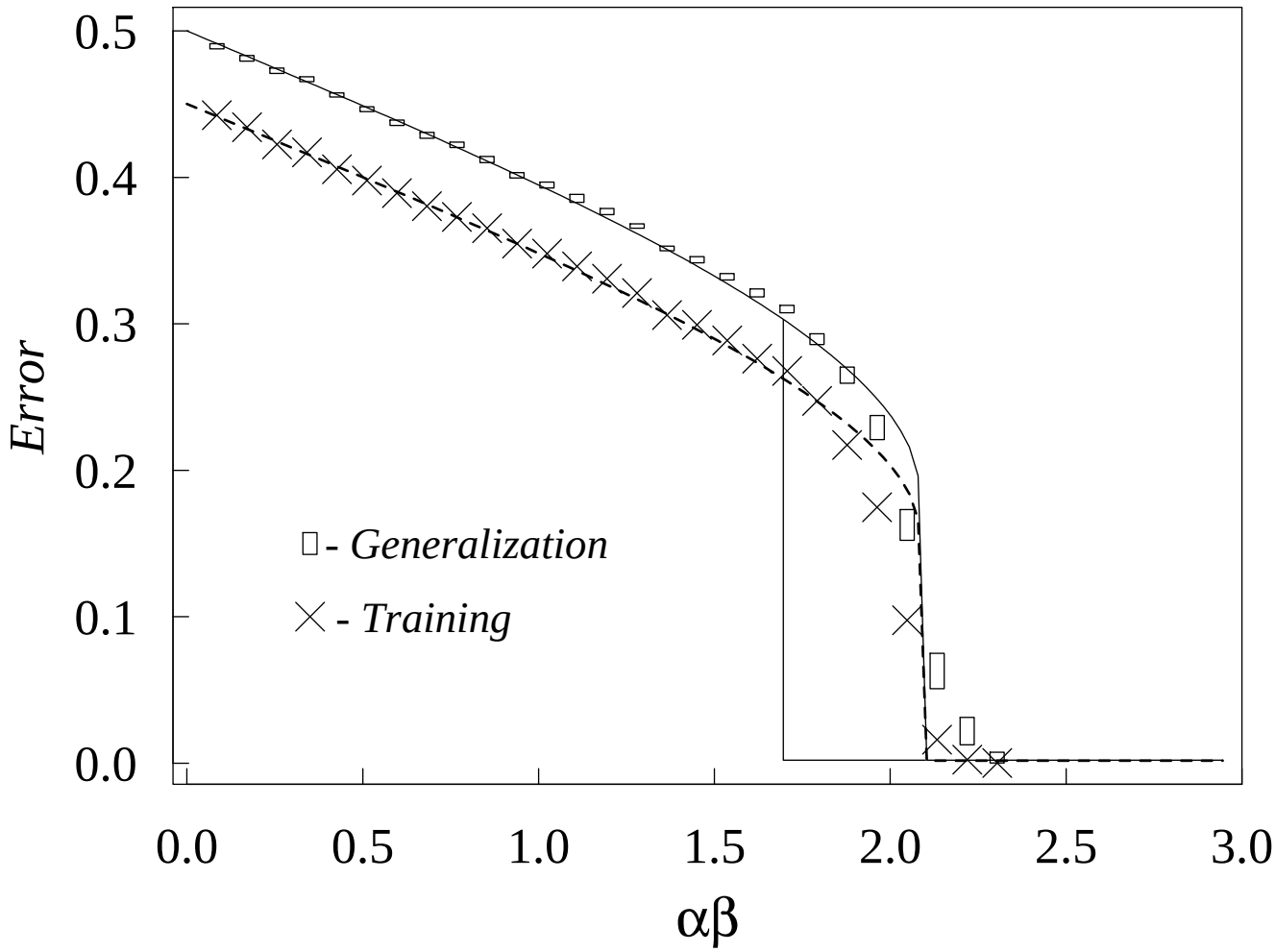
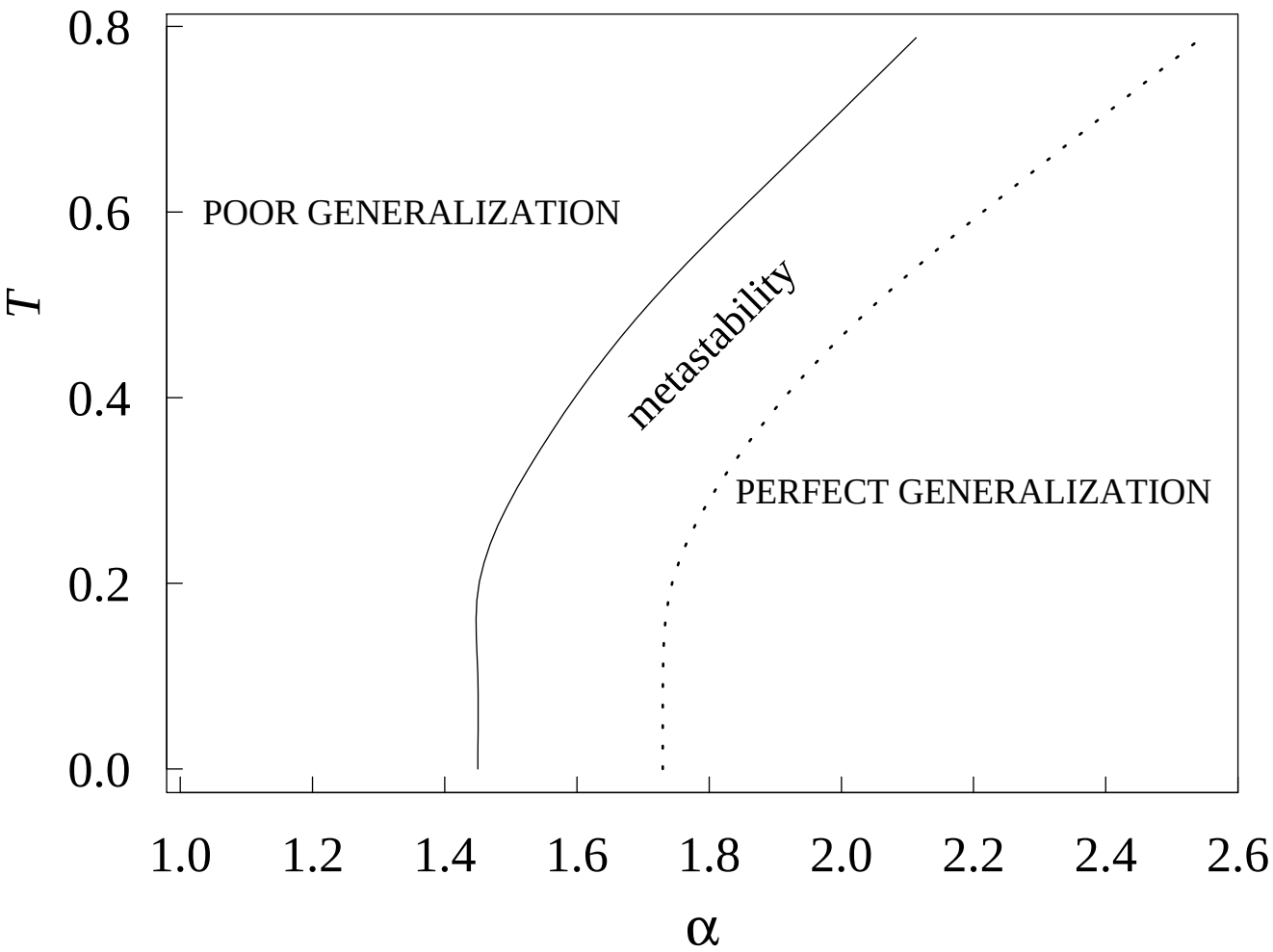
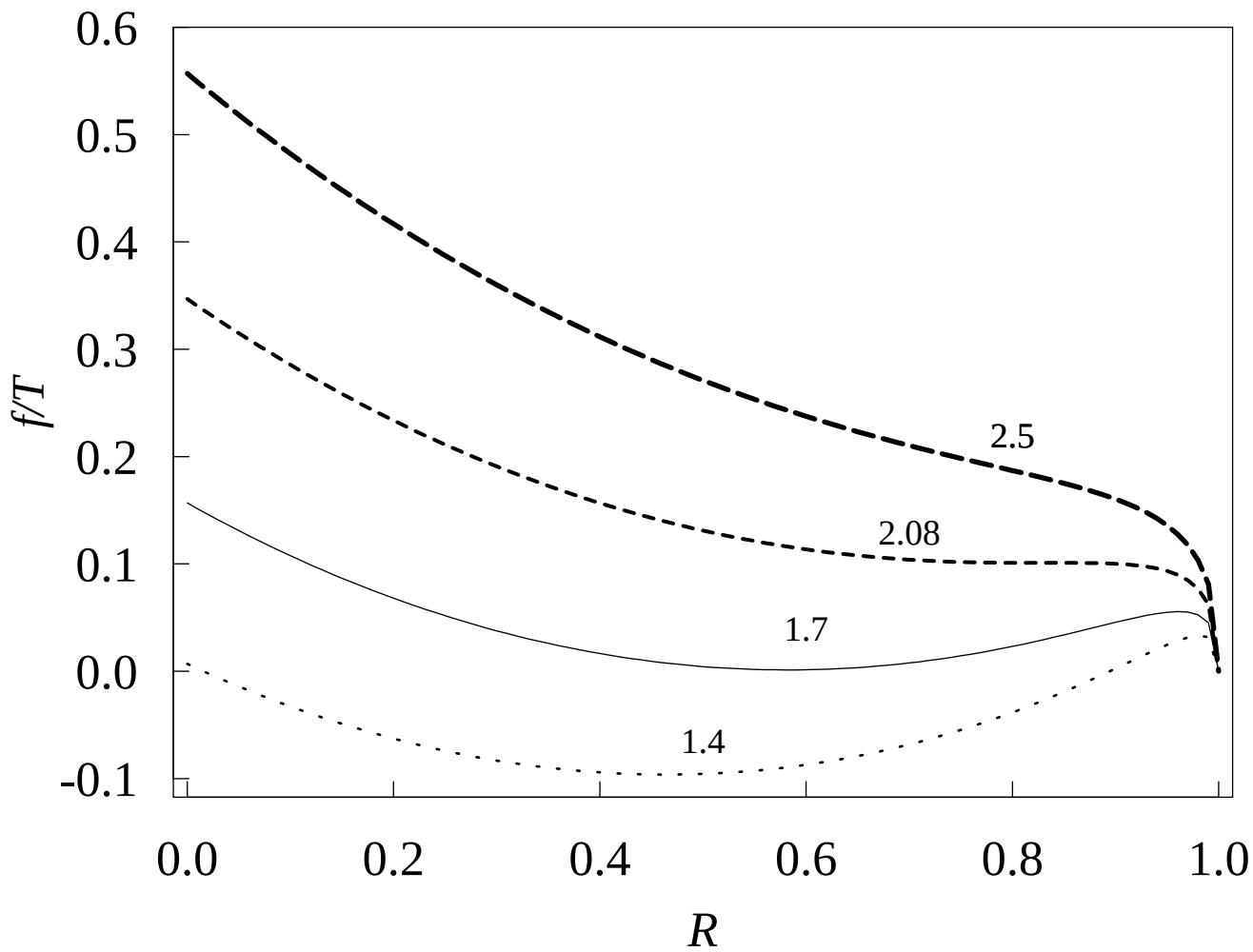
Part I: Figure 7



