

# Vapnik-Chervonenkis entropy of the spherical perceptron

P. Riegler\* and H. S. Seung  
Bell Laboratories, Lucent Technologies  
Murray Hill, NJ 07974

revised December 2, 1996

## Abstract

Perceptron learning of randomly labeled patterns is analyzed using a Gibbs distribution on the set of realizable labelings of the patterns. The entropy of this distribution is an extension of the Vapnik-Chervonenkis (VC) entropy, reducing to it exactly in the limit of infinite temperature. The close relationship between the VC and Gardner entropies can be seen within the replica formalism.

There has been recent progress towards understanding the relationship between the statistical physics and Vapnik-Chervonenkis (VC) approaches to learning theory[1, 2, 3, 4]. The two approaches can be unified in a statistical mechanics based on the VC entropy. This paper treats the case of learning randomly labeled patterns, or the *capacity* problem, and extends some of the results of previous work[5, 6] to finite temperature. As will be explained in a companion paper, this extension is important for treating the *generalization* problem, which occurs in the context of learning patterns labeled by a target rule.

Our general framework is illustrated for the simple perceptron  $\text{sgn}(w \cdot x)$ , which maps an  $N$ -dimensional real-valued input  $x$  to a  $\pm 1$ -valued output. Given a sample  $X = (x_1, \dots, x_m)$  of inputs, the weight vector  $w$  determines a labeling  $L = (l_1, \dots, l_m)$  of the sample via  $l_i = \text{sgn}(w \cdot x_i)$ . The weight vector  $w$  defines a normal hyperplane that separates the positive from the negative examples. The training error of a labeling  $L$  with respect to a reference labeling  $L^0$  is defined by

$$e_t(L, L^0) = \frac{1}{m} \sum_{i=1}^m \frac{1 - l_i l_i^0}{2}, \quad (1)$$

and is just the fraction of different labels in the two labelings. We consider the case in which the reference labeling is chosen at random, and address the issue of

---

\*permanent address: Institut für Theoretische Physik, Universität Würzburg, D-97074 Würzburg, Germany

*capacity*[7, 8]. Namely, how many randomly labeled inputs can the perceptron learn with accuracy  $\epsilon_t$ ? This issue is distinct from, but related to, the issue of *generalization*, which arises when the inputs are labeled by some underlying target rule[9].

The class of perceptrons can be visualized as a sphere in  $N$ -dimensional space, since only the direction of  $w$  matters in  $\text{sgn}(w \cdot x)$ . The labelings of the sample can be visualized geometrically as the cells of a tessellation of this sphere[10]. Each input  $x_i$  cuts the sphere into two hemispheres. The weight vectors of one hemisphere classify the input as positive, and those of the other classify it as negative. A sample of  $m$  inputs cuts the sphere into many cells, as shown in Figure 1. Each cell consists of an equivalence class of weight vectors that give the sample the same labeling. Although there are  $2^m$  possible labelings, not all of them can necessarily be realized.

The volume  $V(L, X)$  of a cell can be written as

$$V(L, X) = \int dw \prod_{i=1}^m \theta(l_i w \cdot x_i) , \quad (2)$$

where the integral is over a uniform measure on the unit sphere. The indicator function  $V^0(L, X) \equiv \lim_{n \rightarrow 0} V^n(L, X)$  takes the value one for all realizable labelings and zero for others.

Using the training error, a Gibbs distribution at inverse temperature  $\beta = 1/T$  can be defined on the set of realizable labelings:

$$P(L) = Z^{-1} V^0(L, X) e^{-m\beta e_t(L, L^0)} , \quad (3)$$

where the partition function is defined by

$$Z = \sum_L V^0(L, X) e^{-m\beta e_t(L, L^0)} . \quad (4)$$

The free energy is defined by

$$-\beta f(\alpha, \beta) = \frac{1}{N} \langle \langle \log Z \rangle \rangle , \quad (5)$$

where the quenched average  $\langle \langle \rangle \rangle$  is over the sample  $X$  and the reference labeling  $L^0$ . The thermodynamic limit  $N, m \rightarrow \infty$  is taken with the ratio  $\alpha \equiv m/N$  fixed, and the free energy normalized by  $N$  to make it an intensive quantity. The entropy is defined as

$$s(\alpha, \beta) = - \sum_L \langle \langle P(L) \log P(L) \rangle \rangle . \quad (6)$$

In the limit of infinite temperature, the Gibbs distribution assigns equal measure to every realizable labeling, so that the entropy is just the average of the logarithm of the number of realizable labelings, as the VC entropy was originally defined [9]. The equations (3) and (6) generalize the VC entropy to finite temperature.

