf{W} \rangle_z \cdot \mathbf{W}^0 \ . \tag{4.19}$$

The definition of the average $\langle \mathbf{W} \rangle_z$ in these two equations reveals the meaning of the parameters $\hat{q}$ and $\hat{R}$,

$$\langle \mathbf{W} \rangle_z \equiv \frac{\int d\mu(\mathbf{W}) \, \mathbf{W} \exp[\mathbf{W} \cdot (\mathbf{z}\sqrt{\hat{q}} + \mathbf{W}^0 \hat{R})]}{\int d\mu(\mathbf{W}) \exp[\mathbf{W} \cdot (\mathbf{z}\sqrt{\hat{q}} + \mathbf{W}^0 \hat{R})]} \ . \tag{4.20}$$

In this equation, the local field $\mathbf{z}\sqrt{\hat{q}} + \mathbf{W}^0 \hat{R}$ acting upon $\mathbf{W}$ consists of two parts. The first is a Gaussian random field with variance $\hat{q}$ originating from the random fluctuations of the examples. The second is the bias towards the teacher weights $\mathbf{W}^0$, with an amplitude $\hat{R}$.

In general, we know from Eq. (2.24) that $\mathbf{W}$ must approach the optimal weight vector $\mathbf{W}^*$ as $\alpha \to \infty$. For a realizable perceptron rule ($\mathbf{W}^* = \mathbf{W}^0$), this means that $R \to 1$. If $\mathbf{W}^*$ is unique, the Gibbs distribution in weight space contracts about it as $\alpha \to \infty$, which means that $q \to 1$. The approach to the optimum is reflected in a

where

$$Dx \equiv \frac{dx}{\sqrt{2\pi}} e^{-x^2/2} \ , \tag{4.7}$$

and R is the overlap of the network with the teacher network, i.e.,

$$R = \frac{1}{N} \mathbf{W} \cdot \mathbf{W}^0 \tag{4.8}$$

(see Appendix C of Part I). The relationship between (4.6) and (4.5) is plain, since in both cases the arguments of g and g_0 are Gaussian random variables with unit variance and cross-correlation R.

The generalization error of the perceptron depends only on its overlap with the teacher, i. e. $\epsilon(\mathbf{W}) = \epsilon(R)$. Learning can be visualized very easily since $R = \cos\theta$, where θ is the angle between $\mathbf{W}$ and $\mathbf{W}^0$. The generalization error goes to zero as the angle between the student and the teacher weight vectors vanishes. Perfect learning corresponds to an overlap $R = 1$ or $\theta = 0$.

B. The Replica Theory of Perceptron Learning

The calculation of $\mathcal{G}_r$, (2.15), for perceptron learning is presented in Appendix A. The dependence of $\mathcal{G}_r$ on the weights is through the order parameters

$$Q_{\mu\nu} = \frac{1}{N} \mathbf{W}^\mu \cdot \mathbf{W}^\nu \ , \tag{4.9}$$

$$R_\mu = \frac{1}{N} \mathbf{W}^\mu \cdot \mathbf{W}^0 \ . \tag{4.10}$$

The first order parameter matrix measures the overlap between the weight vectors of two copies (replicas) of the system, $\mathbf{W}^\mu$ and $\mathbf{W}^\nu$. The second order parameter measures the overlap of the networks with the teacher. The values of these order parameters are obtained by extremizing $\mathcal{G}_r$.

In general, evaluating the saddle point equations for (4.9) and (4.10) requires making an ansatz about the symmetry of the order parameters at the saddle point. The simplest ansatz is the replica symmetric (RS) one,

$$Q_{\mu\nu} = \delta_{\mu\nu} + (1 - \delta_{\mu\nu}) q \ , \tag{4.11}$$

$$R_\mu = R \ . \tag{4.12}$$

In this case the order parameters of the replica theory have simple meanings in terms of thermal and quenched averages. The order parameter R is the expected value of the overlap with the teacher,

$$R = \frac{1}{N} \langle\!\langle \, \langle \mathbf{W} \rangle_T \, \rangle\!\rangle \cdot \mathbf{W}^0 \ . \tag{4.13}$$

IV. STATISTICAL MECHANICS OF PERCEPTRON LEARNING

In this and the following two Sections we will investigate concrete examples of learning in layered neural networks. All the examples will be concerned with a single-layer perceptron. In this Section we outline the general theoretical framework for learning a perceptron rule. In Section V we will study models of learning realizable rules, while Section VI will be devoted to unrealizable rules. The calculations will be done using the replica symmetric ansatz.