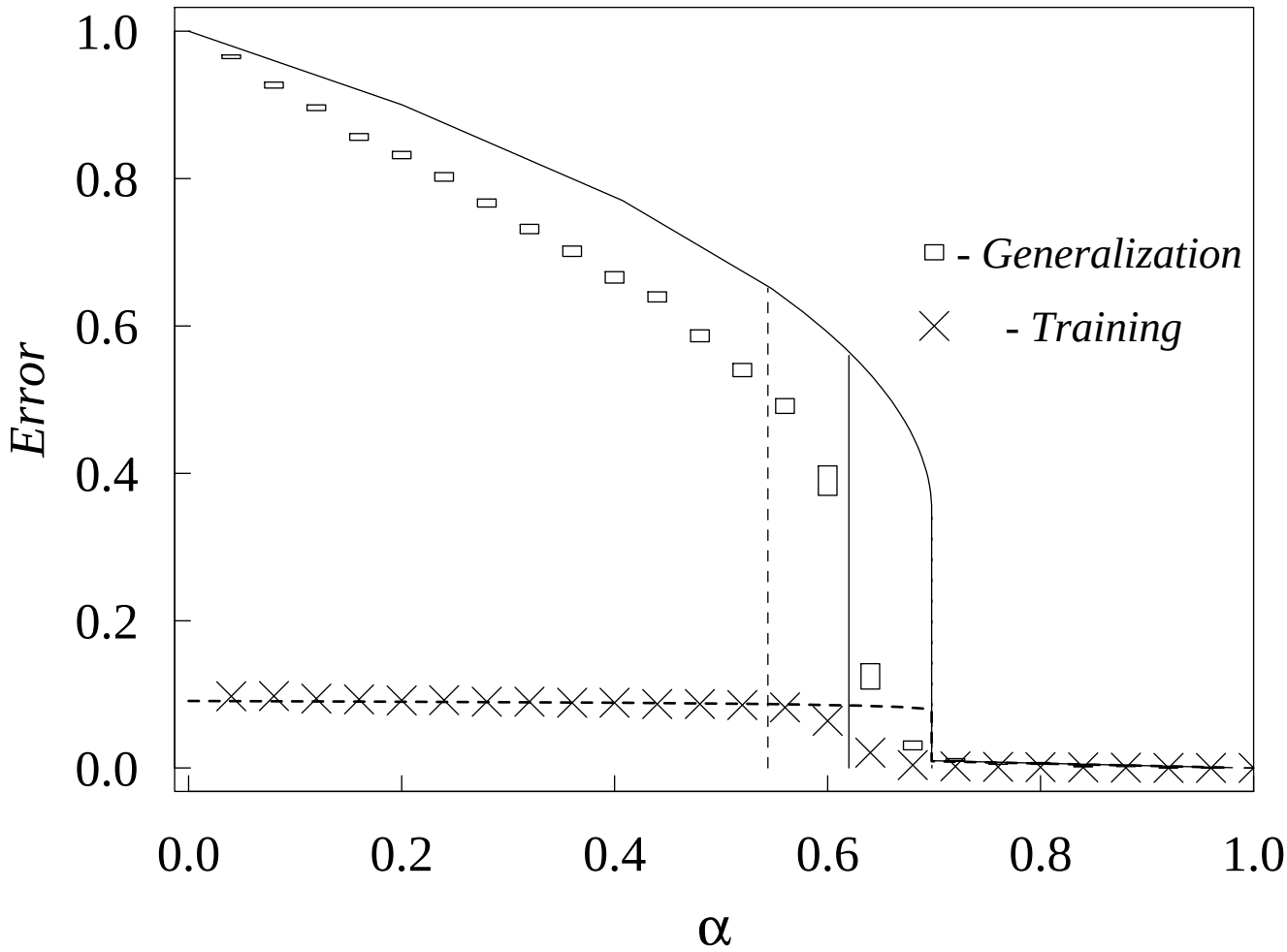
Part I: Figure 6



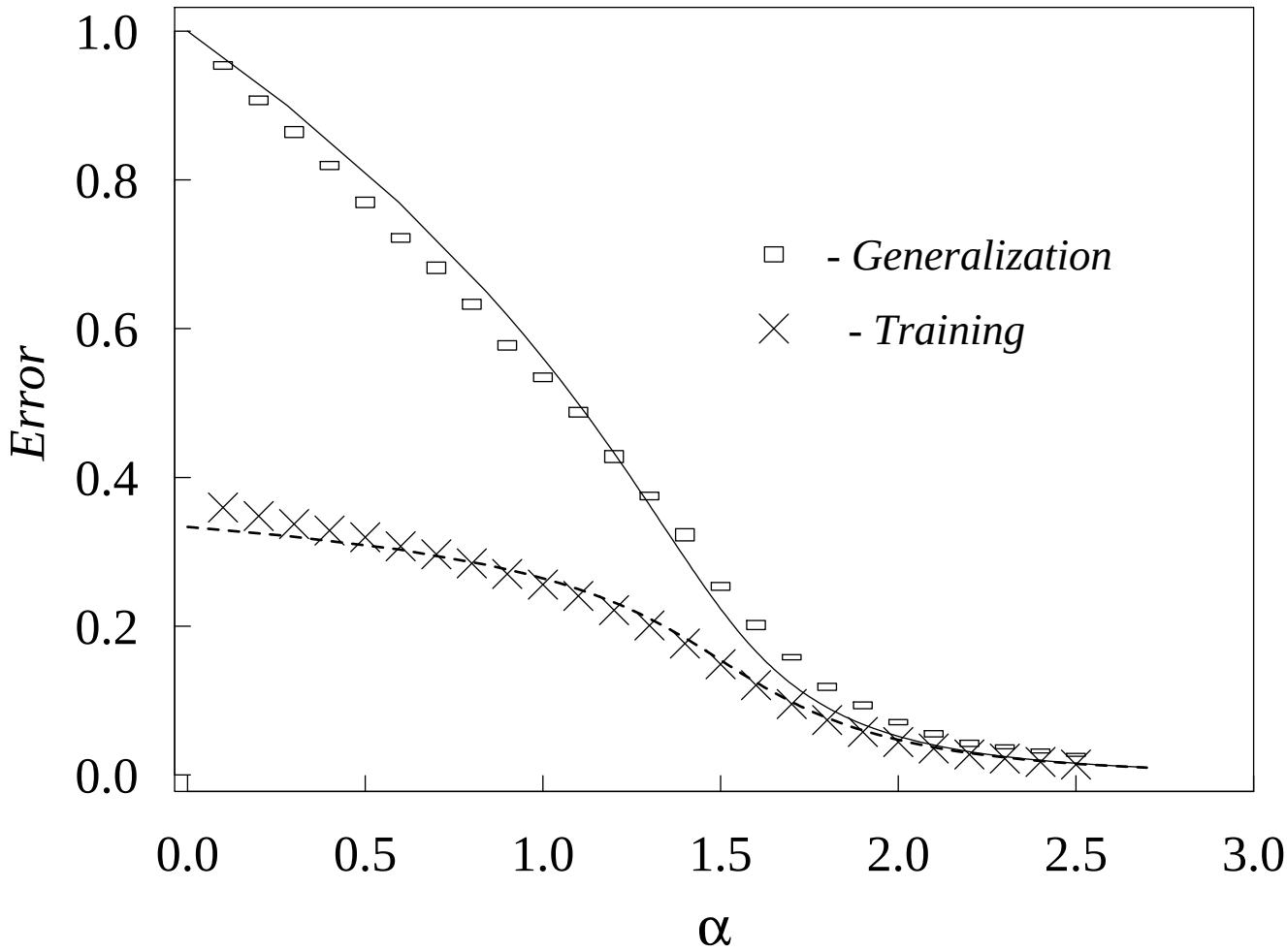
Part I: Figure 5



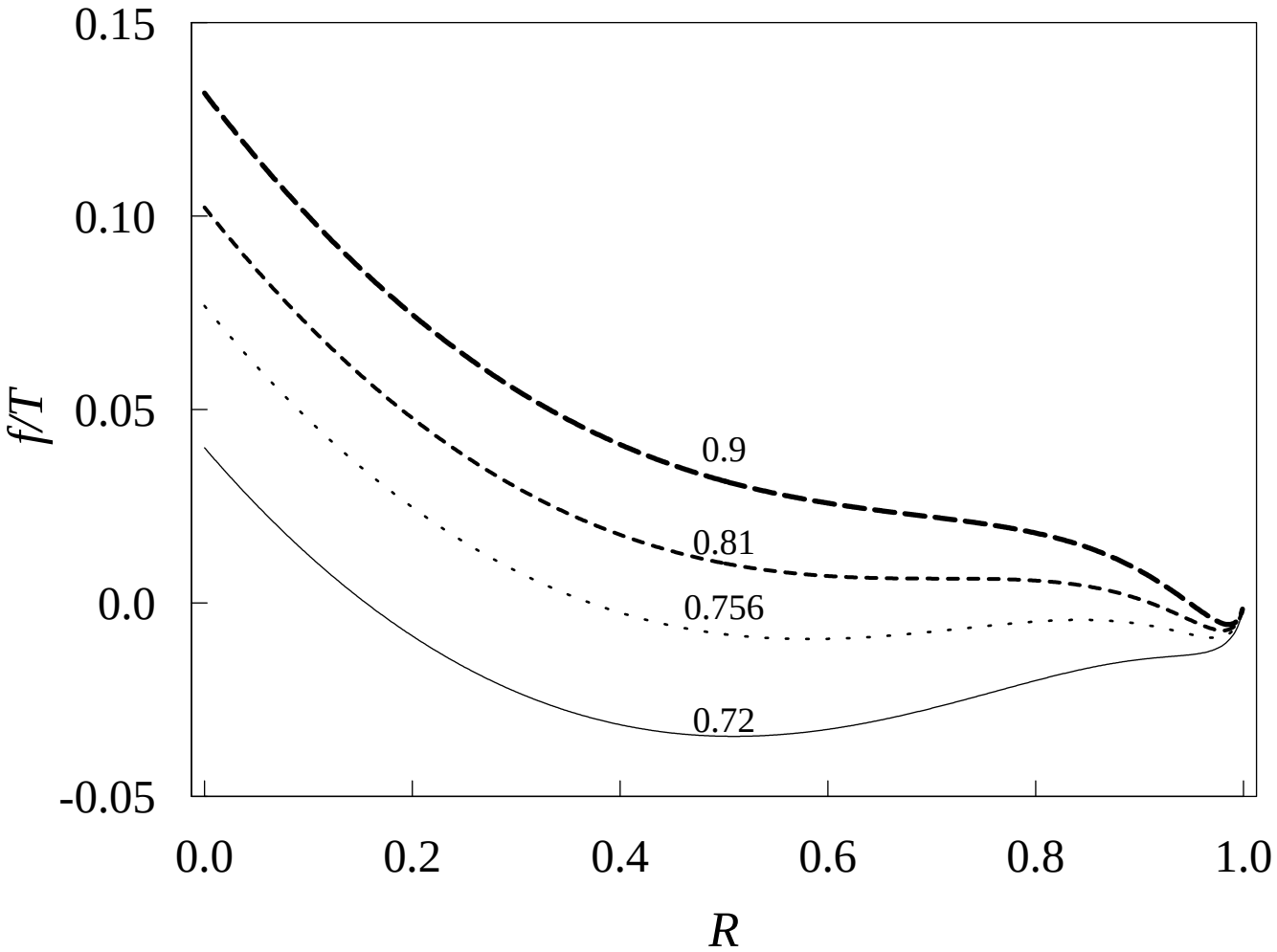
Part I: Figure 4b T=0.2



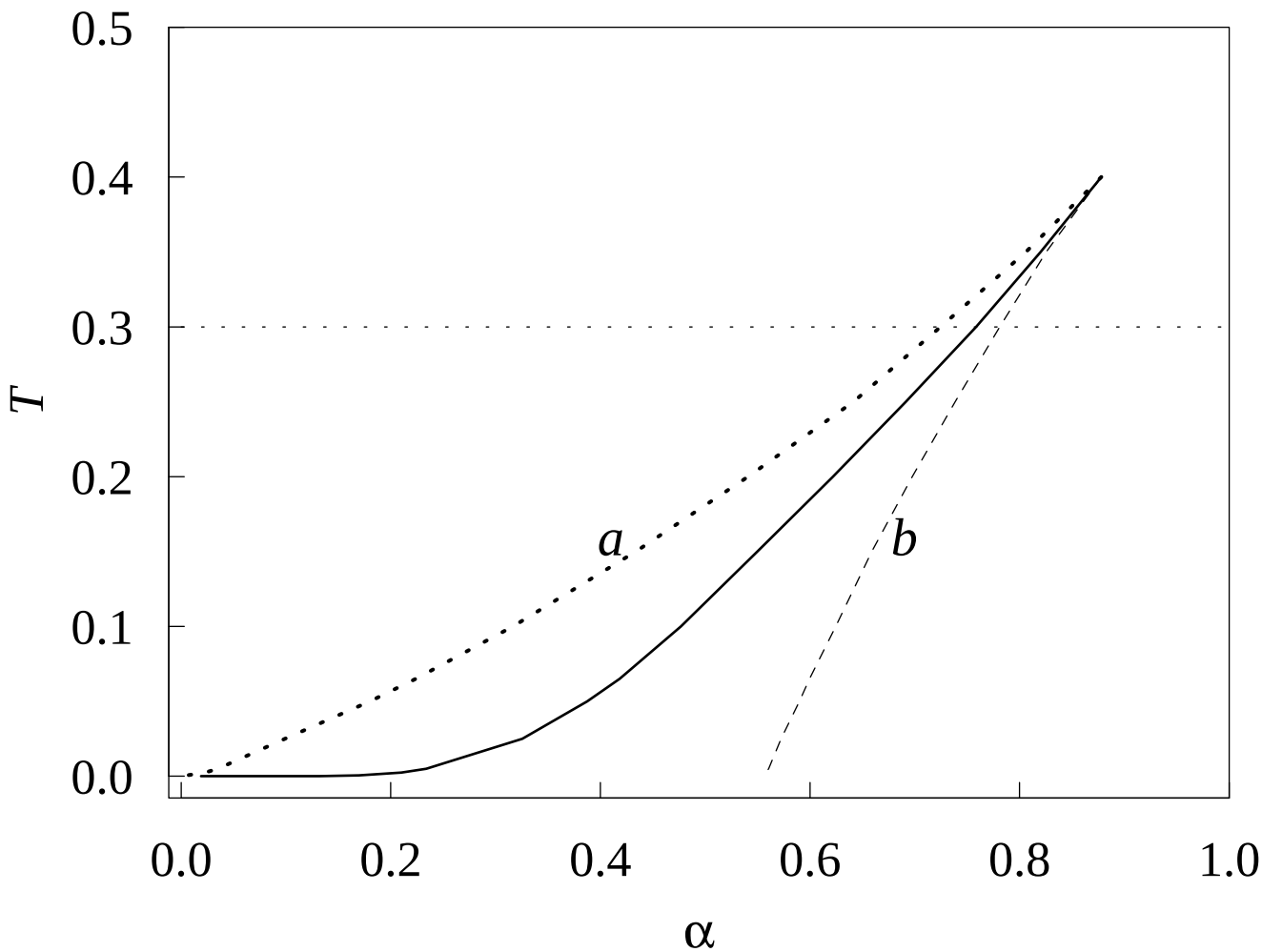
Part I: Figure 4a T=1



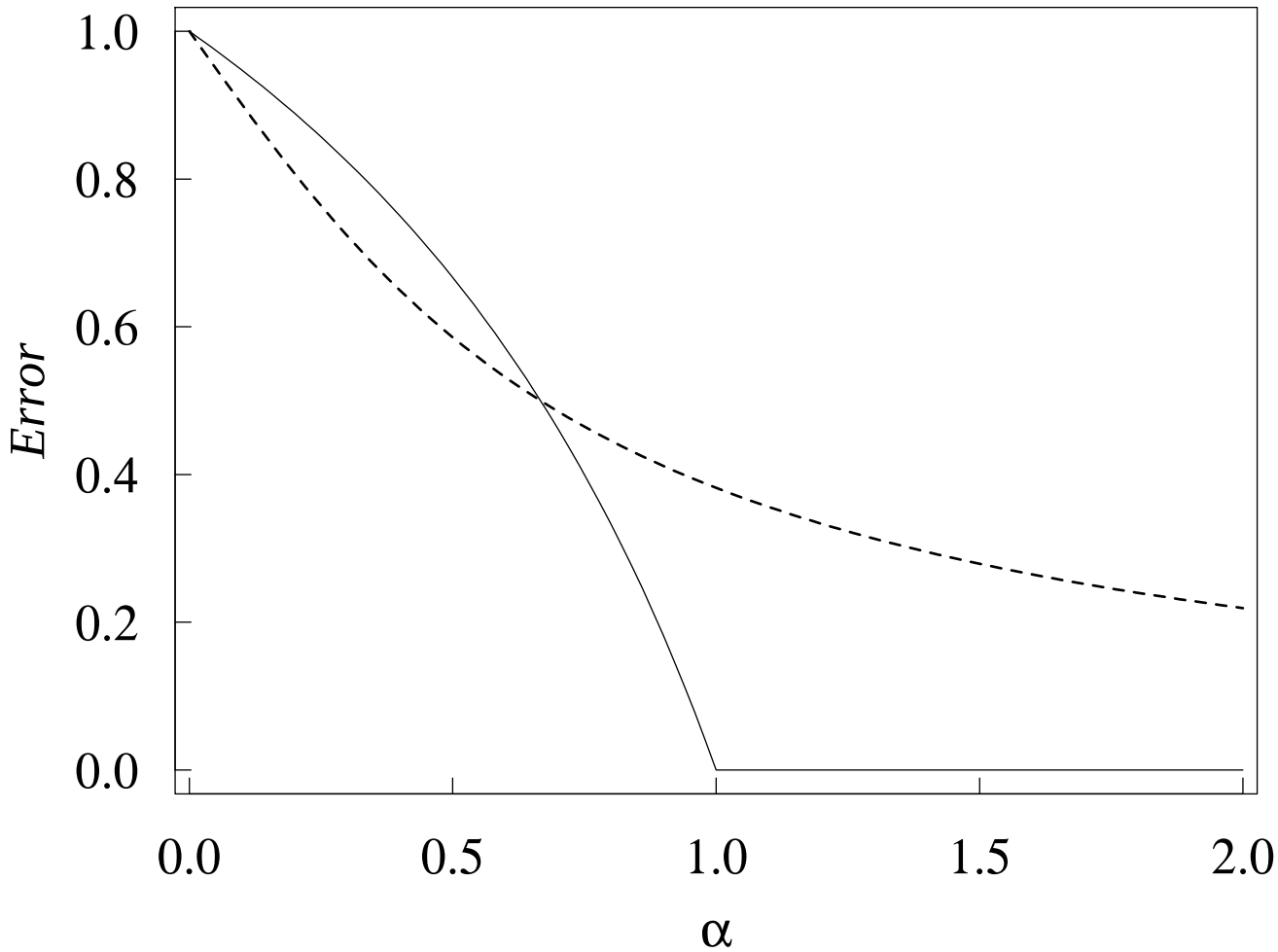
Part I: Figure 3



Part I: Figure 2



Part I: Figure 1



simulations at $T = 5.0$ for $N = 75$, 32 samples. In the simulations, the transition is smeared, occurring gradually between the spinodal and thermodynamic transition points.