The above canonical formulation took  $\alpha$  and  $\beta$  as thermodynamic state variables. The complementary microcanonical formulation replaces  $\beta$  by  $\epsilon_t$ , and defines the entropy as the average of the logarithm of the number of realizable labelings with training error  $\epsilon_t$ [11]. The entropy obeys the upper bound

$$s(\alpha, \epsilon_t) \leq \frac{1}{N} \log \binom{m}{m\epsilon_t} \leq \alpha \mathcal{H}(\epsilon_t), \quad (7)$$

where  $\mathcal{H}(p) = -p \log p - (1-p) \log(1-p)$  is the binary entropy function.

The canonical and microcanonical definitions of the entropy are equivalent in the thermodynamic limit. The formulations are related by a Legendre transformation,

$$s(\alpha, \epsilon_t) = \min_{\beta} \{ \alpha \beta \epsilon_t - \beta f(\alpha, \beta) \}, \quad (8)$$

which implies that the conjugate variables of error and temperature are related via  $\epsilon_t = \partial(\beta f)/\partial\beta$ .

The capacity  $\alpha_c(\epsilon_t)$  of the perceptron is defined as the maximum number of randomly labeled examples that can be learned with error  $\epsilon_t$ . It is found by solving the equation  $s(\alpha, \epsilon_t) = 0$  for  $\alpha$ . For  $\alpha > \alpha_c(\epsilon_t)$ , there is no realizable labeling with error  $\epsilon_t$ .

**Annealed theory** As a preliminary to calculating the free energy (5), we calculate a related quantity, the annealed free energy

$$-\beta f^{ann}(\alpha, \beta) = \frac{1}{N} \log \langle\langle Z \rangle\rangle. \quad (9)$$

The annealed average  $\langle\langle Z \rangle\rangle$  of the partition function can be computed via the replica trick by substituting  $V^n(L, X)$  for  $V^0(L, X)$  in (4), where  $n$  is a positive integer. Assuming replica symmetry, the annealed free energy is given by

$$\begin{aligned} -\beta f^{ann}(\alpha, \beta) &= \lim_{n \rightarrow 0} \min_q \left\{ \frac{n}{2} \log(1-q) + \frac{1}{2} \log \left( 1 + \frac{nq}{1-q} \right) \right. \\ &\quad \left. + \alpha \log \int Dx H^n \left( x \sqrt{\frac{q}{1-q}} \right) + \alpha \log(1 + e^{-\beta}) \right\} \quad (10) \end{aligned}$$

The  $n \rightarrow 0$  limit must be taken differently in the two different regions  $\alpha < 2$  and  $\alpha > 2$  of the phase diagram in Figure 2. The order parameter  $q$  is the typical overlap between two weight vectors from the same cell. Since  $q < 1$  for  $\alpha < 2$ , the limit  $n \rightarrow 0$  eliminates all terms except the last in (10), and the result  $s^{ann} = \alpha \mathcal{H}(\epsilon_t)$  follows by Legendre transformation. In this “linear” phase, the annealed entropy increases linearly with  $\alpha$ , and saturates the bound (7). Finding the value of  $q$  requires expanding the annealed free energy (10) to first order in  $n$ . It increases with  $\alpha$ , since the size of a typical cell is decreasing, but is independent of temperature/error. For  $\alpha > 2$ , the order parameter  $q$  is equal to one, meaning that the typical cell has vanishingly small size. The limits  $n \rightarrow 0$  and  $q \rightarrow 1$  in (10) must be taken with  $n/(1-q)$  held constant. After Legendre transformation of the result, we obtain the annealed entropy in the “sublinear” phase ( $\alpha > 2$ ).