A. General Formulation

The perceptron is a network which sums a single layer of inputs S_j with synaptic weights W_j, and passes the result through a transfer function σ

$$\sigma = g\left(N^{-\frac{1}{2}}\sum_{j=1}^{N} W_j S_j\right) = g\left(N^{-\frac{1}{2}}\mathbf{W}\cdot\mathbf{S}\right) \ , \tag{4.1}$$

where $g(x)$ is a sigmoidal function of x. The normalization $1/\sqrt{N}$ in Eq. (4.1) is included to make the argument of the transfer function be typically of order unity. We assume for simplicity that the network space is restricted to vectors that satisfy the normalization

$$\sum_{i=1}^{N} W_i^2 = N \ . \tag{4.2}$$

The a priori distribution on the input space is assumed to be Gaussian,

$$P(\mathbf{S}) = (2\pi)^{-N/2} e^{-\frac{1}{2}\mathbf{S}\cdot\mathbf{S}} \ . \tag{4.3}$$

In this work, the teacher is always taken to be another perceptron. This means that the target rule is an input-output mapping of the form

$$\sigma^0(\mathbf{S}) = g_0\left(N^{-\frac{1}{2}}\mathbf{W}^0\cdot\mathbf{S}\right) \ , \tag{4.4}$$

where $\mathbf{W}^0$ is a fixed set of N weights W_i^0, and g_0 is a sigmoidal function. We assume that the teacher weights $\mathbf{W}^0$ also satisfy the normalization condition (4.2).

Training is performed by a stochastic dynamics with the training energy (2.1). For each example the error function is taken to be

$$\epsilon(\mathbf{W};\mathbf{S}) = \frac{1}{2}\left[g\left(N^{-\frac{1}{2}}\mathbf{W}\cdot\mathbf{S}\right) - g_0\left(N^{-\frac{1}{2}}\mathbf{W}^0\cdot\mathbf{S}\right)\right]^2 \ . \tag{4.5}$$

The average generalization error is

$$\epsilon(\mathbf{W}) = \langle\!\langle\, \epsilon(\mathbf{W};\mathbf{S})\,\rangle\!\rangle = \int Dx \int Dy\, \frac{1}{2}[g(x\sqrt{1-R^2}+yR) - g_0(y)]^2 \ , \tag{4.6}$$

$$\epsilon_t(T,\alpha) = \epsilon_{\min} + \left(\frac{T}{2} - \frac{\mathrm{Tr}VU^{-1}}{2N}\right)\frac{1}{\alpha} + \mathcal{O}(\alpha^{-2}) \ . \tag{3.14}$$

The above results predict an important relationship between the expected training and generalization errors at $T = 0$. According to Eqs. (3.12) and (3.14) both errors approach the same limit ϵ with a $1/\alpha$ power law. The coefficients of $1/\alpha$ in the two errors are identical in magnitude but different in sign yielding

$$\frac{\partial \epsilon_t}{\partial \alpha} = -\frac{\partial \epsilon_g}{\partial \alpha}, \quad \alpha \to \infty, \quad T = 0 \ . \tag{3.15}$$

This result can be used to estimate the expected generalization error ¿from the measured training error in smooth networks.

A special case occurs when the rule to be learned by the smooth network is realizable. This means that there exists a weight vector $\mathbf{W}^*$ within the allowed weight space that has zero generalization error, i.e.

$$\epsilon_{\min} = \epsilon(\mathbf{W}^*) = \int d\mu(\mathbf{S})\, \epsilon(\mathbf{W}^*;\mathbf{S}) = 0 \ . \tag{3.16}$$

This is equivalent to the statement that $\epsilon(\mathbf{W}^*;\mathbf{S}) = 0$ for all input vectors $\mathbf{S}$, because the error function was defined to be nonnegative. Since $\mathbf{W}^*$ minimizes $\epsilon(\mathbf{W})$, we can also assume $\partial_i \epsilon(\mathbf{W}^*) = 0$, as long as $\mathbf{W}^*$ lies in the interior of the weight space.