FIGURES

FIG. 1. Annealed learning curves at $T = 0$ for the linear/continuous model. The generalization curve shows a second order transition to perfect learning at $\alpha_c = 1$, and the training curve is zero everywhere. Also shown (dashed line) the high T generalization curve, as a function of T/α, for comparison.

FIG. 2. Annealed phase diagram for the linear/discrete model in the (α, T) plane. The dashed lines (a) and (b) are the spinodals, which demarcate the region where there are two metastable phases. The solid line is the thermodynamic transition line, at which the two phases have equal free energies. The lines all terminate at $T = 0.40$, $\alpha = 0.87$.

FIG. 3. Annealed free energy of the linear/discrete model as a function of the overlap R at $T = 0.3$ and $\alpha < 0.730$, $\alpha = 0.730$ (spinodal), $\alpha = 0.756$ (thermodynamic transition), and $\alpha > 0.781$ (spinodal).

FIG. 4. Annealed learning curves and Monte Carlo simulations (averaged over 64 samples) for the linear/discrete model. (a) At $T = 1$ the agreement between simulations ($N = 100$) and the AA is quite good. (b) At $T = 0.2$, the AA is quantitatively inaccurate when compared with the simulated generalization error. However, the AA seems to predict the location of the transition quite accurately.

FIG. 5. High temperature limit of the free energy of the boolean/discrete model as a function of the overlap R for $\alpha/T = 1.69$ (thermodynamic transition), $\alpha/T = 2.08$ (spinodal).

FIG. 6. Annealed phase diagram for the boolean/discrete model. The dashed line is a spinodal above which the low R metastable state vanishes. The solid line is the thermodynamic transition line, at which the low R and $R = 1$ states have equal free energy.

FIG. 7. Annealed learning curves for the boolean/discrete model and Monte Carlo

$$-\beta f = \mathrm{extr}_{R,\hat{R}} \left\{ -R\hat{R} + \frac{1}{N} \ln \int d\mu(\mathbf{W})\, e^{\hat{R}\mathbf{W}\cdot\mathbf{W}^0} - \alpha G_{\mathrm{an}}(R) \right\} \tag{C12}$$

The saddle point equations are

$$\hat{R} = -\alpha \frac{\partial G_{\mathrm{an}}}{\partial R} \ , \tag{C13}$$

$$R = \frac{1}{N} \langle \mathbf{W} \rangle_{\hat{R}} \cdot \mathbf{W}^0 \ , \tag{C14}$$

where

$$\langle \mathbf{W} \rangle_{\hat{R}} \equiv \frac{\int d\mu(\mathbf{W})\, \mathbf{W} \exp(\hat{R}\mathbf{W}\cdot\mathbf{W}^0)}{\int d\mu(\mathbf{W}) \exp(\hat{R}\mathbf{W}\cdot\mathbf{W}^0)} \tag{C15}$$

The order parameter $\hat{R}$ has a natural interpretation: it is the strength of a local field pushing $\mathbf{W}$ in the direction of $\mathbf{W}^0$. Since it increases with α, it forces $\mathbf{W}$ toward $\mathbf{W}_0$ as $\alpha \to \infty$. Upon eliminating $\hat{R}$, these equations reduce to Eq. (4.16).

Equation (C11) can be evaluated quite easily for the case of the Ising constraint $W_i = \pm 1$. Then we can make the replacement

$$\int d\mu(\mathbf{W}) \to \sum_{W_i = \pm 1} \ , \tag{C16}$$

which leads finally to

$$G_0(R) = -R\hat{R} + \ln 2 \cosh \hat{R} \ , \tag{C17}$$

assuming that the teacher weights also satisfy the Ising constraint. Extremizing with respect to $\hat{R}$, we obtain the equation

$$R = \tanh \hat{R} \tag{C18}$$

Eliminating $\hat{R}$, one can finally derive the result (4.20), which is purely a function of R, and is the familiar result for the entropy of the Ising model as a function of the magnetization R.

Calculating G_0 for the spherical distribution is somewhat more complicated. We rewrite the a priori spherical distribution (4.17) as

$$d\mu(\mathbf{W}) = \prod_{i=1}^{N} \frac{dW_i}{\sqrt{2\pi e}} \int_{-i\infty}^{i\infty} \frac{d\lambda}{2\pi i} e^{\lambda(\mathbf{W}\cdot\mathbf{W} - N)} \tag{C19}$$

and then G_0 is determined as a saddle point over $\hat{R}$ and λ,

$$\begin{aligned} G_0 &= -\frac{1}{2} \ln(2\pi e) + R\hat{R} + \lambda + \ln \int dW e^{-\lambda W^2 - \hat{R}W} \\ &= -\frac{1}{2} + R\hat{R} + \lambda - \frac{1}{2} \ln(2\lambda) + \frac{\hat{R}^2}{4\lambda} \end{aligned} \tag{C20}$$

Eliminating $\hat{R}$ and λ finally yields the result (4.18), which was justified previously by geometric arguments.

of variables in the integral (C5) yields the form (4.14). Evaluating the integral for the cases of $g(x) = x$ (linear perceptron) and $g(x) = \mathrm{sign}(x)/\sqrt{2}$ (boolean perceptron), we obtain

$$G_{\mathrm{an}} = \begin{cases} \frac{1}{2}\ln\left[1 + 2\beta(1-R)\right], & \text{linear,} \\ -\ln\left[1 - (1-e^{-\beta})\pi^{-1}\cos^{-1} R\right], & \text{boolean.} \end{cases} \tag{C6}$$

Although derived using the Gaussian a priori input distribution (4.3), the above results for $\epsilon(\mathbf{W})$ and $G_{\mathrm{an}}(\mathbf{W})$ apply also to the case of the discrete inputs $S_i = \pm 1$ in the thermodynamic limit. This insensitivity to input distribution is explained by the central limit theorem, which guarantees that $\mathbf{W}\cdot\mathbf{S}/\sqrt{N}$ and $\mathbf{W}^0\cdot\mathbf{S}/\sqrt{N}$ are Gaussian variables (in the $N \to \infty$ limit) with very weak assumptions about the distribution of $\mathbf{S}$. This assertion may be verified by a straightforward calculation for discrete $\mathbf{S}$, which yields Eqs. (C6) as the leading terms in a saddle point expansion in $1/N$.

Since $G_{\mathrm{an}}(\mathbf{W})$, Eq. (3.4), depends only on the overlap R, the annealed partition function (3.3) can now be rewritten as an integral over R,

$$\langle\!\langle\, Z \,\rangle\!\rangle = \int dR \exp N[G_0(R) - \alpha G_{\mathrm{an}}(R)] \ , \tag{C7}$$

where

$$N G_0(R) = \ln \int d\mu(\mathbf{W})\, \delta(R - N^{-1}\mathbf{W}\cdot\mathbf{W}^0) \tag{C8}$$

is the logarithm of the density of networks with overlap R. In the thermodynamic limit ($N \to \infty$), the integral can be evaluated as

$$-\beta f(T,\alpha) \equiv \frac{1}{N}\ln\langle\!\langle\, Z \,\rangle\!\rangle = \mathrm{extr}_R\{G_0(R) - \alpha G_{\mathrm{an}}(R)\} \ . \tag{C9}$$

Hence the thermodynamic free energy $f(T,\alpha)$ is determined by extremizing the free energy function, $-\beta f(R) \equiv G_0(R) - \alpha G_{\mathrm{an}}(R)$. Differentiating $f(R)$ with respect to R, we obtain the stationarity condition Eq. (4.16).

We can write the stationarity equations in a more revealing form by proceeding further in the evaluation of G_0. The delta function in (C8) can be expanded by introducing another order parameter $\hat{R}$,

$$G_0 = \frac{1}{N}\ln \int_{-i\infty}^{i\infty} \frac{d\hat{R}}{2\pi i} \exp\left(-NR\hat{R} + \ln \int d\mu(\mathbf{W})\, e^{\hat{R}\mathbf{W}\cdot\mathbf{W}^0}\right) \ . \tag{C10}$$

In the thermodynamic limit, this reduces to

$$G_0 = -R\hat{R} + \frac{1}{N}\ln \int d\mu(\mathbf{W})\, e^{\hat{R}\mathbf{W}\cdot\mathbf{W}^0} \ , \tag{C11}$$

where the right hand side must be stationary with respect to the saddle point parameter $\hat{R}$. The free energy can now be written as a saddle point over two order parameters

APPENDIX C: ANNEALED APPROXIMATION FOR PERCEPTRON LEARNING

In this appendix, we give a fuller account of the results that were outlined in Sections IV.A and B. We begin with a derivation of Eq. (4.7) for the average generalization error, which illustrates many of the calculational techniques of this paper. Integrating the error function (4.6) over the a priori input measure (4.3), we obtain

$$\begin{aligned}\epsilon(\mathbf{W}) &= \int D\mathbf{S}\, \epsilon(\mathbf{W};\mathbf{S}) \\ &= \int dx \int dy\, \frac{1}{2}[g(x)-g(y)]^2 \qquad \text{(C1)} \\ &\qquad \times \int D\mathbf{S}\, \delta\left(x - N^{-\frac{1}{2}}\mathbf{W}\cdot\mathbf{S}\right)\delta\left(y - N^{-\frac{1}{2}}\mathbf{W}^0\cdot\mathbf{S}\right) \\ &= \int \frac{dx\, d\hat{x}}{2\pi}\int \frac{dy\, d\hat{y}}{2\pi} e^{ix\hat{x}+iy\hat{y}}\frac{1}{2}[g(x)-g(y)]^2 \\ &\qquad \times \int D\mathbf{S} \exp\left[-iN^{-\frac{1}{2}}(\mathbf{W}\hat{x}+\mathbf{W}^0\hat{y})\cdot\mathbf{S}\right] \ . \qquad \text{(C2)}\end{aligned}$$

The two auxiliary variables x, y are introduced to remove $\mathbf{S}$ from the argument of the g functions, and $\hat{x}, \hat{y}$ are introduced to transform the delta functions into exponentials using the identity

$$\delta(x) = \int \frac{d\hat{x}}{2\pi} e^{ix\hat{x}} \ . \qquad \text{(C3)}$$

When the Gaussian average over $\mathbf{S}$ is now performed, a simple Gaussian integral in $\hat{x}$ and $\hat{y}$ is left. These variables can in turn be integrated out, leaving

$$\epsilon(\mathbf{W}) = \int \frac{dx\, dy}{2\pi\sqrt{1-R^2}} \exp\left(-\frac{x^2+y^2-2xyR}{2(1-R^2)}\right)\frac{1}{2}[g(x)-g(y)]^2 \qquad \text{(C4)}$$

This result has a simple interpretation. It is the average of $\frac{1}{2}[g(x)-g(y)]^2$, where x and y, like $\mathbf{W}\cdot\mathbf{S}/\sqrt{N}$ and $\mathbf{W}^0\cdot\mathbf{S}/\sqrt{N}$ in (4.6), are Gaussian variables with unit variance and cross-correlation R. A simple change of variables in Eq. (C4), a shift followed by a rescaling, yields the form (4.7).

To derive the annealed approximation for perceptron learning, we begin by evaluating Eq. (3.4) for $G_{\rm an}(\mathbf{W})$. The calculation is essentially the same as the previous one for $\epsilon(\mathbf{W})$, and results in a similar formula,

$$\begin{aligned}G_{\rm an}(\mathbf{W}) = -\ln \int \frac{dx\, dy}{2\pi\sqrt{1-R^2}} \exp\left(-\frac{x^2+y^2-2xyR}{2(1-R^2)}\right) \\ \times \exp(-\beta/2[g(x)-g(y)]^2) \ . \qquad \text{(C5)}\end{aligned}$$

Again, the answer depends on $\mathbf{W}$ only through the overlap R. We denote the function of R on the right hand side by $G_{\rm an}$, so that $G_{\rm an}(\mathbf{W}) \equiv G_{\rm an}(N^{-1}\mathbf{W}\cdot\mathbf{W}^0)$. A change

If δE were a quantity of order unity, the cumulants would be functions only of $\alpha\beta$, the only parameter governing the distribution $e^{-\beta E_0}$. Hence Eq. (B6) would be the desired power series in β. In fact, the situation is somewhat more complicated because δE scales like α.

To investigate the scaling of the cumulants with α, we write them as sums over connected correlation functions

$$C_j = \sum_{\mu_1,\ldots,\mu_j=1}^{P} \langle \delta\epsilon_{\mu_1} \cdots \delta\epsilon_{\mu_j} \rangle_c \tag{B10}$$

Counting the P^j terms in this sum, the naive estimate would be that $C_j \sim P^j$. However, the quenched average makes any term containing an unrepeated index vanish, since $\langle\langle\, \delta\epsilon_\mu \,\rangle\rangle = 0$. In other words, if an index appears, it must appear at least twice for the term to be nonvanishing. This means that the cumulants can scale no faster than $C_j \sim P^{\lfloor j/2 \rfloor}$. At the same time, there are also factors of N which make the cumulant extensive, so that it behaves like

$$C_j \sim N\alpha^{\lfloor j/2 \rfloor} c_j(\alpha\beta) \ , \tag{B11}$$

to leading order in α.

In the expansion for the free energy, the term containing C_j only contributes to terms of order $\beta^{\lceil j/2 \rceil}$ or higher in the ultimate high temperature expansion. Hence the free energy can be written in the form

$$-\beta F = \ln Z_0 + \sum_{j=1}^{\infty} \beta^j F_j(\alpha\beta) \tag{B12}$$

where F_j contains contributions from the finite number of cumulants $C_1, \ldots, C_{2j}$.

In general, any quantity A that is finite in the high-T limit possesses a high temperature expansion of the form

$$A(\beta, T/\alpha) = \sum_{i=0}^{\infty} \beta^i A_i(T/\alpha) \tag{B13}$$

In the high temperature limit ($\beta \to 0$, $T/\alpha = const$),

$$A(\beta, T/\alpha) \to A_0(T/\alpha) \ , \tag{B14}$$

which depends only on the effective temperature T/α. In general, the functions A_i can be nonanalytic functions of T/α. In particular, the high-T limit term A_0 can be nontrivial and lead to such behavior as the first order transition in the boolean/discrete model.

In particular, for $\gamma = \beta$, $\langle\!\langle\, \Delta \,\rangle\!\rangle = P(\epsilon_t - \epsilon_g)$, so that we finally obtain

$$\epsilon_t(T,\alpha) < \epsilon_g(T,\alpha) \ , \tag{A4}$$

which was stated without proof in Eq. (2.20).

APPENDIX B: HIGH TEMPERATURE EXPANSION

In this appendix, we show using a cumulant expansion that the free energy (2.13) can be written as a power series in β, with coefficients that are functions of $\alpha\beta$. The zeroth order term of this series is the high temperature limit discussed in Section II.C.

The first step is to separate the energy into random and nonrandom parts. The nonrandom part is

$$E_0 \equiv \langle\!\langle\, E \,\rangle\!\rangle = P\epsilon(\mathbf{W}) = N\alpha\epsilon(\mathbf{W}) \ , \tag{B1}$$

and the random part is

$$\delta E \equiv E - E_0 = \sum_{\mu=1}^{P} \delta\epsilon_\mu \ , \tag{B2}$$

where

$$\delta\epsilon_\mu \equiv \epsilon(\mathbf{W};\mathbf{S}^\mu) - \epsilon(\mathbf{W}) \tag{B3}$$

We now treat δE as a perturbation to the "free Hamiltonian" E_0. The partition function takes the form

$$Z = Z_0 \langle e^{-\beta\delta E}\rangle_0 \ , \tag{B4}$$

Here $\langle\cdots\rangle_0$ denotes the average with respect to the distribution $Z_0^{-1}e^{-\beta E_0}$. The factor

$$Z_0(\beta\alpha) = \int d\mu(\mathbf{W}) e^{-\beta E_0} = \int d\mu(\mathbf{W}) e^{-N\beta\alpha\epsilon(\mathbf{W})} \ , \tag{B5}$$

is the high-T partition function introduced in Section II.C.