To summarize, the annealed entropy in the linear and sublinear phases is

$$s^{ann}(\alpha, \epsilon_t) = \begin{cases} \alpha \mathcal{H}(\epsilon_t), & \alpha \leq 2, \\ \alpha \log \alpha - (\alpha - 1) \log(\alpha - 1) + \alpha \mathcal{H}(\epsilon_t) - \alpha \log 2, & \alpha > 2, \end{cases} \quad (11)$$

and is shown in Figure 3. Either temperature or training error can be used as a state variable in the free energy and entropy. Converting between these variables is done using the thermodynamic relationship

$$\epsilon_t = \frac{\partial(\beta f^{ann})}{\partial \beta} = \frac{1}{e^\beta + 1} \quad (12)$$

Since there is only a single cell with zero training error, we have  $s^{ann}(\alpha, \epsilon_t = 0) = 0$  for  $\alpha < 2$ . For  $\alpha > 2$ , the annealed entropy is negative, because there are no cells with zero training error, with probability approaching one in the thermodynamic limit. Thus  $\alpha_c = 2$  for  $\epsilon_t = 0$ .

For any positive  $\epsilon_t < 1/2$ , the annealed entropy increases linearly with  $\alpha$  for  $\alpha < 2$ . For  $\alpha > 2$  it initially increases sublinearly, but then decreases to zero. The zero entropy line  $\alpha_c(\epsilon_t)$  is depicted in Figure 2. By the upper bound  $s^{ann}(\alpha, \epsilon_t) \geq s(\alpha, \epsilon_t)$ , this is an upper bound for the capacity of the perceptron to store randomly labeled examples with finite error  $\epsilon_t$ . This bound on capacity increases with training error.

For  $\epsilon_t = 1/2$ , the annealed entropy is monotonically increasing. This reflects the obvious point that the perceptron has infinite capacity if the training error is one half. According to (12),  $\epsilon_t = 1/2$  corresponds to infinite temperature, for which the Gibbs distribution is uniform over all realizable labelings.

The annealed entropy is an upper bound on the quenched entropy. How tight a bound is it? According to a classic result in combinatorial geometry[12, 7], the total number of realizable labelings is exactly

$$2 \sum_{i=0}^{N-1} \binom{m-1}{i} \quad (13)$$

for any  $x_1, \dots, x_m$  in general position, and hence does not fluctuate for any well-behaved distribution on the  $x_i$ [7]. The logarithm of this number gives the quenched entropy at infinite temperature, or

$$s(\alpha, T = \infty) = \begin{cases} \alpha \log 2, & \alpha \leq 2, \\ \alpha \log \alpha - (\alpha - 1) \log(\alpha - 1) & \alpha > 2. \end{cases} \quad (14)$$

Comparison with (11) shows that the annealed and quenched entropies are equal at infinite temperature. Since the total number of realizable labelings is nonfluctuating, the order of the logarithm and the expectation in (5) does not matter.

**Quenched theory** At finite temperature, there is no reason to expect that the annealed and quenched entropies are equal. The quenched entropy can be calculated by introducing a second replication to treat the average of the logarithm in the free energy(5). With the replica symmetric ansatz, the quenched

free energy is given by

$$\begin{aligned}
-\beta f &= \lim_{n \rightarrow 0} \min_{q, Q} \left\{ \frac{n}{2} \log(1-q) + \frac{1}{2} \log \left( 1 + n \frac{q-Q}{1-q} \right) + \frac{1}{2} \frac{nQ}{1-q+n(q-Q)} \right. \\
&\quad \left. + \alpha \int Dx \log \int Dy \sum_{\sigma=\pm 1} e^{-\beta(1-\sigma)/2} H^n \left( \frac{\sqrt{Q}x + \sqrt{q-Q}y}{\sigma\sqrt{1-q}} \right) \right\} \quad (15)
\end{aligned}$$

In addition to the order parameter  $q$ , there is a new order parameter  $Q$  corresponding to the typical overlap between weight vectors drawn from different cells with the same training error.