We have $\epsilon(\mathbf{W}^*;\mathbf{S}) \leq \epsilon(\mathbf{W};\mathbf{S})$, since the left hand side is zero, and the right hand side is nonnegative. This implies that

$$\partial_i \epsilon(\mathbf{W}^*;\mathbf{S}) \geq 0 \ . \tag{3.17}$$

for all $\mathbf{S}$. But also

$$\int d\mu(\mathbf{S}) \partial_i \epsilon(\mathbf{W}^*;\mathbf{S}) = \partial_i \epsilon(\mathbf{W}^*) = 0 \ . \tag{3.18}$$

Equations (3.17) and (3.18) together imply that $\partial_i \epsilon(\mathbf{W}^*;\mathbf{S}) = 0$ for all $\mathbf{S}$, which in turn implies $V_{ij} = 0$. Finally, we have the result

$$\epsilon_g(T,\alpha) = \frac{T}{2\alpha} + \mathcal{O}(\alpha^{-2}) \ . \tag{3.19}$$

The same holds for ϵ_t.

At zero temperature, the coefficient of $1/\alpha$ vanishes for ϵ_g and ϵ_t. Furthermore, for a smooth network learning a realizable rule, the higher order terms also vanish at $T = 0$. This implies that there is a finite value α_c for which $\epsilon_g = \epsilon_t = 0$ for all $\alpha > \alpha_c$, at $T = 0$. An example of such behavior will be presented in Section V.A below. Such a state we call a state of *perfect learning*.

It should be noted, however, that a realizable rule with a smooth network is an unrealistic situation. The smoothness requirement, as defined above, implies that the measure of error involves equalities and not inequalities. Therefore to realize a rule would necessitate infinite precision in determining the optimal weights. Unlike the case of discrete problems, learning tasks in smooth networks are generically unrealizable.

Since $N\alpha\mathcal{G}_r$ defines a Gaussian measure in weight space, it is straightforward to calculate the average deviations of the weights from $\mathbf{W}^*$. They are

$$\langle \delta \mathbf{W}^\sigma \rangle = 0 \ , \tag{3.6}$$

$$\begin{aligned} \langle \delta W_i^\sigma \delta W_j^\rho \rangle &= (N\alpha)^{-1} \ (A^{-1})_{ij}^{\sigma\rho} \\ &= \frac{1}{N\alpha} \left[T(U^{-1})_{ij} \delta^{\sigma\rho} + (U^{-1}VU^{-1})_{ij} \right] \ , \end{aligned} \tag{3.7}$$

where we have already taken the $n \to 0$ limit. Equations (3.6) and (3.7) have a simple meaning in terms of the physical system. Equation (3.6) reads

$$\langle\!\langle \, \langle \delta \mathbf{W} \rangle_T \, \rangle\!\rangle = \langle\!\langle \, \langle \mathbf{W} \rangle_T \, \rangle\!\rangle - \mathbf{W}^* = 0 \ . \tag{3.8}$$

The diagonal element (in the replica indices) of Eq. (3.7 yields the average correlations

$$C_{ij} = \langle\!\langle \, \langle \delta W_i \delta W_j \rangle_T \, \rangle\!\rangle = \frac{1}{N\alpha} \left[T(U^{-1})_{ij} + (U^{-1}VU^{-1})_{ij} \right] \ . \tag{3.9}$$

The first term, which is proportional to T, represents the contribution of the thermal fluctuations about $\mathbf{W}^*$. The second term represents the quenched fluctuations due to the random sampling of examples. This interpretation is supported by inspecting the off-diagonal element of Eq. (3.7), which is

$$\langle\!\langle \, \langle \delta W_i \rangle_T \langle \delta W_j \rangle_T \, \rangle\!\rangle = \frac{1}{N\alpha} (U^{-1}VU^{-1})_{ij} \ . \tag{3.10}$$

To evaluate the corrections to the generalization error, we expand Eq. (2.7) in powers of $\delta\mathbf{W}$, yielding

$$\epsilon_g = \epsilon_{\min} + \frac{1}{2} \mathrm{Tr} UC \ . \tag{3.11}$$

Substituting Eq. (3.9) one obtains

$$\epsilon_g(T, \alpha) = \epsilon_{\min} + \left(\frac{T}{2} + \frac{\mathrm{Tr} VU^{-1}}{2N} \right) \frac{1}{\alpha} + \mathcal{O}(\alpha^{-2}) \ . \tag{3.12}$$