Taking the logarithm of both sides and performing the quenched average, we obtain

$$-\beta F = \ln Z_0(\alpha\beta) + \sum_{j=1}^{\infty} \frac{(-\beta)^j}{j!} \langle\!\langle\, C_j \,\rangle\!\rangle \tag{B6}$$

where the C_j are from the cumulant expansion of $\ln\langle e^{-\beta\delta E}\rangle_0$

$$C_1 = \langle\delta E\rangle_0 \tag{B7}$$

$$C_2 = \langle(\delta E)^2\rangle_0 - \langle\delta E\rangle_0^2 \tag{B8}$$

$$C_3 = \ldots \tag{B9}$$

[38] H. Sompolinsky, N. Tishby, and H.S. Seung. Learning from examples in large neural networks. *Phys. Rev. Lett.*, 65(13):1683–1686, 1990.
[39] G. Györgyi. Inference of a rule by a neural network with thermal noise. *Phys. Rev. Lett.*, 64:2957–2960, 1990.
[40] K. Binder and D. W. Heerman. *Monte Carlo Simulation in Statistical Mechanics.* Springer-Verlag, Berlin, 1988.
[41] S. Kirkpatrick, Jr. C.D. Gelatt, and M.P. Vecchi. Optimization by simulated annealing. *Science*, 220:671–680, 1983.
[42] F. Rosenblatt. *Principles of Neurodynamics.* Spartan Books, New York, 1962.
[43] A. Krogh and J. A. Hertz. Dynamics of generalization in linear perceptrons. In *Advances in Neural Information Processing Systems 3*, 1991.
[44] M. L. Minsky and S. Papert. *Perceptrons.* MIT Press, Cambridge, 1988.
[45] H. Sompolinsky et al, *Preprint.*
[46] T.M. Cover. Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition. *IEEE Transactions on Electronic Computers*, 14:326–334, 1965.
[47] E. Gardner. The space of interactions in neural network models. *J. Phys.*, A21:257–270, 1988.
[48] E. Gardner and B. Derrida. Optimal storage properties of neural network models. *J. Phys.*, A21:271–284, 1988.
[49] E. Barkai, D. Hansel, and I. Kanter. Statistical mechanics of a multilayered neural network. *Phys. Rev. Lett.*, 65:2312–2315, 1990.

APPENDIX A: THERMODYNAMIC INEQUALITIES

Let E_0 denote the average value of the training energy $P\epsilon(\mathbf{W})$, as in Eq. (2.18) and δE the difference $E - E_0$. Define the function

$$\Delta(\beta,\gamma) \equiv \frac{\int d\mu(\mathbf{W}) e^{-\gamma E_0 - \beta\delta E}\delta E}{\int d\mu(\mathbf{W}) e^{-\gamma E_0 - \beta\delta E}} \ . \tag{A1}$$

This can be interpreted as the energy of a system at temperature $1/\beta$ with Hamiltonian δE and measure $d\mu(\mathbf{W})e^{-\gamma E_0}$. By the convexity of the free energy, the energy is a decreasing function of β. Hence Δ is bounded above,

$$\Delta(\beta,\gamma) \leq \Delta(\beta=0,\gamma) = \frac{\int d\mu(\mathbf{W}) e^{-\gamma E_0}\delta E}{\int d\mu(\mathbf{W}) e^{-\gamma E_0}} \ . \tag{A2}$$

We now take the quenched average of both sides. In the integrals, the average can be applied directly to δE, since E_0 is independent of the quenched disorder. But $\langle\!\langle\, \delta E \,\rangle\!\rangle = 0$, so the right hand side vanishes, and we have

$$\langle\!\langle\, \Delta(\beta,\gamma) \,\rangle\!\rangle \leq \langle\!\langle\, \Delta(\beta=0,\gamma) \,\rangle\!\rangle = 0 \ . \tag{A3}$$

[17] A.R. Barron. Statistical properties of artificial neural networks. In *28th Conference on Decision and Control*, 1989.

[18] Luc Devroye. Automatic pattern recognition: a study of the probability of error. *IEEE Trans. on Pattern Anal. and Mach. Intelligence*, 1988.

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

[20] 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.

[21] 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.

[22] E. Levin, N. Tishby, and S.A. Solla. A statistical approach to learning and generalization in layered neural networks. *Proceedings of the IEEE—Special Issue on Neural Networks*, 1990.

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

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

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

[26] 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.

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

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

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

[30] D.B. Schwartz, V.K. Samalam, J.S. Denker, and S.A. Solla. Exhaustive learning. *Neural Computation*, 2:374–385, 1990.

[31] T. Hogg J. Shrager and B.A. Huberman. *Science*, 242:414, 1988.

[32] S. Ahmed and G. Tesauro. In D.S. Touretzky, editor, *Advances in Neural Information Processing Systems I*, pages 160–164, San Mateo, CA, 1989. Morgan Kaufmann.

[33] E.L. Thorndike. *Fundamentals of Learning.* Columbia University Press, New York, 1932.

[34] W. Kohler. *Gestalt Psychology.* Liveright, New York, 1929.

[35] H. Sompolinsky and N. Tishby. Learning in a two layer network of edge detectors. *Europhysics Letters*, 1990.

[36] W. Krauth and M. Mézard. Storage capacity of memory networks with binary couplings. *J. Phys. France*, 50:3057–3066, 1989.

[37] S.S. Venkatesh. On learning binary weights. To be in COLT91.

REFERENCES

[1] Richard P. Lippman. Review of neural networks for speech recognition. *Neural Computation*, 1:1–38, 1989.

[2] T.J. Sejnowski and C. Rosenberg. Parallel networks that learn to pronounce english text. *Complex Systems*, 1:145–168, 1987.

[3] Y. LeCun, B. Boser, J.S. Denker, D. Henderson, R.E. Howard, W. Hubbard, and L.D. Jackel. Backpropagation applied to handwritten zip code recognition. *Neural Computation*, 1:541–551, 1989.

[4] N. Qian and T.J. Sejnowski. Predicting the secondary structure of globular proteins using neural network models. *Journal of Molecular Biology*, 202:865–884, 1988.

[5] D. Zipser and R.A. Andersen. A back-propagation programmed network that simulates response properties of a subset of posterior parietal neurons. *Nature*, 331:679–684, 1988.

[6] S.R. Lockery, G. Wittenberg, Jr. W.B. Kristan, and W.G. Cottrell. Function of identified interneurons in the leech elucidated using neural networks trained by back-propagation. *Nature*, 340:468–471, 1989.

[7] D.E. Rumelhart and J.L. McClelland, editors. *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, volume 1: Foundations. MIT Press, Cambridge, MA, 1986.

[8] T. Poggio and F. Girosi. A theory of networks for approximation and learninig. AI Memo 1140, MIT, Cambridge, MA, 1989.

[9] L.G. Valiant. A theory of the learnable. *Communications of the ACM*, 27:1134–1142, 1984.

[10] A. Blumer, A. Ehrenfeucht, D. Haussler, and M.K. Warmuth. Learnability and the vapnik-chervonenkis dimension. *JACM*, 36:929–965, 1989.

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

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

[13] S. Judd. On the complexity of loading shallow neural networks. *Journal of Complexity*, 1:177–192, 1988.

[14] V.N. Vapnik and A.Y. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and its Applications*, 1971.

[15] V.N. Vapnik. *Estimation of Dependences based on Empirical Data.* Springer-Verlag, New York, 1982.

[16] Y.S. Abu-Mostafa. The vapnik-chervonenkis dimension: Information versus complexity in learning. *Neural Computation*, 1:312–317, 1989.

layers, is allowed to vary significantly. The important issue of learning under these circumstances will be discussed elsewhere.

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.

An example of this case is the perceptron with discrete weights and linear output, see Eq. (4.37). When $x > 1$, $\partial s/\partial \epsilon$ remains finite at small ϵ, so there must be a discontinuous transition at some $(\alpha\beta)_c$ to $\epsilon_g = 0$,

$$\epsilon_g = 0 \ , \ \alpha\beta > (\alpha\beta)_c \ , x > 1 \ . \tag{5.4}$$

The boolean perceptron with discrete weights is an example of an entropy with $x = 2$, see Eqs. (4.46)–(4.47).

The above classification is based on the high-T theory. Although extrapolating to finite T is nontrivial, we believe it is possible for large classes of realizable rules. In our perceptron models, nontrivial shapes of learning curves resulted only when the weights were constrained to discrete values. In contrast, for multilayer networks we expect that the exponent x in Eq. (5.1) may be nontrivial even when the weights are allowed to vary continuously, possibly leading to Eqs. (5.3) or (5.4). This will occur, for example, if the optimal solution involves *discrete* internal representations of the hidden neurons, as demonstrated in the two-layer model network for the contiguity problem[35].

The above classification also requires knowledge of the behavior of $s(\epsilon)$ near $\epsilon = 0$. This behavior can be calculated for toy models, where the target rule is known, but is generally impossible to calculate in real-world applications of learning. Even in well-defined simple models, evaluating s may be difficult, since the primary interest is its behavior in the limit $N \to \infty$. For this reason, the simple existence of a gap in the spectrum of ϵ is an insufficient basis for classification. At the very least, the scaling of the gap with N must be determined (see Sec. IV.F). Nevertheless, our results may provide useful hints for understanding or even predicting the behavior of real-world learning problems.

E. Uniqueness of the Optimal Solution

Throughout this and the following paper we have assumed that the optimal network weights, i.e. the components of the weight vector $\mathbf{W}^*$ that globally minimizes the generalization function $\epsilon(\mathbf{W})$, are unique. If they are nonunique, we assume that the degenerate global minima are at least widely separated in network space, forming a discrete set. This is related to our assumption that the training dynamics searches for the values of the network weights within a well-defined architecture. Furthermore, in all the models studied in Part I (except that of Sec. IV.G), the target rule is perfectly realizable within the imposed architecture. Under these conditions it is reasonable to expect that the optimal solution is generally unique up to obvious symmetries. For instance, in multilayer networks the solution may be unique only up to permutation of the hidden neurons. Other symmetries may result from the nature of the rule itself. An example is the updown degeneracy of the optimal solution of the two layer network of edge detectors in Ref. [35]. This uniqueness does not hold if the architecture of the network, e.g. the number of neurons or the number of

C. Results of the Perceptron Models

The results of the specific models studied here indicate that the shapes of learning curves may be very different from the PAC Learning bound of an inverse power law. The shape of the generalization curve depends strongly on the degree of constraint imposed on the network space. In the case of a linear output, the imposition of binary constraints changes a T/α tail (Eq. 4.27) into an exponential $\exp(-2\alpha/T)$ (Eq. 4.37). In addition, at low T there is a discontinuous transition at finite α from poor to good generalization, see Figs. 2 and 4. The superior generalization ability of the constrained network is not surprising. Since the target rule itself was assumed to be realizable in the constrained architecture, imposing the constraints is essentially using prior knowledge (or assumptions) of the nature of the rule to restrict the space of possible networks.

The most dramatic effect of constraining the network is found for the conventional perceptron rule, i.e., one with a Boolean output. In the unconstrained network the generalization error again behaves as an inverse power law (this time even at $T = 0$, Eq. 4.33). On the other hand, in the case of binary weights there is a discontinuous transition at a critical α from poor to perfect learning, Fig. 6. This transition is unique in that it exists even in learning at high temperatures. The collapse of a thermodynamic system at finite T to its ground state above some α_c stems from the singular spectrum of excitation above this state, as discussed in Sec. IV.F.

D. General Classification of Possible Learning Curves

The discontinuous behavior of the Boolean perceptron with discrete weights calls for an understanding of the general conditions which determine whether a given learning task will be achieved gradually or not. Insight into this question is provided by the high-T theory of Sec. II.C. According to Eqs. (2.26)–(2.28), a system can be classified according to the behavior of its entropy $s(\epsilon)$. Recall that $e^{Ns(\epsilon)}$ is defined as the number of networks that yield a given generalization error ϵ. The general form of $s(\epsilon)$ for small ϵ is expected to be

$$s(\epsilon) \approx \epsilon^x \ln \epsilon \ , \epsilon \to 0 \ . \tag{5.1}$$

For a smooth weight space we expect s to diverge logarithmically to $-\infty$, i.e. $x = 0$. This naturally leads via Eq. (2.28) to the inverse power law. For a discrete weight space, we expect the entropy to approach a finite value, i.e. $x > 0$. When $0 < x < 1$ the generalization curve obeys a nontrivial power law

$$\epsilon_g \approx (\alpha\beta)^{-1/(1-x)} \ , \ \alpha\beta \to \infty \ , \ 0 < x < 1 \ . \tag{5.2}$$

When $x = 1$ this law turns into an exponential,

$$\ln \epsilon_g \propto -\alpha\beta \ , \ \alpha\beta \to \infty \ , \ x = 1 \ . \tag{5.3}$$

on T and α is only through the effective temperature T/α. Even in this simple approximation, perceptron models exhibit a rich spectrum of generalization behaviors.

The second approximation, the annealed approximation (AA), reduces to the proper high-T limit, but deviates from it significantly at finite α and T, where the behavior is no longer a function of only the ratio of the two parameters.

In all four realizable perceptron rules studied here, these approximations have predicted correctly the shapes of the learning curves at finite T and large α. Furthermore the AA has yielded interesting results at finite T and α that are qualitatively correct. For instance, the first order transitions predicted by the AA for the perceptrons with discrete weights are clearly observed in the simulations, Figs. 4 and 7, and are in agreement with the full quenched theory, to be derived in Part II.

On the basis of our general arguments in Secs. II and III, we expect that these approximations will also hold for realizable rules in the more complex cases of multilayer networks. This is borne out in recent studies of two-layer networks of local feature detectors[35], and other multilayer systems[45]. Thus these approximations provide powerful theoretical tools for the study of learning from examples, at least for realizable rules.

Our treatment should be contrasted with the difficult problems of the capacity of single- and multilayer networks[46–48]. The capacity problems usually deal with loading random sets of data. In this case the system is highly frustrated and one has to employ the complex methods of spin-glass theory, such as replica symmetry breaking[36, 49]. In the learning problems of the present work, the training set consists of very structured data, generated by a well-defined rule. The underlying structure is represented in our formulation by the generalization function $\epsilon(\mathbf{W})$, Eq. (2.4), which dominates the behavior of the system at least when the number of examples is large. This function would be completely flat, and therefore meaningless, in the case of random training sets.

In fact the high-T approximation ignores completely the fluctuations due to finite sampling of the rule. It is thus useful in studying how learning is affected by the nature of the target rule ($\epsilon(\mathbf{W})$ as well as by the network architecture (e.g., the entropy $s(\epsilon)$, Eq. (2.27). As has been explained in Section III the AA does take into account the effect of randomness induced by the examples, though only approximately. The reasonable results generated by the AA even at finite α and T imply that this randomness may not have major effects (e.g., frustration and other spin-glass phenomena), at least in models where the chosen architecture is compatible with the rule to be learned. This may explain why, in many applications of supervised learning, simple local gradient algorithms seem to yield good results.