In the linear phase ( $\alpha < 2$ ), the entropy and the order parameter  $q$  are the same as in the annealed calculation. The new order parameter  $Q$  is determined by minimization of the  $\mathcal{O}(n^2)$  term of the free energy. As temperature rises from zero to infinity,  $Q$  decreases from  $q$  to zero, as shown in Figure 4. As a function of  $\alpha$ , there is an interesting nonmonotonic behavior of  $Q$  near  $\alpha = 2$ .

Figure 4 also shows the results for the order parameters  $q$  and  $Q$  obtained by simulations. For a fixed but random set of  $\alpha N$  pattern vectors and labels a random fraction  $\epsilon_t$  of the labels was flipped and a first solution found by linear programming. In order to obtain  $q$ , 50 further solutions within the same cell were generated by Monte Carlo sampling: Starting from a solution within the cell a direction was picked at random and the two boundaries of the cell in this direction were determined. Then a point on the arc connecting the boundary points was chosen at random. This procedure was repeated 100000 times, and a Monte Carlo sample was taken every 2000 steps, for a total of 50 samples. To compute the value of  $Q$ , 50 cells were generated by flipping a fraction  $\epsilon_t$  of the labels for a fixed set of patterns, and Monte Carlo sampling was performed for each of these cells. The results shown in figure 4 were obtained by averaging over 100 independent experiments, in each of which a set of pattern vectors was drawn randomly. The deviation of  $Q$  at  $\alpha = 2$  from the theoretical result might be due to finite size effects. In general we observe that the fluctuations of  $Q$  increase as  $\alpha \rightarrow 2$ .

In the sublinear phase ( $\alpha > 2$ ), the order parameter  $q = 1$ , so that the typical cell has become vanishingly small. As in the annealed case, the limits  $n \rightarrow 0$  and  $q \rightarrow 1$  must be taken with the ratio  $n/(1-q)$  held constant, yielding

$$\begin{aligned}
-\beta f &= \min_{u, v} \left\{ -\log u + \frac{1}{2}(u^2 - 1)v^2 \right. \\
&\quad \left. + \alpha \int Dx \log(1 + (e^{-\beta} - 1)H(vx)) \right. \\
&\quad \left. + u(1 + (e^{-\beta} - 1)H(-uvx)) \exp(-(u^2 - 1)v^2 x^2/2) \right\} \quad (16)
\end{aligned}$$

where  $u \equiv 1/\sqrt{1+n(1-Q)/(1-q)}$  and  $v \equiv \sqrt{Q/(1-Q)}$ . The resulting entropy is graphed in Figure 3, and is smaller than the annealed bound (11). The zero entropy line  $\alpha_c(\epsilon_t)$  is shown in the phase diagram of Figure 2, and is at lower  $\alpha$  than the zero entropy line from the annealed theory.

**Gardner vs. VC entropy** Previous work on the statistical mechanics of learning from examples has utilized a Gibbs distribution on the space of functions, as pioneered by Elizabeth Gardner[8, 13]. In the Gardner formulation, the definition of capacity depends on whether the function class is continuous or discrete. For continuous function classes like the spherical perceptron, the Gardner entropy diverges to  $-\infty$  at capacity[8, 14, 15]. For discrete function classes like the Ising perceptron, the Gardner entropy vanishes at capacity[16, 17]. Here we have taken a different approach involving a Gibbs distribution on the set of realizable labelings induced by the function class. The VC entropy vanishes at capacity, regardless of whether the function class is continuous or discrete. This is because the set of labelings is finite, for both finite and infinite function classes.

Replica calculations of the Gardner and VC entropies are very closely related. The  $\mathcal{O}(n)$  term in the expansion of the annealed free energy (10) is the same as the free energy of the Gardner calculation. The replica symmetric VC free energy (15) resembles the one-step replica symmetry breaking (RSB) Gardner free energy[14, 15], and gives a very similar value for the capacity. However, it is not exactly the same. Probably a full RSB calculation is necessary to obtain the correct capacity.