The $1/\alpha$ expansion for the average training error can be evaluated by calculating first the corrections to the average free energy. Substituting Eq. (3.2) in Eq. (2.14) yields, after taking the $n \to 0$ limit

$$\begin{aligned} \beta F &= N\alpha\epsilon_{\min} + \frac{1}{2} \mathrm{Tr} \ln A \\ &= \frac{1}{2} N\alpha \left(\mathrm{Tr} \ln(\beta U) - \beta \mathrm{Tr} VU^{-1} \right) \ . \end{aligned} \tag{3.13}$$

Using Eq. (2.11) one obtains

This implies that $\mathbf{W}^*$ minimizes both the generalization error $\epsilon(\mathbf{W})$ and $\mathcal{G}_r$ in the $n \to 0$ limit. Hence we conclude that for any fixed temperature the training and generalization errors both approach the optimal value $\epsilon_{min} = \epsilon(\mathbf{W}^*)$ as $\alpha \to \infty$,

$$\epsilon_g \to \epsilon_{\min}, \qquad \epsilon_t \to \epsilon_{\min}, \qquad \alpha \to \infty \tag{2.24}$$

In the following Section, α will be used as a control parameter in a saddle point expansion to calculate the approach to the optimum for smooth networks.

III. SMOOTH NETWORKS

We define smooth networks to be those with continuously varying weights and error functions $\epsilon(\mathbf{W};\mathbf{S})$ that are at least twice differentiable with respect to $\mathbf{W}$ in the vicinity of the optimal weight vector $\mathbf{W}^*$. According to this definition, whether a network is smooth depends on both the smoothness of the weight space and the smoothness of the error function $\epsilon(\mathbf{W};\mathbf{S})$. In a smooth network, neither the output neuron nor the hidden neurons are saturated at the optimal $\mathbf{W}^*$. We now use the replica formalism to derive the asymptotic shape of the learning curves in these networks.

As stated above, the integrals over the weight space are dominated, as $\alpha \to \infty$, by the optimal weight vector $\mathbf{W}^*$, which minimizes both $\mathcal{G}_r$ (in the $n \to 0$ limit) and $\epsilon(\mathbf{W})$. At finite large α, the leading corrections to ϵ_g come ¿from the immediate neighborhood of $\mathbf{W}^*$. In a smooth network we can expand $\mathcal{G}_r$ in powers of

$$\delta W_i^\sigma \equiv W_i^\sigma - W_i^* \ . \tag{3.1}$$

The linear terms vanish since $\mathbf{W}^*$ is a minimum of $\mathcal{G}_r$. The leading corrections are

$$\mathcal{G}_r \approx \mathcal{G}_r^{\min} + \frac{1}{2} \sum_{ij\sigma\rho} \delta W_i^\sigma A_{ij}^{\sigma\rho} \ \delta W_j^\rho \ , \tag{3.2}$$

where

$$A_{ij}^{\sigma\rho} = \beta U_{ij} \delta^{\sigma\rho} - \beta^2 V_{ij} \ . \tag{3.3}$$

The matrix U_{ij} is the Hessian of the error function at the optimal weight vector $\mathbf{W}^*$, i.e.,

$$U_{ij} = \int d\mu(\mathbf{S}) \, \partial_i \partial_j \epsilon(\mathbf{W}^*, \mathbf{S}) \ . \tag{3.4}$$

The symbol ∂_i denotes $\partial / \partial W_i$. The matrix V_{ij} is

$$V_{ij} = \int d\mu(\mathbf{S}) \, \partial_i \epsilon(\mathbf{W}^*, \mathbf{S}) \partial_j \epsilon(\mathbf{W}^*, \mathbf{S}) \ . \tag{3.5}$$

D. Spin glasses and replica symmetry breaking

In some cases, the coupling between replicas produce only minor changes in the learning curves. In others, such terms can lead to the appearance of qualitatively new phases at low temperatures. These phases are conveniently described by the properties of the matrix

$$Q_{\mu\nu} = \frac{1}{N}\langle \mathbf{W}^{\mu} \cdot \mathbf{W}^{\nu} \rangle \ , \qquad \mu \neq \nu \tag{2.21}$$

which measures the expected overlap of the weights of two copies of the system. Since the replicated Hamiltonian (2.15) is invariant under permutation of the replica indices, one naively would expect that $Q_{\mu\nu} = q$ for all $\mu \neq \nu$. The physical interpretation of q would then be

$$q = N^{-1} \langle\!\langle \, \langle \mathbf{W} \rangle_T \cdot \langle \mathbf{W} \rangle_T \, \rangle\!\rangle \ , \tag{2.22}$$

and is known as the Edwards Anderson parameter in spin glasses theory[20]. The high temperature phase indeed possesses this *replica symmetry*. However, as the temperature is lowered, a spontaneous *replica symmetry breaking* (RSB) can occur, signaling the appearance of a spin-glass phase. In this phase, the expected values of correlations among different replicas do depend on the replica indices.