It should be emphasized that the above approximations may be useful not only for studying specific "toy" models but also in generating general approximate predictions that could perhaps be useful in applied research on neural networks. An example is Eq. (3.13), which provides a way of estimating the generalization error from the observed training error for a multilayer network with a boolean output.

determines the quality of the approximation. In the present case we can obtain a measure of this distortion by calculating the average of an input vector, e.g. $\langle \mathbf{S}^1 \rangle_{\rm an}$. In the a priori Gaussian distribution of inputs (4.3), the average value of an input is of course zero. On the other hand evaluating this average using the a posteriori Gibbs distribution (3.16), implied by the AA. we find

$$\langle \mathbf{S} \rangle = -\frac{1}{\sqrt{N}} \frac{\beta\theta(1-R)}{1+2\beta(1-R)} \mathbf{W}^0 \ . \tag{4.59}$$

implying that the inputs are biased towards $-\theta\mathbf{W}^0$. It is important to note that the magnitude of the bias is small—down by $\sqrt{N}$ from the magnitude of $\mathbf{S}$. However, it is enough to push $\mathbf{W}$ towards $-\mathbf{W}^0$, leading to overlaps as low as $R = -1$. Finally, note that in the realizable case $\theta = 0$, the average of $\mathbf{S}$ is zero. However even in this case the second moments of $\mathbf{S}$ are in general distorted, except when $R = 1$. This is consistent with the general expectation that the distortions of the input distribution, implied by the AA, are relatively small in realizable rules. But they vanish only in the limit of $\alpha \to \infty$.

V. DISCUSSION

A. Learning at Finite Temperature

In this paper we have studied the process of learning from examples with a stochastic training dynamics. The level of noise in the dynamics is denoted by the temperature T. One of the most important results of our analysis is that learning at finite temperature is possible, and sometimes advantageous. For any finite T, as the number of examples increases, the network weights approach their optimal values, namely the values that minimize the generalization error. In Part I we have focused mainly on realizable rules. When trained with a fixed number of examples, our realizable models have average generalization errors that increase with T. Thus from purely equilibrium considerations working at $T = 0$ is better. On the other hand, the lower the temperature the longer it may take to reach equilibrium. This is particularly true for highly nonlinear models, such as the boolean perceptron with discrete weights. Although the critical number of examples per weight $\alpha_c(T)$ increases with T in this model, the time it takes to converge to the optimal state increases dramatically as T is lower.

B. The High Temperature and Annealed Approximations

We have presented two approximate theoretical approaches to the problem of learning from examples in layered networks. The first approximation replaces the training energy by the number of examples times the generalization error, and becomes exact in the limit of learning with high thermal noise level. The dependence

$$g_0(x) = x + \theta \ . \tag{4.52}$$

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

$$g(x) = x \ . \tag{4.53}$$

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

The average of the error function

$$\epsilon(\mathbf{W};\mathbf{S}) = \frac{1}{2N}[(\mathbf{W} - \mathbf{W}_0)\cdot\mathbf{S} - \sqrt{N}\theta]^2 \tag{4.54}$$

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

$$\epsilon(\mathbf{W}) = 1 - R + \frac{1}{2}\theta^2 \ . \tag{4.55}$$

At $R = 1$, the generalization error takes on its minimum value, $\epsilon_{\rm min} = \theta^2/2$. Recall that according to Eq. (2.21), the average training and generalization errors both approach $\epsilon_{\rm min}$ as $\alpha \to \infty$, at all T.

In Section II.D we noted that the AA may yield the wrong $\alpha \to \infty$ limit, for unrealizable, non-Boolean rules. The present model is a simple example of this phenomenon. In the present case, $G_{\rm an}$, Eq. (4.14), is

$$G_{\rm an}(R) = \frac{1}{2}\ln[1 + 2\beta(1-R)] + \frac{1}{2}\frac{\beta\theta^2}{1 + 2\beta(1-R)} \ . \tag{4.56}$$

The resulting annealed free energy, Eq. (4.13), is

$$\begin{aligned} -\beta f(R) = &\frac{1}{2}\ln(1-R^2) - \frac{\alpha}{2}\ln\left[1 + 2\beta(1-R)\right] \\ &-\frac{\alpha}{2}\frac{\beta\theta^2}{1 + 2\beta(1-R)} \ . \end{aligned} \tag{4.57}$$

In the $\alpha \to \infty$ limit R is determined by minimizing $G_{\rm an}$, yielding

$$R = \begin{cases} 1, & \theta^2 \le T \ , \\ 1 - \frac{1}{2}(\theta^2 - T), & T < \theta^2 < T + 4 \ , \\ -1, & T + 4 \le \theta^2 \ . \end{cases} \tag{4.58}$$

According to these results, when $\theta^2 > 4$ and $T = 0$, the weight vector $\mathbf{W}$ approaches $-\mathbf{W}^0$ as $\alpha \to \infty$!

To understand the origin of this gross failure of the annealed approximation we recall from Section II.D that the annealed approximation can be viewed as the exact theory for a stochastic training dynamics in both weight and input space, leading to the Gibbs distribution (3.16) in both $\mathbf{W}$ and $\mathbf{S}$ space. The magnitude of the resulting distortion in the aposteriori distribution of the inputs relative to their a priori one

2. Annealed Approximation

Summing Eqs. (4.20) and (4.31) we obtain for the annealed free energy

$$-\beta f = \alpha \ln\left(1 - \frac{1-e^{-\beta}}{\pi}\cos^{-1} R\right) - \frac{1-R}{2}\ln\frac{1-R}{2} - \frac{1+R}{2}\ln\frac{1+R}{2} \ , \tag{4.50}$$

The behavior of this free energy is qualitatively similar to that of the high-T limit. The state $R = 1$ is always a local minimum of f with $f = 0$. In addition to it there is another local minimum at high temperatures and small α, given by the saddle point equation

$$\tanh^{-1} R = \frac{\alpha(1-e^{-\beta})}{\pi\sqrt{1-R^2}}\left(1 - \frac{1-e^{-\beta}}{\pi}\cos^{-1} R\right)^{-1} \tag{4.51}$$

Fig. 6 exhibits the annealed phase diagram, with a thermodynamic transition line to $R = 1$, and also a spinodal line above which the metastable phase vanishes. At high temperatures, the transition lines become linear with slopes given by the high-T theory, Eqs. (4.42)–(4.43) above. At $T = 0$ the AA predicts that for $0 < \alpha < 1.45$ there are states with small values of R that have zero training energy, i.e. there are degenerate ground states. For $\alpha > 1.45$ these poor learning states are no longer compatible with the examples, and hence their training error is positive. They are, however, still metastable and starting from random initial weights ($R \approx 0$) the system will converge to one of these states. At $\alpha > 1.73$ the poor learning states lose their stability and the only metastable state is $R = 1$. The annealed result $\alpha_c = 1.448$ was already derived by Gardner and Derrida[25].

3. Numerical Simulations

Figure 7 compares annealed learning curves with simulations for $N = 75$, averaged over 32 samples. At $T = 5$, there is a good fit to the AA. The only difference is that in the simulations, the transition is smeared over the region between the spinodal point and the thermodynamic transition. Due to the finite size of the system, some networks are able to escape from the metastable phase even below the spinodal point, and consequently the simulations end up sampling a mixture of both the metastable and thermodynamic phases.

G. 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 example of Section IV.A. However, in the present case the transfer function of the teacher has a threshold, i.e., the rule is given by Eq. (4.5) with a transfer function

$$T > 0.59\alpha \ . \tag{4.42}$$

In this regime the optimal state $R = 1$ is only metastable. If the initial network has R which is close to 1 the learning dynamics will converge fast to the state $R = 1$. However starting from a random initial weight vector $R \approx 0$ the system will not converge to the optimal state.

For $T/\alpha < 0.59$ the equilibrium state is $R = 1$ although there is still a local minimum, i.e., a solution of Eq. (4.41) with $R < 1$. Finally for

$$T < 0.48\alpha \ , \tag{4.43}$$

there is no solution with $R < 1$ to Eq. (4.41). In this regime (beyond the spinodal), starting from any initial condition the system converges fast to the optimal state.

The collapse of the system to the energy ground state at finite temperature is an unusual phenomenon. The origin of this behavior is the square root singularity of $\epsilon(R)$ at $R = 1$. This singularity implies that a state characterized by $\delta R \equiv 1 - R \ll 1$ has an energy which is proportional to

$$E \propto N\sqrt{\delta R} \ . \tag{4.44}$$

This big increase in energy cannot be offset by the gain in entropy, which is proportional to

$$\delta N G_0(R) \propto N(\delta R) \ln(\delta R) \ . \tag{4.45}$$

This effect can be nicely seen using the microcanonical description. According to Eq. (2.28) above, a smooth low temperature limit exists provided that

$$\lim_{\epsilon \to \epsilon_{\min}} \partial s(\epsilon)/\partial \epsilon = \infty \ . \tag{4.46}$$

On the other hand, Eq. (4.20) implies that in the present case

$$\partial s(\epsilon)/\partial \epsilon \approx -\epsilon \ln \epsilon \ , \epsilon \to 0 \ . \tag{4.47}$$

Thus the rate of increase in entropy is too small to give rise to thermal fluctuations below some critical temperature.

It is instructive to apply the above argument to the case of states that differ from the ground state by a flip of a single weight. According to Eq. (4.44) the energy of such states is

$$E \propto \sqrt{N} \ , \tag{4.48}$$

whereas the entropy associated with such an excitation is only

$$\delta N G_0(R) \propto \ln N \ . \tag{4.49}$$

It should be emphasized, however, that examining the spectrum of the first excitations is not generally sufficient for determining the thermodynamic behavior at any finite T, where the relevant states are those with energy of order NT.

We have used the Metropolis Monte Carlo algorithm to simulate learning in the linear/discrete perceptron[40]. This algorithm is a standard technique for calculating thermal averages over Gibbs distributions. The simulations were performed for multiple samples, i.e. different training sets drawn randomly from a common input distribution. The quenched average was performed over these samples, and error bars were calculated from the standard error of measurement of the sample distribution. In general, fewer samples were required for larger N, because of self-averaging. Our simulations used discrete inputs $S_i = \pm 1$, rather than continuous, to take advantage of the speedup offered by integer arithmetic. As discussed in Appendix C, both types of inputs should give the same learning curves in the thermodynamic limit. In Fig. 4 simulations with $N = 100$ averaged over 64 samples are compared with the AA learning curves at $T = 1$ and $T = 0.2$. At the higher T the prominent feature is the smooth decrease of ϵ_g to zero with an exponential tail, and the annealed curves fit the simulations quite well. At the lower T, although the fit is not as good, the AA nevertheless captures the prominent feature of the generalization curve, the sudden transition to $R \approx 1$. The transition occurs at $\alpha \approx 0.65$, which is quite close to the the annealed spinodal point $\alpha_o(T = 0.2) = 0.70$. The roundness of the transition in the simulations is consistent with the expected smearing of the discontinuity in a finite system.

F. Boolean Output with Discrete Weights

The boolean/discrete model, first studied by Gardner and Derrida[25], exhibits a striking feature. It has a first order transition from a state of poor learning to a state of perfect learning[38, 27]. Unlike the linear/discrete perceptron discussed above, this model's transition persists at all temperatures. The occurrence of this remarkable transition can be understood using the high temperature limit.

1. The High-Temperature Limit

In the high-T limit the energy of the system is given simply as $N\alpha\epsilon(R)$. Hence, using Eq. (4.12) for $\epsilon(R)$, the free energy is simply

$$-\beta f = -\frac{\alpha\beta}{\pi}\cos^{-1} R - \frac{1-R}{2}\ln\frac{1-R}{2} - \frac{1+R}{2}\ln\frac{1+R}{2} \ , \tag{4.40}$$

This free energy is shown in Fig. 5 for various values of α/T. In contrast to previous models, the state $R = 1$ is a local minimum of f for all values of T and α. For small values of α/T the state $R = 1$ is only a local minimum of f, as can be seen in Fig. 5. The global minimum is given by the solution of the saddle-point equation

$$R = \tanh\left(\frac{\alpha\beta}{\pi\sqrt{1-R^2}}\right) \ . \tag{4.41}$$

This state of poor learning $R < 1$ is the equilibrium state for

that is the global minimum of f) changes discontinuously. In particular, the value of R drops discontinuously upon crossing the line from above. The solid line starts from the origin $\alpha = T = 0$. This implies that for all α there is a unique equilibrium state at low T. As $T \to 0$ this state approaches 1 for all α, again in agreement with Eq. (4.34). Note however that for small α the line approaches the origin as

$$T_c(\alpha) \approx e^{-1/2\alpha} \quad , \quad \alpha \to 0 \ . \tag{4.39}$$

This implies that for a small number of examples even a small noise in the dynamics will generate a transition to the low R state. The first order transition line terminates at the point $T = 0.40, \alpha = 0.87$. Above these values the behavior of the system is smooth with a unique equilibrium phase.

2. Metastability

The above phase diagram indicates that there is a region at low α and T in which there are two states that are local minima of f. The dashed lines marking the boundary of metastability are called *spinodal lines*. The origin of this metastability is evident in Fig. 3 where graphs of the free energy for $T = 0.3$ (see the horizontal dashed line in Fig. 2) and various values of α are shown. For $\alpha < 0.730$ there is only a single local minimum. When we increase α to 0.730 another local minimum appears at higher R. This is called a *spinodal point* of the full spinodal line *(a)* in Fig. 2. However, the original minimum is still the global minimum, so it is still the *thermodynamic phase*. The new minimum is *metastable*. Only at $\alpha = 0.756$, does the new local minimum becomes the global minimum. Since a system in true thermodynamic equilibrium finds the global minimum of its free energy, we identify $\alpha = 0.756$ as the location of a thermodynamic transition. The transition is first order, since R jumps discontinuously from 0.583 to 0.972. However, the local minimum of lower R still exists for $0.756 < \alpha < 0.781$. Only above $\alpha = 0.781$, a second spinodal point, does the low R minimum vanish, leaving only the high R one.

This second spinodal point is of particular importance. Starting from initially random weights ($R \approx 0$), the system should converge fast to the nearest metastable state, which is the state with low R as long as such a state exists. The time required to cross the free energy barrier to the thermodynamic high-R phase scales as $t \approx e^{-\beta N \Delta f}$ where Δf is the barrier of the free energy per weight between the two states, see Fig. 3. Thus the time to reach the true equilibrium state will grow exponentially with the system size. Hence for large networks, the transition to the high-R state will not occur at the equilibrium (solid) line but rather at the spinodal line *(b)*. It is important to note that unlike the equilibrium transition line, the spinodal line terminates at $T = 0$ at a finite value of α, $\alpha = 0.556$. This implies that in spite of Eq. (4.34), a finite value of α is required to learn in a finite time.

3. Numerical simulations

E. Linear Perceptron with Discrete Weights

Imposing binary constraints on the weights of a perceptron with a linear output modifies its learning performance drastically. Let us first consider the zero temperature behavior. The weight vector must satisfy the same P homogeneous linear equations as before, Eqs. (4.21), but now it is constrained to $W_i = \pm 1$. For almost every continuous input $\mathbf{S}$, the equation $(\mathbf{W} - \mathbf{W}^0) \cdot \mathbf{S} = 0$ cannot be satisfied unless $\mathbf{W} = \mathbf{W}^0$. Hence just one example guarantees perfect learning at zero temperature, i. e.

$$\alpha_c = 0 \ . \tag{4.34}$$

However, this argument does not exclude the existence of local minima of E. In this case the main effect of increasing the number of examples is to smooth the energy surface, in order to make the unique global minimum dynamically accessible for large networks. In order to investigate these aspects, study of the finite T version of the problem is extremely useful.

