1 On the Annealed VC Entropy for Margin Classifiers: A Statistical Mechanics Study

Manfred Opper
Neural Computing Research Group, Aston University, Birmingham B4 7ET, UK.
opperm@aston.ac.uk

Using techniques from Statistical Physics, the annealed VC entropy for hyperplanes in high dimensional spaces is calculated as a function of the margin for a spherical Gaussian distribution of inputs.

1.1 Introduction

The Vapnik Chervonenkis (VC) approach to statistical learning theory [9, 10] allows to express the complexity of a family of statistical predictors in terms of entropic quantities, the so called VC entropies. For the case of a binary classifier, these entropies give the logarithm of the number of different classifications of a set of input points which are realizable by the family of classifiers. Upper bounds on these entropies can be expressed by a single combinatorial quantity, the VC dimension. Classifiers with large VC complexities can have a large deviation between empirical error and generalization error, which for the case of empirical risk minimzation may lead to strong overfitting, when not enough training data are available. For the case of learning in neural networks, the VC approach has been critizised for overestimating the complexities and giving too pessimistic bounds for a practical application in model selection.

effective dimensions

Recently, for margin classifiers and support vector machines it has been shown that if global VC dimensions are replaced by effective, data dependent dimensions (the so called *fat-shattering dimensions*[1]), reliable estimates for optimally generalizing models can be obtained [10]. In these cases, the effective VC dimensions depend on the size of the by which positive and negative training inputs can be separated. Besides from general bounds and simulations, such results may be further understood from another approach to computational learning theory which

has its origin in statistical mechanics. Using techniques from the theory of disordered systems, a huge variety of results for the typical learning behaviour of large neural networks have been obtained in the last years. For a review see e.g. [12],[5] and [7]. The approach enables exact calculations for generalization errors and other properties of neural networks (assuming specific 'nice' distributions of examples) in the limit where the dimension of input space and the size of the set of examples are both very large. Although some of these techniques (like many in the field of Theoretical Physics) have not been made fully rigorous sofar, this approach yields often new important results on which other, more general methods can be tested.

In the following, I will present a calculation of the *annealed* VC entropy for classifications by hyperplanes (perceptrons) as a function of the margin. The method follows a recent publication [6] which aimed at calculating the capacity of a toy neural network. This latter model can be interpreted as a problem of unsupervised learning in a perceptron where the output variables must be chosen in such a way that the margin between positive and negative examples is maximal.

1.2 VC–Entropy

Let us assume a training set of ℓ input/output pairs $(\mathbf{x}_1, y_1) \ldots, (\mathbf{x}_\ell, y_\ell)$ for a binary classifier which are drawn independently at random from a fixed distribution. In the following, $\mathbf{x}_1^\ell$ stands for the set of inputs $\mathbf{x}_1, \ldots, \mathbf{x}_\ell$ and y_1^ℓ for the sequence of outputs $y_1, \ldots, y_\ell$. The VC approach enables us to bound the deviations between the training error $E_t(\mathbf{x}_1^\ell, y_1^\ell, c)$ (the number of misclassifications on the training set) and the generalization error $e_g(c)$ (the probability of a misclassification) over a family of classifiers $c \in \mathcal{F}$. E.g., it has been shown [10] that

$$\Pr\left(\sup_{c\in\mathcal{F}} |E_t(\mathbf{x}_1^\ell, y_1^\ell, c) - e_g(c)| > \varepsilon\right) \leq 4\exp\left[H_{ann}(2\ell) - \ell\varepsilon^2\right] \tag{1.1}$$

annealed entropy

where the is defined as

$$H_{ann}(\ell) = \ln\left\langle \mathcal{N}(\mathbf{x}_1^\ell)\right\rangle, \tag{1.2}$$

and where $\mathcal{N}(\mathbf{x}_1^\ell) \leq 2^\ell$ is the number of classifications (or) of ℓ inputs which are realizable by going through the classifiers $c \in \mathcal{F}$, and the brackets $\langle \ldots \rangle$ denote expectations with respect to the distribution of the inputs.

Perceptrons classify inputs $\mathbf{x} \in \mathbf{R}^N$ by hyperplanes via $y = \mathrm{sgn}(\mathbf{w}\cdot\mathbf{x} + b) \in \{-1, +1\}$ (the weight vector $\mathbf{w} \in \mathbf{R}^N$ is normal to the class separating plane and $b \in \mathbf{R}$ is a bias). Throughout this paper we will be concerned with hyperplanes *through the origin* i.e. $b = 0$ only. For this case, it is well known that

$$\mathcal{N}(\mathbf{x}_1^\ell) = 2\sum_{i=0}^{N-1}\binom{\ell-1}{i}, \tag{1.3}$$

independently of the position of the inputs $\mathbf{x}_i$ (as long as they are in general position).