Formally, the spin glass phase is characterized by a non-trivial dependence of quantities such as $Q_{\mu\nu}$ on the replica indices. Physically, the spin glass phase is marked by the existence of many degenerate ground states of the energy (or free energy) which are well separated in configuration space. The different values of $Q_{\mu\nu}$ represent the distribution of overlaps among pairs of these ground states. This degeneracy is not connected with any simple physical symmetry, but is a result of strong frustration in the system. Furthermore, these degenerate ground states occupy disconnected regions in configuration space that are separated by energy barriers that diverge with N. Such barriers are important in the context of learning, since they lead to anomalously slow learning dynamics. One of the aims of studying the specific examples below will be to investigate the existence and role of spin-glass phases[23, 19].

E. The Large α limit

The replica formalism can also be used to investigate the behavior at a large number of examples, i.e., the $\alpha \to \infty$ limit. From Eq. (2.14) it is clear that the free energy, and the training and generalization errors are all weight space integrals that are dominated by the minimum of $\mathcal{G}_r$, as $\alpha \to \infty$. Denoting this minimum by $\mathbf{W}^{\sigma} = \mathbf{W}^{*}$, we find

$$\begin{aligned} \mathcal{G}_r^{\min} &= -\ln \int d\mu(\mathbf{S}) \exp(-\beta n \epsilon(\mathbf{W}^{*}; \mathbf{S})) \\ &= \beta n \epsilon(\mathbf{W}^{*}) + \mathcal{O}(n^2). \end{aligned} \tag{2.23}$$

$$\epsilon_t(T,\alpha) = \lim_{n\to 0} \frac{1}{n} \int d\mu(\mathbf{W}^\sigma) \frac{\partial \mathcal{G}_r[\mathbf{W}^\sigma]}{\partial \beta} e^{-N\alpha \mathcal{G}_r[\mathbf{W}^\sigma]} \ . \tag{2.17}$$

The simplicity of the replica formulation lies in the fact that only the *number* of examples remains as a simple prefactor in Eq. (2.14). All other example dependence has been removed, so that the replicated Hamiltonian $\mathcal{G}_r$ depends only on the form of $\epsilon(\mathbf{W};\mathbf{S})$ and on the nature of the a priori measure on the input space, $d\mu(\mathbf{S})$. Equations (2.14) and (2.15) also make explicit the scaling of the problem. Since $\epsilon(\mathbf{W};\mathbf{S})$ is defined to be of order 1, $\mathcal{G}_r$ itself is of order 1 times n. Thus, as the integral on the weight space in Eq. (2.14) is nN dimensional, where N is the number of degrees of freedom in weight space, P must scale as N.

The annealed approximation studied in Part I amounts to replacing $\langle\!\langle \ln Z \rangle\!\rangle$ by $\ln\langle\!\langle Z \rangle\!\rangle$. It can be obtained from Eq. (2.13) by setting $n = 1$ instead of taking the limit $n \to 0$. The replicated Hamiltonian $\mathcal{G}_r[\mathbf{W}]$ with $n = 1$ reduces to the annealed expression $G_{an}(\mathbf{W})$, Eq. (3.4) of Part I.

C. Replica theory and the High-T limit

The replica theory provides a simple derivation of the high-T limit, which was studied in Part I. Since $\mathcal{G}_r$ is an intensive quantity independent of P, the high-T limit can be derived by simply expanding it in powers of β. The leading terms are

$$P\mathcal{G}_r[\mathbf{W}^\sigma] = N\left[\alpha\beta \sum_{\sigma=1}^{n} \epsilon(\mathbf{W}^\sigma) - \frac{1}{2}\alpha\beta^2 \sum_{\sigma,\rho=1}^{n} g(\mathbf{W}^\sigma,\mathbf{W}^\rho) + \mathcal{O}(\beta^3)\right] \ , \tag{2.18}$$

where

$$g(\mathbf{W}^\sigma,\mathbf{W}^\rho) = \int d\mu(\mathbf{S})\, \epsilon(\mathbf{W}^\sigma;\mathbf{S})\epsilon(\mathbf{W}^\rho;\mathbf{S}) - \epsilon(\mathbf{W}^\sigma)\epsilon(\mathbf{W}^\rho) \ . \tag{2.19}$$

Note that g measures the correlations in the errors of two different weights on the same example.