1. Equilibrium

Using Eqs. (4.14)–(4.15), and Eq. (4.20), the annealed free energy is

$$-\beta f = -\frac{1}{2}\alpha \ln[1 + 2\beta(1-R)] - \frac{1-R}{2}\ln\frac{1-R}{2} - \frac{1+R}{2}\ln\frac{1+R}{2} \ , \tag{4.35}$$

which is extremized when

$$\tanh^{-1} R = \frac{\alpha\beta}{1 + 2\beta(1-R)} \ . \tag{4.36}$$

For large α, the learning curve

$$\epsilon_g \equiv 1 - R \approx 2e^{-2\alpha/T} \tag{4.37}$$

has an exponential tail. Thus the learning curve decreases much faster than the $1/\alpha$ law that holds for continuous weights, Eq. (4.28). Also, the prefactor of α in the exponent diverges as $T \to 0$, signaling the fact that

$$\epsilon_g = 0 \quad , \quad T = 0 \ , \tag{4.38}$$

for all $\alpha \geq 0$, in agreement with Eq. (4.34).

Besides the asymptotic behavior this model differs from the linear/continuous one by possessing a rich phase diagram in the T-α plane, associated with a discontinuous phase transition at low T. The phase diagram predicted by the AA is shown in Fig. 2. The dashed lines marked *(a)* and *(b)* bound the region where there are two locally stable solutions of Eq. (4.36) for R. The solid line marks where the two solutions have equal free energy. Upon crossing this line the equilibrium state (i.e., the state

In the high-T limit we have

$$\epsilon_t = \epsilon_g = 1 - R = 1 + \frac{T}{2\alpha} - \sqrt{1 + \left(\frac{T}{2\alpha}\right)^2} \tag{4.27}$$

Hence for all T, the asymptotic behavior of R is

$$R = 1 - \frac{T}{2\alpha} \ , \quad \alpha \to \infty \ , \tag{4.28}$$

The asymptotic behavior of the training error ϵ_t is the same as that of ϵ_g.

At low T and finite α, $\epsilon_g \geq \epsilon_t$. In particular at $T = 0$, the AA predicts

$$\epsilon_g = \frac{2 - 2\alpha}{2 - \alpha} \ , \ \alpha \leq 1 \ , \tag{4.29}$$

with a second order transition to perfect learning at $\alpha_c = 1$. On the other hand the training error is zero, at $T = 0$, for all α. These $T = 0$ learning curves are shown in Fig. 1.

D. Boolean Output with Continuous Weights

The boolean perceptron with continuous weights corresponds to the original perceptron studied in[42, 44]. At zero temperature, weight vectors with zero training energy satisfy the inequalities

$$(\mathbf{W} \cdot \mathbf{S}^l)(\mathbf{W}^0 \cdot \mathbf{S}^l) > 0 \ . \tag{4.30}$$

Since these inequalities do not constrain the weight space as much as the equalities (4.21), this model requires more examples for good generalization then did the preceding linear model. The quenched theory of this model has been previously studied in detail[26]. Here we present the simpler annealed approximation.

In this model $g(x) = \mathrm{sign}(x)$ and Eq. (4.14) reads

$$G_{\mathrm{an}}(R) = -\ln\left(1 - \frac{1 - e^{-\beta}}{\pi}\cos^{-1} R\right) \tag{4.31}$$

yielding for the free energy

$$-\beta f = \frac{1}{2}\ln(1 - R^2) + \alpha \ln\left(1 - \frac{1 - e^{-\beta}}{\pi}\cos^{-1} R\right) \tag{4.32}$$

Evaluating R by minimizing the free energy, we find that there is no transition to $R = 1$ at any T and α. The state $R = 1$ is approached only asymptotically as $\alpha \to \infty$. The corresponding generalization error vanishes at large α as

$$\epsilon_g = \frac{1}{1 - e^{-\beta}}\frac{1}{\alpha} + \mathcal{O}(\alpha^{-2}) \tag{4.33}$$

Unlike the previous linear/continuous model, the $1/\alpha$ power law persists even at $T = 0$: there is no transition to perfect learning.

In the quenched theory discussed in Part II, there appear order parameters that represent explicitly the role of fluctuations induced by the sample to sample variation of the examples. It is useful to note that although the AA lacks these order parameters, it nevertheless does incorporate implicitly the effect of the sample fluctuations. This can be seen from the fact that the average training error predicted by the theory will be in general less than ϵ_g. This is a simple manifestation of the overtraining of the training set. Finally, as discussed above for general problems, we expect that the AA may provide a reasonable approximation at least for (1) the case of a realizable rule, i.e., that $\mathbf{W}^0$ is contained within the allowed configuration space of networks; and (2) the boolean case, i.e., $g(x) = \mathrm{sign}(x)$.

C. Linear Perceptron with Continuous Weights

The case of a perceptron with a quadratic error function Eq. (4.9) defined on a continuous weight space is particularly simple, and has been studied by Krogh and Hertz[43]. The zero temperature behavior can be understood from simple considerations. Weight vectors with zero training energy satisfy

$$(\mathbf{W} - \mathbf{W}^0) \cdot \mathbf{S}^l = 0 \ , \quad l = 1, \ldots, P \ . \tag{4.21}$$

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. (4.21), $\mathbf{W} = \mathbf{W}^0$, for almost every random choice of examples. Hence there is a zero temperature transition to *perfect learning* ($\epsilon_g = 0$) at

$$\alpha_c = 1 \ . \tag{4.22}$$

Evaluating Eq. (4.14), with $g(x) = x$ yields

$$G_{\mathrm{an}} = \frac{1}{2} \ln \left[1 + 2\beta(1 - R) \right] \ . \tag{4.23}$$

Adding this to (4.18) yields the annealed free energy of Eq. (4.13),

$$-\beta f(R) = \frac{1}{2} \ln(1 - R^2) - \frac{\alpha}{2} \ln \left[1 + 2\beta(1 - R) \right] \ , \tag{4.24}$$

which is extremized when

$$\frac{R}{1 - R^2} = \frac{\alpha\beta}{1 + 2\beta(1 - R)} \ . \tag{4.25}$$

The generalization error is related to R via Eq. (4.10). The training error is given by

$$\epsilon_t = \frac{1}{\alpha} \frac{\partial \beta f}{\partial \beta} = \frac{1 - R}{1 + 2\beta(1 - R)} \ . \tag{4.26}$$

Solving for R one then evaluates the average generalization error via (4.7). Likewise the average training error is evaluated by differentiating $G_{\rm an}$ with respect to β, as in Eq. (3.6).

We will consider the case of a *spherical constraint* and that of an *Ising constraint.* In the perceptron with a spherical constraint, the a priori measure $d\mu(\mathbf{W})$ is uniform on the sphere of radius $\sqrt{N}$, given by the normalization condition, Eq. (4.2). We may write the measure more formally as

$$d\mu(\mathbf{W}) \equiv \prod_{i=1}^{N} \frac{dW_i}{\sqrt{2\pi e}} \delta(\mathbf{W}\cdot\mathbf{W} - N) \ , \tag{4.17}$$

which is normalized to $\int d\mu(\mathbf{W}) = 1$. In this case, the fraction of weight space with an overlap R is simply the volume of the $N-2$ dimensional sphere with radius $\sqrt{1-R^2}$. Hence the entropy $G_0(R)$, Eq. (4.15), is (in the limit of large N)

$$G_0(R) = \frac{1}{2}\ln(1-R^2) \ , \tag{4.18}$$

a result that is derived in more detail in Appendix C. The entropy diverges as $R \to 1$, as the fraction of weight space with overlap R approaches zero. Such a divergence is typical of a continuous weight space.

The Ising perceptron corresponds to a network with binary valued weights $W_i = \pm 1$, or

$$d\mu(\mathbf{W}) \equiv \prod_{i=1}^{N} dW_i[\delta(W_i - 1) + \delta(W_i + 1)] \ . \tag{4.19}$$

The entropy of Ising networks with an overlap R is given by

$$G_0(R) = -\frac{1-R}{2}\ln\frac{1-R}{2} - \frac{1+R}{2}\ln\frac{1+R}{2} \ , \tag{4.20}$$

a result derived in Appendix C. It approaches zero as $R \to 1$, meaning that there is exactly one state with $R = 1$. This nondivergent behavior is typical of discrete weight spaces.

To conclude, the picture emerging from the AA is extremely simple. The properties of the system can be expressed in terms of a single order parameter, namely, the overlap R. The stochastic fluctuations in the value of R can be neglected in the limit of large N. Hence the system almost always converges to a unique value of R given by the minimum of the free energy $f(R)$. Depending on the details of the problem, $f(R)$ can have more than one local minimum. If this happens, the equilibrium properties of the system are determined by the unique global minimum of f. The local minima represent metastable states. Starting from a value of R near one of these minima the system is likely to converge rapidly to the local minimum. It will remain in this state for a time that scales exponentially with the network size. Hence for large networks the local minima of f can be considered as stable states.

the teacher weight vectors vanishes. Perfect learning corresponds to an overlap $R = 1$ or $\theta = 0$.

In the following we discuss perceptrons with either linear or Boolean outputs, and weights that are either binary or obey a spherical constraint.

For the *linear perceptron*, the transfer function is $g(x) = x$. The error function, Eq. (4.6), is in this case a quadratic function in weight space,

$$\epsilon(\mathbf{W};\mathbf{S}) = \frac{1}{2N}[(\mathbf{W} - \mathbf{W}^0)\cdot\mathbf{S}]^2 \ . \tag{4.9}$$

Averaging this function over the whole input space, we find

$$\epsilon(\mathbf{W}) = 1 - R, \tag{4.10}$$

in accord with Eq. (4.7) with a linear g.

A second output function to be considered is the *Boolean* output $g(x) = \mathrm{sign}(x)$, which corresponds to the original perceptron model studied by Rosenblatt[42]. The *Boolean perceptron* $\sigma = \mathrm{sign}(\mathbf{W}\cdot\mathbf{S})$ separates the input space in half via a hyperplane perpendicular to the weight vector. The error function, Eq. (4.6), is (up to a factor of 2)

$$\epsilon(\mathbf{W};\mathbf{S}) = \Theta[-(\mathbf{W}\cdot\mathbf{S})(\mathbf{W}^0\cdot\mathbf{S})] \ , \tag{4.11}$$

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

$$\epsilon(\mathbf{W}) = \frac{1}{\pi}\cos^{-1} R \tag{4.12}$$

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

B. The Annealed Approximation for Perceptron Learning

The annealed free energy of perceptron learning is shown in Appendix C to be

$$-\beta f = G_0(R) - \alpha G_{\mathrm{an}}(R) \ , \tag{4.13}$$

$$G_{\mathrm{an}}(R) = -\ln \int Dx \int Dy \ \exp\left(-\frac{\beta}{2}\left[g\left(x\sqrt{1-R^2} + yR\right) - g(y)\right]^2\right) \ , \tag{4.14}$$

$$G_0(R) = N^{-1}\ln \int d\mu(\mathbf{W})\,\delta(R - N^{-1}\mathbf{W}\cdot\mathbf{W}^0) \ . \tag{4.15}$$

The function $NG_0(R)$ is the logarithm of the density of networks with overlap R, so we will sometimes refer to it as the "entropy," even though it is not the same as the thermodynamic entropy $s = -\partial f/\partial T$. The properties of the system in the large N limit are obtained by minimizing the free energy f, which yields

$$\frac{\partial G_0(R)}{\partial R} = \alpha\frac{\partial G_{\mathrm{an}}(R)}{\partial R} \ . \tag{4.16}$$

is a search through weight space for the perceptron that best approximates a target rule. 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,

$$d\mu(\mathbf{S}) = \prod_{i=1}^{N} DS_i \ , \tag{4.3}$$

where Dx denotes the normalized Gaussian measure

$$Dx \equiv \frac{dx}{\sqrt{2\pi}} e^{-x^2/2} \ . \tag{4.4}$$

In most of this section we consider only the case of a perfectly realizable target rule. This means that it is another perceptron of the form

$$\sigma^0(\mathbf{S}) = g\left(\frac{1}{\sqrt{N}}\mathbf{W}^0 \cdot \mathbf{S}\right) \ , \tag{4.5}$$

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

Training is performed by a stochastic dynamics of the form (2.7) with the training energy function (2.1). For each example the error function is taken to be

$$\epsilon(\mathbf{W};\mathbf{S}) = \frac{1}{2}\left[g\left(N^{-\frac{1}{2}}\mathbf{W}\cdot\mathbf{S}\right) - g\left(N^{-\frac{1}{2}}\mathbf{W}^0\cdot\mathbf{S}\right)\right]^2 \ . \tag{4.6}$$

The generalization function is

$$\epsilon(\mathbf{W}) = \int D\mathbf{S}\, \epsilon(\mathbf{W};\mathbf{S}) = \int Dx \int Dy\, \frac{1}{2}[g(x\sqrt{1-R^2} + yR) - g(y)]^2 \tag{4.7}$$

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

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

(see Appendix C). The relationship between (4.6) and (4.7) is plain, since in both cases the arguments of g are Gaussian random variables with unit variance and cross-correlation R.

It is important to note that in perceptron learning, the generalization function of a network 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 function goes to zero as the angle between the student and

where $\mathbf{W}^\dagger$ minimizes $G_{\rm an}$. In general, this vector is not necessarily the same as the vector $\mathbf{W}^*$ which minimizes $\epsilon(\mathbf{W})$. Hence, there is no guarantee that the AA correctly predicts the value of the optimal generalization error or the values of the optimal weights, except for two special cases. One is the case of a realizable rule, for which $\epsilon(\mathbf{W}^*;\mathbf{S}) = 0$ for all inputs $\mathbf{S}$. Clearly the minimum of $G_{\rm an}$ in Eq. (3.4) then occurs at $G_{\rm an}(\mathbf{W}^*) = 0$. The second is the case of a network whose output layer consists of a single boolean output unit, as discussed above. From (3.7) it is evident that the minimum of $G_{\rm an}$, in this case, coincides with the minimum of ϵ_g, and hence $\mathbf{W}^\dagger = \mathbf{W}^*$. To summarize, the correct limit $\epsilon_g \to \epsilon_{\rm min}$, stated in Eq. (2.21), can be violated by the AA for unrealizable, non-Boolean rules.

With respect to the training error, the AA for unrealizable rules is also inadequate: the correct limit $\epsilon_t \to \epsilon_{\rm min}$ is typically violated, even for the Boolean case, and the limit $\epsilon_t \to \epsilon_g$ does not hold either. For Boolean rules, this can be seen from the expression (3.12) for the training error. In the $T \to 0$ limit, the annealed training error approaches

$$\lim_{T\to 0} \epsilon_t(T,\alpha) = \min_{\mathbf{W},\mathbf{S}} \epsilon(\mathbf{W};\mathbf{S}) \tag{3.19}$$

since annealing both the weights and examples at zero temperature minimizes the training energy with respect to all variables. Often the right hand side is zero, so that the AA predicts $\epsilon_t(T=0,\alpha) = 0$ for all α. This is a gross violation of the correct limit $\epsilon_t \to \epsilon_{\rm min}$. Hence the fact that the inequality $\epsilon_t < \epsilon_g$ holds in the AA may be a spurious effect of the distortion of the input distribution, rather than an honest accounting for the randomness of the examples.

In short, we expect the AA to be adequate at least qualitatively for realizable rules but inadequate, except at high temperatures, for unrealizable ones.

IV. LEARNING OF A PERCEPTRON RULE

In this Section we apply the AA to study several models of a single-layer perceptron learning a realizable perceptron rule. We then demonstrate the breakdown of the AA in a simple model of an unrealizable perceptron rule. Discussion of the full quenched theory of these models will be deferred to Part II.