Only the learning of randomly labeled examples has been analyzed here. When the examples are drawn from a target function, the issue of generalization to examples not seen during training is of great importance[9, 18]. This issue can also be addressed with a statistical mechanics of VC entropy, as will be discussed elsewhere.

**Acknowledgments** This work was supported by Bell Laboratories, the Deutsche Forschungsgemeinschaft and a travel grant by the Studienstiftung des deutschen Volkes.

## References

- [1] J. M. R. Parrondo and C. Van den Broeck. Vapnik-chervonenkis bounds for generalization. *J. Phys.*, A26:2211–2223, 1993.
- [2] A. Engel and C. Van den Broeck. Systems that can learn from examples: replica calculation of uniform convergence bounds for the perceptron. *Phys. Rev. Lett.*, 71:1772–1775, 1993.
- [3] D. Haussler, M. Kearns, H. S. Seung, and N. Tishby. Rigorous learning curve bounds from statistical mechanics. In *Proceedings of the Seventh Annual ACM Conference on Computational Learning Theory*, pages 76–87, New York, 1994. ACM.
- [4] M. Opper. Learning and generalization in a two-layer neural network: The role of the vapnik-chervonenkis dimension. *Phys. Rev. Lett.*, 72:2113–2116, 1994.

- [5] A. Engel and M. Weigt. Multifractal analysis of the coupling space of feedforward neural networks. *Phys. Rev.*, E53:R2064–7, 1996.
- [6] R. Monasson and D. O’Kane. Domains of solutions and replica symmetry breaking in multilayer neural networks. *Europhys. Lett.*, 27:85–90, 1994.
- [7] T. M. Cover. Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition. *IEEE Trans. Electronic Comput.*, 14:326–334, 1965.
- [8] E. Gardner. The space of interactions in neural network models. *J. Phys.*, A21:257–270, 1988.
- [9] V. N. Vapnik. *Estimation of Dependences based on Empirical Data*. Springer-Verlag, New York, 1982.
- [10] M. Opper and D. Haussler. Generalization performance of Bayes optimal classification algorithm for learning a perceptron. *Phys. Rev. Lett.*, 66:2677–2680, 1991.
- [11] H. S. Seung. Annealed theories of learning. In J.-H. Oh, C Kwon, and S. Cho, editors, *Neural networks: the statistical mechanics perspective*, pages 32–41, Singapore, 1995. World Scientific.
- [12] L. Schläfli. Theorie der vielfachen Kontinuität (1852). In *Gesammelte Mathematische Abhandlungen*, volume 1, pages 177–392. Birkhäuser, Basel, 1950.
- [13] E. Gardner and B. Derrida. Optimal storage properties of neural network models. *J. Phys.*, A21:271–284, 1988.
- [14] R. Erichsen and W. K. Thuemann. Optimal storage of a neural network model: a replica symmetry breaking solution. *J. Phys.*, A26:L61–L68, 1993.
- [15] P. Majer, A. Engel, and A. Zippelius. Perceptrons above saturation. *J. Phys.*, A26:7405–16, 1993.
- [16] W. Krauth and M. Mézard. Storage capacity of memory networks with binary couplings. *J. Phys. (Paris)*, 50:3057–3066, 1989.
- [17] W. Krauth and M. Opper. Critical storage capacity of the  $j = \pm 1$  neural network. *J. Phys.*, 22:L519–L523, 1989.
- [18] H. S. Seung, H. Sompolinsky, and N. Tishby. Statistical mechanics of learning from examples. *Phys. Rev.*, A45:6056–6091, 1992.