The general form of Eq. (2.18) is similar to that of a spin-glass replica Hamiltonian[21,22]. The first term is the one that survives the high-T limit. It represents the non-random part of the training energy. Taking into account only this contribution leaves the different replicas uncoupled, and hence F reduces to its high-T limit

$$-\beta F \to \ln \int d\mu(\mathbf{W}) e^{-N\beta\alpha\epsilon(\mathbf{W})} \ , \tag{2.20}$$

in which the training energy becomes proportional to the generalization function, i.e. $E(\mathbf{W}) \to P\epsilon(\mathbf{W})$. The high-$T$ limit was discussed in detail in Section II.C of Part I. As T decreases the second term of Eq. (2.18) becomes important. This term is a coupling between different replicas which originates from the randomness of the examples.

$$F(T,\alpha) = -T\langle\!\langle \ln Z \rangle\!\rangle, \tag{2.8}$$

$$S(T,\alpha) = -\int d\mu(\mathbf{W})\langle\!\langle P(\mathbf{W}) \ln P(\mathbf{W}) \rangle\!\rangle. \tag{2.9}$$

They are related by the identity

$$F = P\epsilon_t - TS \ . \tag{2.10}$$

Knowing F the expected training error can be evaluated via

$$\epsilon_t = \frac{1}{N\alpha}\frac{\partial(\beta F)}{\partial\beta} \ , \tag{2.11}$$

and the entropy by

$$S = -\frac{\partial F}{\partial T} \ . \tag{2.12}$$

B. The Replica Method

To evaluate the correct behavior at all T one has to evaluate quenched averages such as Eq. (2.8) and its derivatives. Such averages are commonly studied using the replica method[20]. The average free energy is written as

$$-\beta F = \langle\!\langle \ln Z \rangle\!\rangle = \lim_{n\to 0}\frac{1}{n}\ln\langle\!\langle Z^n \rangle\!\rangle \ . \tag{2.13}$$

One first evaluates Eq. (2.13) for integral n and then analytically continues to $n = 0$. Using Eqs. (2.1) and (2.4) we obtain

$$\langle\!\langle Z^n \rangle\!\rangle = \int \prod_{\sigma=1}^{n} d\mu(\mathbf{W}^\sigma) \exp(-N\alpha\mathcal{G}_r[\mathbf{W}^\sigma]) \ , \tag{2.14}$$

where the replicated Hamiltonian is

$$\mathcal{G}_r[\mathbf{W}^\sigma] = -\ln\int d\mu(\mathbf{S}) \exp\left(-\beta\sum_{\sigma=1}^{n}\epsilon(\mathbf{W}^\sigma;\mathbf{S})\right). \tag{2.15}$$

The average generalization error (2.7) can be rewritten using replicas as

$$\begin{aligned}\epsilon_g(T,\alpha) &= \lim_{n\to 0}\langle\!\langle Z^{n-1}\int d\mu(\mathbf{W})\epsilon(\mathbf{W})e^{-\beta E(\mathbf{W})}\rangle\!\rangle \ , \\ &= \lim_{n\to 0}\int\prod_{\sigma=1}^{n} d\mu(\mathbf{W}^\sigma)\epsilon(\mathbf{W}^1)e^{-N\alpha\mathcal{G}_r[\mathbf{W}^\sigma]} \ ,\end{aligned} \tag{2.16}$$

and the average training error (2.6) as

$$E(\mathbf{W}) = \sum_{l=1}^{P} \epsilon(\mathbf{W}; \mathbf{S}^l) \ , \tag{2.1}$$

is the sum of all the errors on a *training set* of *examples* $(\mathbf{S}^l, \sigma_0(\mathbf{S}^l))$, with $l = 1, \ldots, P$. The inputs $\mathbf{S}^l$ are chosen at random from the entire input space according to some a priori normalized measure $d\mu(\mathbf{S})$. As discussed in Part I, the proper thermodynamic limit is achieved when the number of examples P scales linearly with the number of weights N, i.e.

$$\alpha \equiv \frac{P}{N} \tag{2.2}$$

remains finite as $N \to \infty$.