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 of order unity. Learning

B. The Annealed Approximation as a Dynamics in Example Space

The above annealed results only approximate the learning procedure described in Eq. (2.7). However, they can be viewed as the exact theory for a dynamic process where both the weights and the examples are updated by a stochastic dynamics, similar to Eq. (2.7), involving the same energy function, i.e.,

$$\frac{\partial \mathbf{W}}{\partial t} = -\nabla_{\mathbf{W}} E + \boldsymbol{\eta}(t) \ , \tag{3.14}$$

$$\frac{\partial \mathbf{S}^l}{\partial t} = -\nabla_{\mathbf{S}^l} E + \boldsymbol{\eta}_l(t), \tag{3.15}$$

Here E is a function of both $\mathbf{W}$ and $\mathbf{S}^l$. This dynamic process leads to a Gibbs probability distribution both in weight space and in input space

$$\mathcal{P}_{\rm an}(\mathbf{W};\mathbf{S}^l) = Z_{\rm an}^{-1} e^{-\beta E(\mathbf{W};\mathbf{S}^l)} \ , \tag{3.16}$$

where

$$Z_{\rm an} = \int d\mu(\mathbf{W}) \int \prod_{l=1}^{P} d\mu(\mathbf{S}^l) \exp(-\beta E(\mathbf{W})) \ , \tag{3.17}$$

which is exactly the annealed partition function.

From the perspective of Eqs. (3.14)–(3.15) the AA represents the behavior of a system with a distorted measure of the input space. The fact that we will find it to be a good approximation in several nontrivial cases reflects the robustness of the performance of the networks in these cases to moderate distortions of the input measure. The effect of reducing temperature is also clear. The bigger β is, the larger the distortion of the a priori input measure due to the Boltzmann factor. Consequently, one expects that deviations from the AA may be important at low T.

C. How good is the Annealed Approximation?

First we note that $G_{\rm an} \to \beta\epsilon(\mathbf{W})$ as $\beta \to 0$. Thus the AA is valid at high temperatures, since it reduces to the high-T limit described above. At lower temperatures the AA deviates significantly from the high-T theory. In particular, it does seem to incorporate the important effect of overtraining, in that ϵ_g, Eq. (3.5), is in general larger than ϵ_t, Eq. (3.6). On the other hand, except for the high-T limit, the results of the AA are in general not exact.

To obtain some insight into the quality of the AA at finite temperatures we examine its behavior in the limit of large α. From Eq. (3.7) it follows that in the AA the asymptotic value of the generalization error is

$$\lim_{\alpha\to\infty} \epsilon_g(T,\alpha) = \epsilon(\mathbf{W}^\dagger) \ , \tag{3.18}$$

A. Single Boolean Output

A particularly simple case is that of an output layer consisting of a single boolean output unit. In this case $\epsilon(\mathbf{W};\mathbf{S}) = 1$ or 0 only, so that

$$G_{\rm an}(\mathbf{W}) = -\ln[1-(1-e^{-\beta})\epsilon(\mathbf{W})] \ . \tag{3.7}$$

Since $G_{\rm an}$ depends on $\mathbf{W}$ only through $\epsilon(\mathbf{W})$, which is of order 1, we can write a microcanonical form of the AA, analogous to what was done for the high-T limit in Eqs. (2.25)–(2.28). The annealed partition function $\langle\!\langle\, Z \,\rangle\!\rangle$ takes the form

$$\langle\!\langle\, Z \,\rangle\!\rangle = \int d\epsilon \, \exp N[G_0(\epsilon) - \alpha G_{\rm an}(\epsilon)] \ , \tag{3.8}$$

where

$$G_{\rm an}(\epsilon) \equiv -\ln[1-(1-e^{-\beta})\epsilon] \tag{3.9}$$

$$G_0(\epsilon) \equiv N^{-1} \ln \int d\mu(\mathbf{W}) \delta(\epsilon - \epsilon(\mathbf{W})) \ . \tag{3.10}$$

The function $NG_0(\epsilon)$ is the logarithm of the density of networks with generalization error ϵ. At finite temperature, it is different from the annealed entropy $S_{\rm an} \equiv -\partial F_{\rm an}/\partial T$, which is the logarithm of the density of networks with training error ϵ. However, since $\epsilon_t = \epsilon_g$ in the high temperature limit, G_0 approaches $s_{\rm an}$.

In the thermodynamic limit ($N \to \infty$), the integral (3.8) is dominated by its saddle point. Thus at any given α and T the value of the average generalization error is given by minimizing the free energy $f(\epsilon)$, where $-\beta f \equiv G_0 - \alpha G_{\rm an}$. This leads to the implicit equation

$$\left.\frac{\partial G_0}{\partial \epsilon}\right|_{\epsilon=\epsilon_g} = -\frac{\alpha(1-e^{-\beta})}{1-(1-e^{-\beta})\epsilon_g} \ , \tag{3.11}$$

which is analogous to the high-T result Eq. (2.28). It is interesting to note that in this case the AA predicts a simple relation between the training and generalization errors. Using Eq. (3.6) above, one obtains

$$\epsilon_t = \frac{e^{-\beta}\epsilon_g}{1-(1-e^{-\beta})\epsilon_g} \tag{3.12}$$

where ϵ_g is the average generalization error given by (3.11), or equivalently

$$\epsilon_g = \frac{\epsilon_t}{e^{-\beta} + (1-e^{-\beta})\epsilon_t} \tag{3.13}$$

To the extent that the annealed approximation is valid, this relation could be used in actual applications to estimate the generalization error from the measured training error.

Thus the properties of the system in the high-T limit are determined by the dependence of the entropy on generalization error.

From the theoretical point of view, the high-T limit simply characterizes models in terms of an effective energy function $\epsilon(\mathbf{W})$ which is often a rather smooth function of $\mathbf{W}$. The smoothness of the effective energy function also implies that the learning process at high temperature is relatively fast. One does not expect to encounter many local minima, although a few large local minima may still remain, as will be seen in some of the models below. Another feature of learning at high temperature is the lack of *overtraining*. One measure of overtraining of examples is the difference between the expected training and generalization errors, i.e., $\epsilon_g - \epsilon_t$. From Eq. (2.23) and the definitions Eqs. (2.11)–(2.12) it follows that $\epsilon_t = \epsilon_g$ in the high-T limit. Of course the price that one pays for learning at high temperature is the necessity of a large training set, as α must be at least of order T.

III. THE ANNEALED APPROXIMATION

Another useful approximate method for investigating learning in neural networks is the annealed approximation, or AA for short. It consists of replacing the average of the logarithm of Z, Eq. (2.13), by the logarithm of the average of Z itself. Thus the annealed approximation for the average free energy $F_{\rm an}$ is

$$-\beta F_{\rm an} = \ln \langle\!\langle\, Z \,\rangle\!\rangle \ . \tag{3.1}$$

Using the convexity of the logarithm function, it can be shown that the annealed free energy is a lower bound for the true quenched value,

$$F_{\rm an} \leq F \ . \tag{3.2}$$

Whether this lower bound can actually serve as a good approximation will be examined critically in this work.

Using Eqs. (2.10) and (2.1) one obtains

$$\langle\!\langle\, Z \,\rangle\!\rangle = \int d\mu(\mathbf{W}) e^{-PG_{\rm an}(\mathbf{W})} \ , \tag{3.3}$$

$$G_{\rm an}(\mathbf{W}) = -\ln \int d\mu(\mathbf{S})\, e^{-\beta\epsilon(\mathbf{W};\mathbf{S})} \ . \tag{3.4}$$

The generalization and training errors are approximated by

$$\epsilon_g = \frac{1}{\langle\!\langle\, Z \,\rangle\!\rangle} \int d\mu(\mathbf{W}) \epsilon(\mathbf{W}) e^{-PG_{\rm an}(\mathbf{W})} \ , \tag{3.5}$$

$$\epsilon_t = \frac{1}{\langle\!\langle\, Z \,\rangle\!\rangle} \int d\mu(\mathbf{W}) \frac{\partial G_{\rm an}(\mathbf{W})}{\partial \beta} e^{-PG_{\rm an}(\mathbf{W})} \ . \tag{3.6}$$

In this limit E can simply be replaced by its average Eq. (2.18), and the fluctuations δE, coming from the finite sample of randomly chosen examples, can be ignored. To see this we note that δE is of order $\sqrt{P}$. The leading contribution to βF from the term $\beta \delta E$ in Z is proportional to $\beta^2 \langle\langle (\delta E)^2 \rangle\rangle \approx N\alpha\beta^2$. This is down by a factor of β compared to the contribution of the average term, which is of the order $N\alpha\beta$. Thus in this limit, the equilibrium distribution of weights is given simply by

$$\mathcal{P}_0(\mathbf{W}) = Z^{-1} \exp(-N\beta\alpha\epsilon(\mathbf{W})) \ . \tag{2.23}$$

where

$$Z_0 = \int d\mu(\mathbf{W}) \exp(-N\beta\alpha\epsilon(\mathbf{W})) \ . \tag{2.24}$$

The subscript 0 signifies that the high temperature limit is the zeroth order term of a complete high temperature expansion, derived in Appendix B.

In the high-T limit, it is clear from Eq. (2.23) that all thermodynamic quantities, including the average training and generalization errors, are functions only of the *effective temperature* T/α. It should be emphasized that the present case is unlike most $T \to \infty$ limits in statistical mechanics, in which all states become equally likely, regardless of energy. Here the simultaneous $\alpha \to \infty$ limit guarantees nontrivial behavior, with contributions from both energy and entropy. In particular, as the effective temperature T/α decreases, the network approaches the optimal ("ground state") weights $\mathbf{W}^*$ that minimize $\epsilon(\mathbf{W})$. This behavior is similar to the $T =$ finite, $\alpha \to \infty$ limit mentioned in (2.21).

It is sometimes useful to discuss the microcanonical version of the statistical mechanics of learning in the high-T limit. Equation (2.24) can be written as

$$Z_0 = \int d\epsilon \exp(-N\beta\alpha f(\epsilon)) \ , \tag{2.25}$$

where the free energy per weight of all networks whose generalization error equals ϵ is

$$f(\epsilon) = \epsilon - \frac{T}{\alpha} s(\epsilon) \ . \tag{2.26}$$

The function $s(\epsilon)$ is the entropy per weight of all the networks with $\epsilon(\mathbf{W}) = \epsilon$, i.e.,

$$s(\epsilon) = N^{-1} \ln \int d\mu(\mathbf{W}) \, \delta(\epsilon(\mathbf{W}) - \epsilon) \ . \tag{2.27}$$

In the large N limit the expected generalization error is simply given by

$$\beta\alpha = \partial s / \partial \epsilon \ . \tag{2.28}$$

$$S = -\frac{\partial F}{\partial T} \ . \tag{2.17}$$

Formally our results will be exact in the thermodynamic limit, i. e. when the size of the network approaches infinity. The relevant scale is the total number of degrees of freedom, namely the total number of (independently determined) synaptic weights N. For the limit $N \to \infty$ to be well defined we envisage that the problem at hand as well as the network architecture allow for a uniform scale-up of N. However, our results should provide a good approximation to the behavior of networks with a fixed large size.

The correct thermodynamic limit requires that the energy function be extensive, i. e. proportional to N. The consequences of this requirement can be realized by averaging Eq. (2.1) over the example sets, yielding

$$\langle\!\langle \, E(\mathbf{W}) \, \rangle\!\rangle = P\epsilon(\mathbf{W}). \tag{2.18}$$

Hence, assuming that $\epsilon(\mathbf{W})$ is of order 1, the number of examples should scale as

$$P = \alpha N, \tag{2.19}$$

where the proportionality constant α remains finite as N grows. This scaling guarantees that both the entropy and the energy are proportional to N. The balance between the two is controlled by the noise parameter T, which remains finite in the thermodynamic limit.

Finally, using the definitions Eqs. (2.11) and (2.12) and the convexity of the free energy one can show that

$$\epsilon_t(T, \alpha) < \epsilon_g(T, \alpha) \tag{2.20}$$

for all T and α (see Appendix A). When the inequality holds it reflects the phenomenon of overtraining of the examples mentioned above. However, as the number of examples P increases, the energy scales like P and becomes proportional to the generalization function, as in Eq. (2.18). This implies that, for any fixed temperature, increasing α yields the desired limits

$$\epsilon_g \to \epsilon_{\min}, \quad \epsilon_t \to \epsilon_{\min}, \quad \alpha \to \infty \ . \tag{2.21}$$

This result will be derived in Part II using the replica formalism.

C. High Temperature Limit

A simple and interesting limit of the learning theory is that of high temperatures. This limit is defined so that both T and α approach infinity, but their ratio remains constant:

$$\beta\alpha = \text{finite} \ , \quad \alpha \to \infty \ , \quad T \to \infty \ . \tag{2.22}$$

stochastic dynamics provides a way of generating a useful measure, namely a Gibbs distribution, over the space of the solutions.

In the present work, we study only long-time properties. As is well known, Eqs. (2.7) generate at long times a Gibbs probability distribution. In our case it is

$$\mathcal{P}(\mathbf{W}) = Z^{-1} e^{-\beta E(\mathbf{W})} \ , \tag{2.9}$$

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

$$Z = \int d\mu(\mathbf{W}) \exp(-\beta E(\mathbf{W})), \tag{2.10}$$

We have incorporated the contribution from $V(\mathbf{W})$ into the a priori normalized measure in weight space, $d\mu(\mathbf{W})$. The powerful formalism of equilibrium statistical mechanics may now be applied to calculate thermal averages, i.e., averages with respect to $P(\mathbf{W})$. They will be denoted by $\langle \cdots \rangle_T$. In the thermodynamic limit, such average quantities yield information about the typical performance of a network, governed by the above measure, independent of the initial conditions of the learning dynamics.

Even after the thermal average is done, there is still a dependence on the P examples $\mathbf{S}_l$. Since the examples are chosen randomly and then fixed, they represent *quenched* disorder. Thus to explore the typical behavior we must perform a second, quenched average over the distribution of example sets, denoted by $\langle\!\langle \cdots \rangle\!\rangle \equiv \int \prod_l d\mu(\mathbf{S}^l)$.

The *average training* and *generalization errors* are given by

$$\epsilon_t(T,P) \equiv P^{-1} \langle\!\langle \, \langle E(\mathbf{W}) \rangle_T \, \rangle\!\rangle \ , \tag{2.11}$$

$$\epsilon_g(T,P) \equiv \langle\!\langle \, \langle \epsilon(\mathbf{W}) \rangle_T \, \rangle\!\rangle \ . \tag{2.12}$$

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

$$F(T,P) = -T \langle\!\langle \ln Z \rangle\!\rangle, \tag{2.13}$$

$$S(T,P) = -\int d\mu(\mathbf{W}) \langle\!\langle \, \mathcal{P}(\mathbf{W}) \ln \mathcal{P}(\mathbf{W}) \, \rangle\!\rangle. \tag{2.14}$$

They are related by the identity

$$F = P\epsilon_t - TS \ . \tag{2.15}$$

Knowing F the expected training error can be evaluated via

$$\epsilon_t = \frac{1}{P} \frac{\partial(\beta F)}{\partial \beta} \ , \tag{2.16}$$

and the entropy by

$$\epsilon_{\min} = \min_{\mathbf{W}} \epsilon(\mathbf{W}) > 0 \ . \tag{2.6}$$

Unrealizable rules occur in two basic situations. In the first, the data available for training are corrupted with noise, making it impossible for the network to reproduce the *data* exactly, even with a large training set. This case has been considered by several authors. Particularly relevant works are Refs. [26] and [39], which show that even with noisy data the *underlying target rule* itself can be reproduced exactly in the limit. We will not address this case explicitly. A second situation, which we do consider, is when the network architecture is restricted in a manner that does not allow an exact reproduction of the target rule itself.