For more complicated types of classifiers, exact expressions for the VC entropies are hard to obtain. A remarkable and general combinatorial theorem, proved for the first time by Vapnik and Chervonenkis in the 1960's [11], however gives the general bound

$$\mathcal{N}(\mathbf{x}_1^\ell) \leq \sum_{i=0}^{h} \binom{\ell}{i}, \tag{1.4}$$

in terms of a single number h, the VC dimension of the family $\mathcal{F}$ of classifiers. This result allows us to obtain distribution independent bounds.

If we restrict the family of perceptrons to the subclass of all those which achieve a margin γ larger than some positive value κ, i.e.

$$\gamma = \max_{||\mathbf{W}||=1} \min_i \; y_i(\mathbf{w} \cdot \mathbf{x}_i) > \kappa, \tag{1.5}$$

the corresponding VC dimension can be much smaller. A bound on the corresponding VC dimension h_κ was given in [9]

VC dimension for margin

$$h_\kappa = \min\left(\left[\frac{R^2}{\kappa^2}\right] + 1, N\right), \tag{1.6}$$

where R is the radius of the minimal sphere containing all inputs. It is not directly possible to implement this bound on the VC dimension into the confidence bound (1.1), when the margin is not fixed in advance but taken from a classifier trained on a specific sample. This is because the bound (1.1) requires a fixed, *a priori* chosen (nonrandom) family of classifiers. Somewhat more complicated bounds have been proved recently for the data dependent case [8]. Nevertheless, we expect that our calculations of the annealed entropy for a fixed margin may also give at least a qualitative picture for the data dependent case.

1.3 The Thermodynamic Limit

We will show that one can obtain exact expressions for the annealed entropy for a fixed margin for the case of a simple , provided we specialize to the so called the 'thermodynamic limit' of large input dimension N, and assume the scaling $\ell, N \to \infty$, keeping $\lambda = \frac{\ell}{N}$ fixed. To see that such a limit makes sense, we set $\lambda_{VC} = \frac{h}{N}$ and apply standard bounds on binomials in terms of binary entropies together with a Laplace approximation on the sum (approximated by an integral) to show that (1.4) yields

$$\lim_{N\to\infty} \frac{1}{N} H_{ann}(\lambda N) \tag{1.7}$$
$$\leq \begin{cases} \lambda \ln(2) & \text{for } \lambda \leq 2\lambda_{VC} \\ -\lambda\left[\frac{\lambda_{VC}}{\lambda}\ln(\frac{\lambda_{VC}}{\lambda}) + (1-\frac{\lambda_{VC}}{\lambda})\ln(1-\frac{\lambda_{VC}}{\lambda})\right] & \text{for } \lambda > 2\lambda_{VC}. \end{cases}$$

This bound becomes an equality for perceptrons without margin. It shows an interesting threshold phenomenon. If $\lambda > 2\lambda_{VC} \doteq \lambda_c$, then only an exponentially small fraction of all $2^{\lambda N}$ classifications can be realized. With probability approaching 1 in the , a *random* choice of output labels y_1^ℓ can not be realized by the classifier, when $\lambda > \lambda_c$. This result relates the *capacity* λ_c of the family of classifiers, via $\lambda_c \leq 2\lambda_{VC}$, to its VC–dimension. Implemented into (1.1), we also have with probability one, that deviations ε between generalization error and training error larger than $\sqrt{\frac{H_{ann}(2l)}{l}}$ will not occur. Hence, since $\lim_{N\to\infty} \frac{H_{ann}(2\lambda N)}{\lambda N}$ is bounded by a quantity of the order $\frac{\ln\lambda}{\lambda}$ for large λ, the possible deviations ε will become arbitrarily small as λ grows large. For the family of perceptrons through the origin, $h = N$, $\lambda_{VC} = 1$ and the capacity $\lambda_c = 2$.

spherical distribution

In the following, we will assume that the inputs are drawn independently from the spherical Gaussian distribution

$$f(\mathbf{x}) = (2\pi)^{-N/2} e^{-\frac{1}{2}||\mathbf{x}||^2}. \tag{1.8}$$

In this case, we get $||x||^2/N \to 1$ with probability one as $N \to \infty$. Hence, heuristically, we expect that (1.6) applies to this case with $R \approx \sqrt{N}$ so that the capacity λ_c should be bounded by a term which is of the order of $\frac{1}{\kappa^2}$.

1.4 An Expression for the Annealed Entropy

In this section, we present the basic ideas of a calculation for the annealed entropy for classification with a margin. We begin with the obvious fact, that the number of dichotomies can be rewritten in terms of $\theta(y_1^\ell, \mathbf{x}_1^\ell) \in \{0, 1\}$ as

$$\mathcal{N}(\mathbf{x}_1^\ell) = \sum_{y_1^\ell \in \{-1,1\}^\ell} \theta(y_1^\ell, \mathbf{x}_1^\ell), \tag{1.9}$$