We assume a stochastic training algorithm that results in a Gibbs distribution of networks

$$P(\mathbf{W}) = Z^{-1} e^{-\beta E(\mathbf{W})} \ , \tag{2.3}$$

independent of initial conditions. The variance of the noise in the training procedure is just the temperature $T = 1/\beta$ of the Gibbs distribution. The normalization factor Z is the partition function

$$Z = \int d\mu(\mathbf{W}) \exp(-\beta E(\mathbf{W})), \tag{2.4}$$

where $d\mu(\mathbf{W})$ is the normalized a priori measure on the network space.

The goal of supervised learning is to find a network that performs well on all inputs, not just those in the training set. Such a network minimizes the *generalization function*

$$\epsilon(\mathbf{W}) = \int d\mu(\mathbf{S})\, \epsilon(\mathbf{W}; \mathbf{S}) \ , \tag{2.5}$$

where $d\mu(\mathbf{S})$ is the a priori probability measure on the input space.

The average training and generalization errors are given by

$$\epsilon_t(T, \alpha) \equiv \frac{1}{N\alpha} \langle\!\langle\, \langle E(\mathbf{W}) \rangle_T \,\rangle\!\rangle \ , \tag{2.6}$$

$$\epsilon_g(T, \alpha) \equiv \langle\!\langle\, \langle \epsilon(\mathbf{W}) \rangle_T \,\rangle\!\rangle \ . \tag{2.7}$$

The thermal average $\langle \cdots \rangle_T$ is with respect to the Gibbs distribution $P(\mathbf{W})$ of networks, for a given realization of examples. The quenched average $\langle\!\langle \cdots \rangle\!\rangle \equiv \int \prod_l d\mu(\mathbf{S}^l)$ is over the a priori distribution of examples. For fixed temperature, the graphs of the above average generalization and training errors versus α will be referred to as *generalization* and *training curves*, respectively. Both graphs will be referred to as *learning curves*.

The free energy F and entropy S of the network are given by

smooth networks, the asymptotic approach of the generalization error to its minimal value is an inverse power law in the number of examples. However, in nonsmooth networks the shape of the generalization curve depends on the details of the problem, e.g. the constraints on the weight space, and the discreteness of the neural output states, as demonstrated by the specific models studied in the following Sections.

In Section IV we outline the replica theory of perceptron learning. It assumes that the system is in a *replica symmetric* phase[19]. This means that the thermodynamic state of the system is unique. We discuss briefly the conditions under which the replica symmetric ansatz holds. In Section V the replica formalism is applied to the same models of learning realizable rules that were studied in Part I. The replica symmetric theory confirms qualitatively the results of Part I. A new feature from the replica theory is the existence of metastable spin glass phases in the models with discrete weights. The spin glass phases are characterized by *replica symmetry breaking*, which signals the existence of degenerate local minima in the network weight space that are separated by high energy barriers. Since these spin glass phases are metastable, they do not affect equilibrium properties, but may have significant consequences for learning dynamics.

In Section VI, models of unrealizable perceptron rules are analyzed. We focus mostly on cases of *weight mismatch* where the weights of the *teacher perceptron*, which generates the rule, are outside the allowed range of the trained network. Unlike the realizable models, the true shapes of the learning curves are very different from the predictions of the annealed approximation. We also show that at low temperatures the equilibrium phase of these systems is in general a spin-glass phase even for a large number of examples. Studying the temperature dependence of the generalization error at a fixed number of examples, we show that unrealizable rules exhibit the phenomenon of *overtraining*. This means that the generalization error is optimized by training at a non-zero temperature. Section VII contains a summary of our results and a discussion of the main issues.

II. GENERAL THEORY

A. Learning from Examples

Let $\sigma = \sigma(\mathbf{W};\mathbf{S})$ denote the output of a feedforward network with inputs $\mathbf{S}$, synaptic weights $\mathbf{W}$, and scalar output σ. The goal of learning is to find the set of weights $\mathbf{W}^*$ such that $\sigma(\mathbf{W};\mathbf{S})$ best approximates (or realizes exactly) some target function, or "rule" $\sigma_0(\mathbf{S})$.