B. Learning at Finite Temperature

We consider a stochastic learning dynamics that is a generalization of Eq. (2.3). The weights evolve according to a relaxational Langevin equation

$$\frac{\partial \mathbf{W}}{\partial t} = -\nabla_{\mathbf{W}} E(\mathbf{W}) - \nabla_{\mathbf{W}} V(\mathbf{W}) + \boldsymbol{\eta}(t), \tag{2.7}$$

where $\boldsymbol{\eta}$ is a white noise with variance

$$\langle \eta_i(t) \eta_j(t') \rangle = 2T \delta_{ij} \delta(t - t') \ . \tag{2.8}$$

We have added also a potential $V(\mathbf{W})$ that represents possible constraints on the range of weights. This term depends on the assumptions about the a priori distribution of $\mathbf{W}$ and does not depend on the examples. The above dynamics tends to decrease the energy, but occasionally the energy may increase due to the influence of the thermal noise. At $T = 0$, the noise term drops out, leaving the simple gradient descent equation (2.3). The above equations are appropriate for continuously varying weights. We will also consider weights that are constrained to discrete values. In such cases the analog of (2.7) is a discrete-time Monte Carlo algorithm, similar to that used in simulating Ising systems[40].

In simulated annealing algorithms for optimization problems, thermal noise has been used to prevent trapping in local minima of the energy[41]. The temperature is decreased slowly so that eventually at $T \approx 0$ the system settles to a state with energy near the global energy minimum. Although thermal noise could play the same role in the present training dynamics, it may play a more essential role in achieving good learning. Since the ultimate goal is to achieve good generalization, reaching the global minimum of the training energy may not be necessary. In fact, in some cases training at fixed finite temperature may be advantageous, as it may prevent the system from *overtraining*, namely finding an accurate fit to the training data at the expense of good generalization abilities. Finally, often there are many nearly degenerate minima of the training error, particularly when the available data set is limited in size. In these cases it is of interest to know the properties of the ensemble of solutions. The

Training is then accomplished by minimizing the energy, for example via gradient descent

$$\frac{\partial \mathbf{W}}{\partial t} = -\nabla_{\mathbf{W}} E(\mathbf{W}) \ . \tag{2.3}$$

Gradient descent schemes have been criticized because they tend to become trapped in local minima of the training energy. Putting such difficulties aside, there is another potential problem with the training energy: it measures the network's performance on a limited set of examples. The true goal of supervised learning is to find a network that performs well on all inputs, not just those in the training set. The performance of a given network $\mathbf{W}$ on the whole input space is measured by the *generalization function*. It is defined as the average error of the network over the whole input space, *i.e.,*

$$\epsilon(\mathbf{W}) = \int d\mu(\mathbf{S})\, \epsilon(\mathbf{W};\mathbf{S}) \ . \tag{2.4}$$

Like the gradient descent procedure (2.3), any training algorithm determines some trajectory $\mathbf{W}(t)$ in weight space. The performance of a training algorithm can be monitored by following the generalization and training errors as functions of time. Such graphs can be found in many empirical studies of neural network learning, and are analogous to the *learning curves* used by psychologists to quantify the performance of human and animal subjects in the laboratory[33].

But in this work, we are not so much concerned with the training process as with the final result. We assume that the training has already been completed, e.g. the training error has been reduced as much as possible. What will this minimal value be? And what will the generalization error be? The answers to these questions depend on the number of examples P. If P is small, it is easy to load the examples onto the network with little training error, but good generalization can hardly be expected with such a sparse sampling of the inputs. If P is increased, the training energy becomes a denser sampling of the input space, increasing the chances of good generalization. We define the *generalization* and *training curves* to be the graphs of the generalization and training errors versus P. Both graphs will be referred to as learning curves. We expect that training procedures using the energy (2.1) should be able to achieve optimal generalization in the limit $P \to \infty$, i.e. the generalization curve will asymptotically approach the minimum possible value of $\epsilon(\mathbf{W})$ within the specified weight space.

We distinguish between learning of *realizable rules* and *unrealizable rules*. Realizable rules are those target functions $\sigma_0(\mathbf{S})$ that can be completely realized by at least one of the networks in the weight space. Thus in a realizable rule there exists a weight vector $\mathbf{W}^*$ such that

$$\epsilon(\mathbf{W}^*, \mathbf{S}) = 0, \quad \text{for all } \mathbf{S} \ , \tag{2.5}$$

or equivalently, $\epsilon(\mathbf{W}^*) = 0$. An *unrealizable* rule is a target function for which

above-mentioned annealed approximation. We first derive some general results in the context of these approximations and then proceed to analyze perceptron models of learning realizable rules. For realizable rules, the AA gives a good understanding of generalization performance, without the technical complexity of the replica method. However, as shall be illustrated with a simple model, the AA typically breaks down for unrealizable rules. In Part II we present in detail the replica theory of learning from examples and some of its general consequences. Then we derive exact results for the learning models of realizable rules that were treated approximately in Part I, and finally we investigate unrealizable perceptron rules.

The outline of Part I is the following. In Section II, the general statistical mechanical theory as well as the high-T limit are formulated. In Section III the annealed approximation is described and critically examined. In Section IV, we describe the statistical mechanics of perceptron models of learning and study in detail five concrete examples of perceptron learning. In Section V the main results are summarized and discussed. Some of the material in this work previously appeared in preliminary form in Ref. [38].

II. GENERAL THEORY

A. Learning from Examples

We consider a network with M input nodes S_i $(i = 1, \ldots, M)$, N synaptic weights W_i $(i = 1, \ldots, N)$, and a single output node $\sigma = \sigma(\mathbf{W}; \mathbf{S})$. The quantities $\mathbf{S}$ and $\mathbf{W}$ are M- and N-component vectors denoting the input states and the weight states, respectively. For every $\mathbf{W}$, the network defines a map from $\mathbf{S}$ to σ. Thus the weight space corresponds to a class of functions, a class that is constrained by the architecture of the network. Learning can be thought of as a search through weight space to find a network with desired properties.

In supervised learning, the weights of the network are tuned so that it approximates as closely as possible a target function $\sigma_0(\mathbf{S})$. One way of achieving this is to provide a set of *examples* consisting of P input-output pairs $(\mathbf{S}^l, \sigma_0(\mathbf{S}^l))$, with $l = 1, \ldots, P$. We assume that each input $\mathbf{S}^l$ is chosen at random from the entire input space according to some normalized a priori measure denoted $d\mu(\mathbf{S})$. The examples can be used to construct a *training energy*

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

where the *error function* $\epsilon(\mathbf{W}; \mathbf{S})$ is some measure of the deviation of the network's output $\sigma(\mathbf{W}; \mathbf{S})$ from the target output $\sigma_0(\mathbf{S})$. The error function should be zero whenever the two agree, and positive everywhere else. A popular choice is the quadratic error function

$$\epsilon(\mathbf{W}; \mathbf{S}) = \frac{1}{2}[\sigma(\mathbf{W}; \mathbf{S}) - \sigma_0(\mathbf{S})]^2 \ . \tag{2.2}$$

puter experiments of supervised learning[21, 31, 32]. In other cases, however, one observes a rather sharp improvement when a critical number of examples is reached[20, 21, 30].

These seemingly conflicting behaviors have analogies in psychological studies of animal learning. The dichotomy between gradual and sudden learning lay at the heart of the long-standing controversy between the behaviorist[33] and the Gestalt[34] approaches to learning in the cognitive sciences. In this debate the underlying assumption has been that a learning process that is based on incremental modifications of the internal structure of the system can yield only gradual improvements in performance. The sudden appearance of concept understanding was therefore related to preexisting strong biases towards the learnt concept, or to mysterious holistic learning mechanisms.

In the present study we show that in large systems, a sudden emergence of good generalization ability can arise even within the framework of incremental microscopic training algorithms. We analyze the conditions under which such *discontinuous transitions to perfect learning* occur. Also, we study the asymptotic forms of learning curves in cases where they are smooth. Other issues addressed in this work include:

1. the consequences of the annealed approximation for learning in large networks and the scope of its validity,
2. the possible emergence of spin-glass phenomena associated with the frustration and randomness induced by the random sampling of examples,
3. how the nonlinearities inherent in the network operation affect its performance,
4. and the effect of stochastic training (noise in the learning dynamics) on generalization performance.

We address these issues by combining general results from the SM formulation with detailed analytical and numerical studies of specific models. In this work, we consider only single-layer perceptron models, which are significantly poorer in computational capabilities than multilayer networks. Somewhat surprisingly, even these simple models exhibit nontrivial generalization properties. Some of these properties have been shown to exist also in two-layer models also[35]. Furthermore the strong nonlinearities generated in multilayer systems can be mimicked to some extent by strong constraints on the range of values of the perceptron weights. Indeed, even the loading problem in a perceptron with binary weights is a hard problem both theoretically and computationally (see e.g. Mézard and Krauth[36] and also [37]). Here we study learning from examples in a perceptron with real-valued weights as well as with binary weights.

A proper treatment of learning from examples requires the use of sophisticated techniques from the SM of disordered systems, in particular the replica method. In order to facilitate the understanding of the results, we present in Part I two relatively simple approximations. One is a *high temperature* limit, and the other is the

as the Vapnik-Chervonenkis (VC) dimension[14–16]. Generally, they are independent of the specific learning algorithm or distribution of examples. The generality of the PAC approach is also its main deficiency, since it is dominated by the worst case, atypical behavior. Our statistical mechanical approach thus differs considerably from the PAC learning theory in that it can provide precise quantitative predictions for the typical behavior of specific learning models.

The SM formalism can also be applied to certain learning models for which few PAC results are yet known. Despite recent works which extend the original PAC framework[12, 17], most PAC theorems apply to *realizable* tasks, namely tasks that can be performed perfectly by the network, given enough examples. In many real life problems the target task can only be approximated by the assumed architecture of the network, so the task is *unrealizable*. In addition, many of the PAC learning results are limited to networks with threshold decision elements, although in many applications analog neurons are used. The SM approach is close in its spirit, though not in its scope and results, to the Bayesian information-theoretic approach, recently applied also to continuous networks[18, 17].

A SM approach to learning from examples was first proposed by Carnevali and Patarnello[19], and by Denker et al.[20], and further elaborated by Tishby et al.[21]][22]]. Studies of learning a classification task in a perceptron can be found in Hansel and Sompolinsky[23] and del Giudice et al.[24], using spin-glass techniques. Gardner and Derrida[25] and Györgyi and Tishby[26, 27] have used these methods for studying learning of a perceptron rule. Related models have been studied in Refs. [28]–[29]. However the extent of applicability of results gained from these specific toy models to more general circumstances has remained unknown.

Recently an interesting attempt to characterize generic generalization performance has been put forward by Schwartz et al.[30]. This work, however, suffers from two basic deficiencies. First, the analysis relies on an approximation whose validity is not addressed. It turns out that this approximation is closely related to the well-known *annealed approximation* (AA) in the statistical mechanics of random systems. Although the AA simplifies enormously the theoretical analysis of these complex systems, in most interesting cases it is known to be unreliable, sometimes even in its qualitative predictions. The second problem is that no attention has been given to the dependence of performance on system size. In fact the behavior of large systems may be quite different from that of *small-size* ones, and the analysis of emergent behavior becomes more involved with increased system size. Moreover, even medium finite size systems behave remarkably close to the asymptotic large size theory, as our simulations demonstrate.

In the present study we attempt to characterize the generic behaviors of learning from examples in large layered networks. In particular we investigate the expected rate of improvement of the generalization with an increasing number of examples, denoted by the *generalization curve*. The PAC theory bounds the generalization curve by an inverse power law. Such a gradual improvement has also been observed in com-

I. INTRODUCTION

In recent years, many attempts have been made to train layered feedforward neural networks to perform computational tasks, such as speech recognition[1] and generation[2], handwriting recognition[3], protein structure prediction[4], and many others. These networks have also been used as models for neurobiological systems[5, 6], and may even serve as useful metaphors for cognitive processes associated with learning, generalization, and concept formation[7].

Learning in neural networks, as well as in other parametric models[8], has also attracted considerable theoretical interest. The activity in this area has centered on two issues. The first is the question of representation, or *realizability*. Given a network of some architecture and size, is there a set of weights that makes the network perform the desired task? The second is the question of *learning*. Given that such a network exists, can its structure and parameters be found with a reasonable amount of time, computational resources, and training data?

Here we focus on the question of learning. We further restrict our scope to supervised learning from examples, which relies on a training set consisting of examples of the target task. The training algorithm uses the examples to find a set of network weight values that perform the task well. The most widely used class of training algorithms works by optimizing a suitable cost function that quantifies the error on the training set.

Such learning algorithms have several potential difficulties. For example, the algorithms may become trapped in local minima that are far from optimal. Furthermore, finding good minima may require prohibitively long convergence times. Finally, there is no guarantee that a good performance on a training set also leads to a good performance on novel inputs. This last issue, the ability of adaptive systems to *generalize* from a limited number of examples, is the focus of the present work. Understanding the determinants of generalization ability is crucial for devising machine learning strategies, as well as for obtaining insight into learning processes in biological systems.

Our study is based on a statistical mechanical (SM) formulation of learning in neural networks. The training procedure is assumed to be stochastic, leading to a Gibbs distribution of network weights. The performances of the system on the training set as well as on novel inputs are calculated as appropriate *thermal* averages on the Gibbs distribution in weight space and *quenched* averages on the sampling of examples. These averages provide an accurate account of the *typical* behavior of large networks.

The currently dominant approach in computational learning theory is based on Valiant's learning model and on the notion of *Probably Almost Correct* (PAC) learning[9, 10]. The main achievements of this approach are general bounds on the probability of error on a novel example for a given size of the training set[11, 12], as well as classification of learning problems according to their time complexity[9, 13]. Most of these (sample complexity) combinatorial bounds depend on the specific structure of the model and the complexity of the task only through a single number, known

Submitted to Physical Review A

Statistical Mechanics of Learning From Examples I. General Formulation and Annealed Approximation

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 equilibrium statistical mechanics. Two simple approximations to the exact quenched theory are presented: the high temperature limit and the annealed approximation. Within these approximations, we study four models of perceptron learning of *realizable* target rules. In each model, the target rule is perfectly realizable because it is another perceptron of identical architecture. We focus on the *generalization curve*, i.e. the average generalization error as a function of the number of examples. The case of continuously varying weights is considered first, for both linear and boolean output units. In these two models, learning is gradual, with generalization curves that asymptotically obey inverse power laws. Two other model perceptrons, with weights that are constrained to be discrete, exhibit *sudden* learning. For a linear output, there is a first-order transition occurring at low temperatures, from a state of poor generalization to a state of good generalization. Beyond the transition, the generalization curve decays exponentially to zero. For a boolean output, the first order transition is to *perfect generalization* at all temperatures. Monte Carlo simulations confirm that these approximate analytical results are quantitatively accurate at high temperatures and qualitatively correct at low temperatures. For *unrealizable* rules the annealed approximation breaks down in general, as we illustrate with a final model of a linear perceptron with unrealizable threshold. Finally, we propose a general classification of generalization curves in models of realizable rules.

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

Typeset Using *REVTEX*