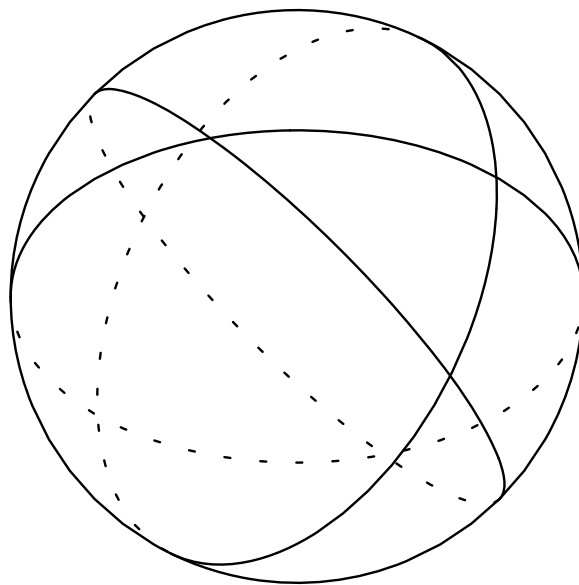


Figure 1: Geometry of the perceptron function class. Since the perceptron is parametrized by the direction of an  $N$ -dimensional weight vector, the class of perceptrons can be thought of as an  $N$ -dimensional sphere. Each input divides the sphere into two halves, consisting of weight vectors that classify the input as positive and negative respectively. A sample of inputs divides the sphere into many cells, where each cell corresponds to a distinct labeling of the sample. The VC entropy is defined as the expectation of the logarithm of the number of cells.

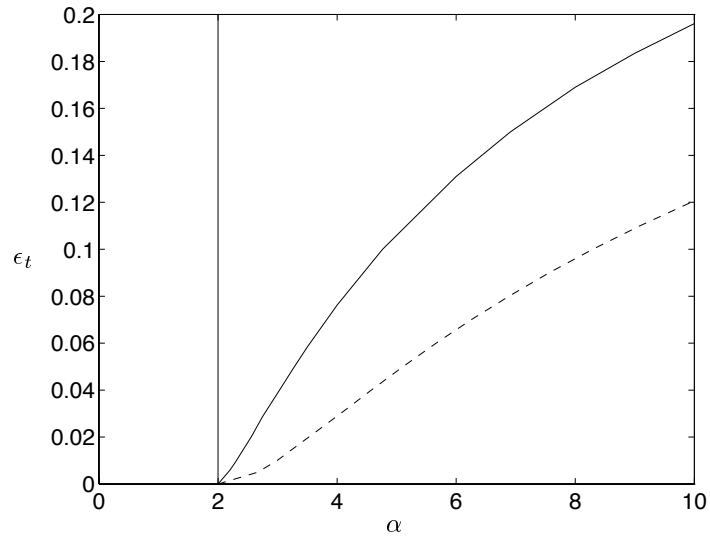


Figure 2: Phase diagram. For  $\alpha < 2$ , the VC entropy increases linearly with  $\alpha$ . For  $\alpha > 2$ , the behavior of the VC entropy is sublinear. The zero entropy line  $\alpha_c(\epsilon_t)$  marks the point beyond which there are no realizable labelings. The solid line is from the quenched entropy, and the dashed line from the annealed.