The *error function* $\epsilon(\mathbf{W};\mathbf{S})$ measures the deviation of the network output $\sigma(\mathbf{W};\mathbf{S})$ from the target rule $\sigma_0(\mathbf{S})$ on a given input $\mathbf{S}$. It should be zero whenever the network and the target rule agree, and positive everywhere else.

The *training energy*

I. INTRODUCTION

The statistical mechanics of learning and generalization in layered neural networks has been a subject of fundamental interest to both practitioners and theoreticians[1–7]. Understanding the process of learning from examples requires studying the competition between the randomness in the sampling of examples and the underlying order affected by the rule that generates the input-output relations of the examples. Consequently, the natural theoretical framework for studying this problem is the replica theory, which was developed for spin-glasses and other statistical mechanical systems with quenched (i.e. static) disorder. In Part I of this work we presented and studied two approximate treatments of learning in layered networks. One, the high-temperature limit, completely ignores the randomness of the examples and focuses on the effects of the underlying rule as well as the network architecture on the system's performance. The second is the annealed approximation. It incorporates in an approximate simplified fashion some of the effects of the randomness. In this Part we present the replica theory of problem.

Application of the replica method to study learning a classification task in a single layer network was suggested by Gardner and Derrida[8] and further pursued by del Giudice et al.[9], Hansel and Sompolinsky[10], and Györgyi and Tishby[11, 12]. Other studies, directly related to the present work, have recently been published[13–18].

In the present work we study the replica theory of learning both *realizable* and *unrealizable* rules. In the case of realizable rules, there exists a set of synaptic weights that make the trained network realize the target rule perfectly. In the unrealizable case, no such set of weights exists, either because of architectural constraints or the presence of noise in the data. The purpose of our study is three-fold:

1. derive general properties of learning from examples in layered networks;

2. derive the exact behavior of models of *realizable perceptron rules*, and compare with the approximate results of Part I;

3. and study the properties of learning *unrealizable perceptron rules*, where the above mentioned approximations fail completely, as discussed in Part I.

In Section II we begin with a brief summary of the statistical mechanical formulation of learning from examples with a stochastic training dynamics, introducing the Gibbs distribution on the space of networks. We first show that, at any fixed temperature, the weight vector with optimal generalization properties is approached smoothly as the number of examples goes to infinity. Thus the presence of noise in the training dynamics does not destroy the potential for learning.

In Section III, we study the properties of *smooth* networks, i.e. networks with smooth neuronal outputs and continuously varying weights. Interestingly enough, such networks have a universal asymptotic shape of the generalization (and training) error as a function of the number of examples per weight. It is shown that for general

Submitted to Physical Review A

Statistical Mechanics of Learning From Examples II. Quenched Theory and Unrealizable Rules

H. S. Seung, H. Sompolinsky
Racah Institute of Physics, Hebrew University, Jerusalem 91904, Israel

N. Tishby
AT&T Bell Laboratories, 600 Mountain Ave., Murray Hill, NJ 07974

Learning from examples in feedforward neural networks is studied using the replica method. We focus on the *generalization curve*, which is defined as the average generalization error as a function of the number of examples. For smooth networks, i.e. those with continuously varying weights and smooth transfer functions, the generalization curve is found to asymptotically obey an inverse power law. This implies that generalization curves in smooth networks are generically gradual. In contrast, for discrete networks, discontinuous learning transitions can occur. We illustrate both gradual and discontinuous learning with four single-layer perceptron models. In each model, a perceptron is trained on a *perfectly realizable* target rule, i.e. a rule that is generated by another perceptron of identical architecture. The replica method yields results that are qualitatively similar to the approximate results derived in Part I for these models. We study another class of perceptron models, in which the target rule is *unrealizable* because it is generated by a perceptron of mismatched architecture. In this class of models, the quenched disorder inherent in the random sampling of the examples plays an important role, yielding generalization curves that differ from those predicted by the simple annealed approximation of Part I. In addition this disorder leads to the appearance of equilibrium spin glass phases, at least at low temperatures. Unrealizable rules also exhibit the phenomenon of *overtraining*, in which training at zero temperature produces inferior generalization to training at nonzero temperature.

P.A.C.S. numbers: 87.10+e, 02.50+f, 05.20-y

Typeset Using *REVTEX*