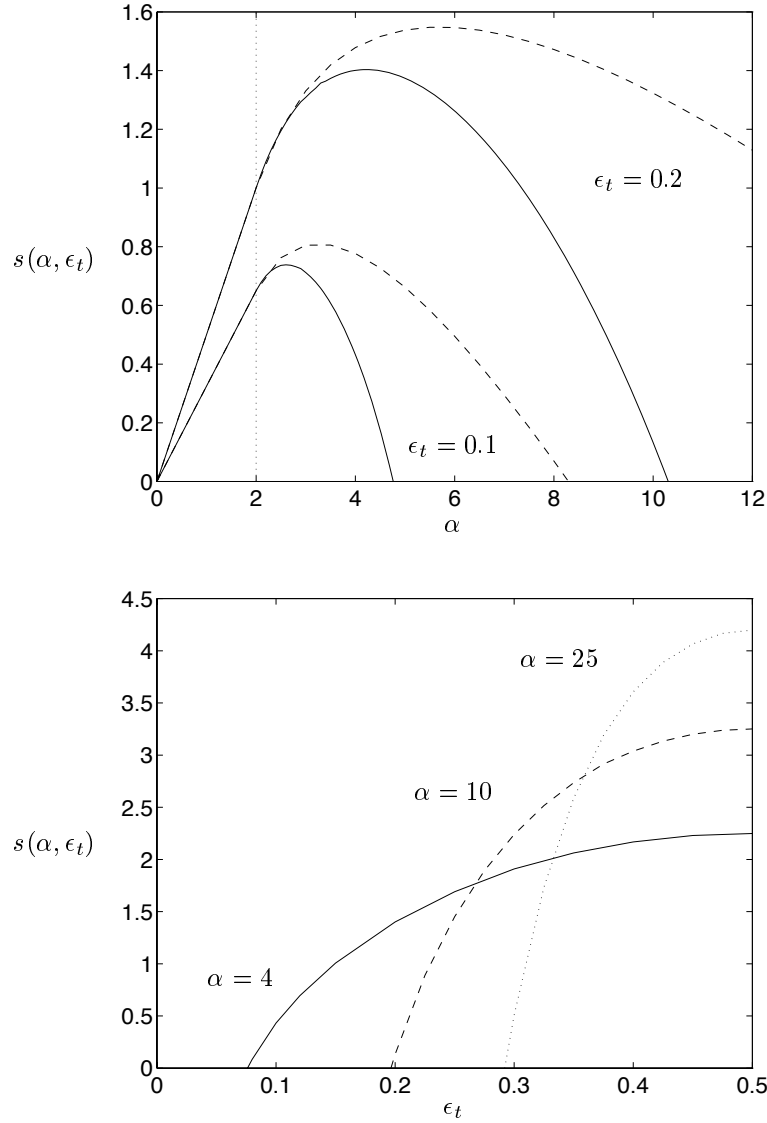


Figure 3: Vapnik-Chervonenkis entropy. (a) VC entropy vs.  $\alpha$  for several values of  $\epsilon_t$ . The entropy increases linearly for  $\alpha < 2$ , saturating its upper bound (7). Above  $\alpha > 2$ , the entropy increases sublinearly and then decreases to zero for any training error  $\epsilon_t$  strictly less than  $1/2$ . The quenched entropy is drawn with solid lines, and the annealed with dashed lines. (b) VC entropy vs.  $\epsilon_t$  for several values of  $\alpha$ . For  $\alpha > 2$  the range of realizable  $\epsilon_t$  shrinks. In the limit of infinite  $\alpha$ , the only realizable training error is  $1/2$ .

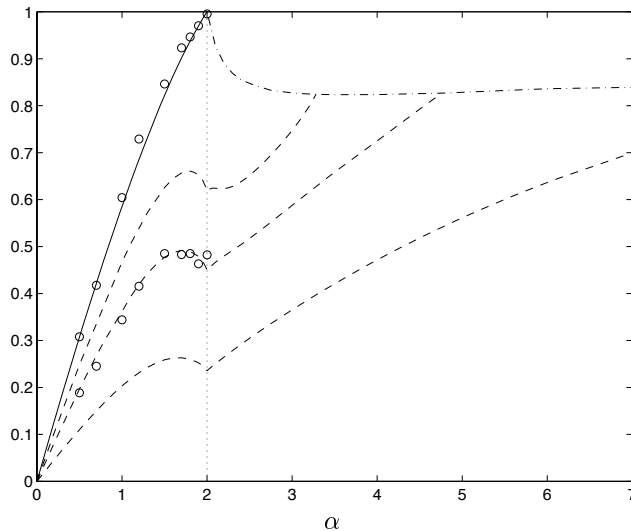


Figure 4: Order parameters. The order parameter  $q$  (solid line) is the typical overlap between two weight vectors from the same cell, and is independent of temperature/error. It increases continuously from zero until it reaches unity at  $\alpha = 2$ . The order parameter  $Q$  (dashed line) is the typical overlap between two weight vectors from different cells. Its behavior as a function of  $\alpha$  is shown for different values of the training error  $\epsilon_t$  ( $\epsilon_t = 0.05, 0.1, 0.2$ ). It lies between  $q$  and 0, and is lower for higher training errors. The dashed-dotted line represents the values of  $Q$  at capacity. Simulations are shown for  $\epsilon_t = 0, 0.1$  and were obtained for a system of size  $N = 100$  averaged over 100 independent runs. Error bars are of the size of the symbols or smaller.