where $\theta(y_1^n, \mathbf{x}_1^\ell) = 1$, if the labels are realizable with a margin κ and 0 else. In the next step, we have to average (1.9) over the distribution (1.8). By symmetry, all $2^{\lambda N}$ terms in the sum (1.9) give the same contribution and we can restrict ourselves to the case $y_i = 1$, for all $i = 1, \ldots, \ell$. We will denote the corresponding decision variable by $\theta(\mathbf{x}_1^\ell)$. Our basic idea for a construction of such a decision variable is based on the Kuhn Tucker conditions and the feasibility conditions on the primal/ dual quadratic optimization problem which is equivalent to (1.5). These conditions are expressed in terms of Lagrange multipliers α_i and read

$$\mathbf{w} = \sum_i \frac{y_i \alpha_i \, \mathbf{x}_i}{\sqrt{N}} \tag{1.10}$$

$$\frac{y_i}{\sqrt{N}}(\mathbf{w} \cdot \mathbf{x}_i) \geq 1 \tag{1.11}$$

$$\alpha_i \geq 0 \tag{1.12}$$

$$\sum \alpha_i \left(\frac{y_i}{\sqrt{N}}(\mathbf{w} \cdot \mathbf{x}_i) - 1 \right) = 0, \tag{1.13}$$

for $i = 1, \ldots, \ell$. We have rescaled all quantities by $\sqrt{N}$ such that for $N \to \infty$, the typical size of the α_i and the components of $\mathbf{w}$ remain of order 1. The last condition states that positive α_i (corresponding to *support vectors*) satisfy $\frac{y_i}{\sqrt{N}}(\mathbf{w} \cdot \mathbf{x}_i) = 1$. The resulting margin γ is given by

$$\gamma^2 = N/||\mathbf{w}||^2 = N/\sum_i \alpha_i. \tag{1.14}$$

It is useful to introduce auxiliary variables s_i by

$$\alpha_i = s_i \Theta(s_i). \tag{1.15}$$

Inserting the first equation into the second, also setting $y_i = 1$, the set (1.10) - (1.13) can be replaced by the single equation

$$s_i \Theta(-s_i) + \sum_j C_{ij}\alpha_j - 1 = 0 \tag{1.16}$$

with the matrix $C_{ij} = \frac{1}{N}(\mathbf{x}_i \cdot \mathbf{x}_j)$. $\Theta(x)$ is the unit step function which is 1 for $x \geq 0$ and 0 else. Introducing Dirac $\delta-$ distributions for the condition (1.16), we

decision variable

can write

$$\theta(\mathbf{x}_1^\ell) = \int_{-\infty}^{\infty} \prod_{i=1}^{\ell} ds_i \; \Theta(1/\kappa^2 - \frac{1}{N}\sum_i \alpha_i) \; \det(A) \times \prod_{i=1}^{\ell} \delta\left(s_i\Theta(-s_i) + \sum_j C_{ij}\alpha_j - 1\right). \tag{1.17}$$

Obviously, the integral is only different from zero, if the condition (1.16) is fulfilled with a margin above κ. The matrix A guarantees proper normalization and is given by $A_{ij} = C_{ij}\Theta(s_j)$ for $i \neq j$, and $A_{ii} = \Theta(-s_i) + C_{ii}\Theta(s_i)$. Since $C_{ii} \to 1$, with probability one as $N \to \infty$, we may also simply set $A_{ii} = C_{ii}$.

As a result, we have expressed the decision variable $\theta(\mathbf{x}_1^\ell)$ as a high dimensional integral, reminiscent of partition functions in statistical physics.

1.5 Evaluation in the Thermodynamic Limit

The basic strategies employed in the statistical mechanics approach consist in the following steps: Exchanging average and integrations, the average over inputs is performed first. Subsequently, the high dimensional integrations are decoupled by introducing auxiliary (low dimensional) integrations and are carried out. Finally, the low dimensional integrals are performed in the limit $N \to \infty$ by the method. By the fact that we are calculating the annealed average $\langle \mathcal{N}(\mathbf{x}_1^\ell) \rangle$ rather than the *quenched* average $\langle \ln(\mathcal{N}(\mathbf{x}_1^\ell)) \rangle$, more sophisticated methods (such as the 'replica trick') are not needed.

It is convenient to decompose the decision variable $\theta(\mathbf{x}_1^\ell)$ into contributions from

the different margins above κ

$$\theta(\mathbf{x}_1^\ell) = \int_0^{1/\kappa^2} Z(q, \mathbf{x}_1^\ell)\, dq \tag{1.18}$$

where now

$$Z(q, \mathbf{x}_1^\ell) = \int_{-\infty}^{\infty} \prod_{i=1}^{\ell} ds_i\, \delta(q - \frac{1}{N}\sum_i \alpha_i)\ \det(A) \times \prod_{i=1}^{\ell} \delta\left(s_i \Theta(-s_i) + \sum_j C_{ij}\alpha_j - 1\right). \tag{1.19}$$

To average over Z, we perform the expectation over the distribution of $\mathbf{x}_i$ first, before we carry out the integrations over the s_i. We have to average over a product of two terms, the determinant and the part with the $\delta-$ distributions. It is easy to see that $\det(A) = \det(B)$, where B is the submatrix of C, which contains all those elements C_{ij}, for which s_i and s_j are positive, i.e. for which both $\mathbf{x}_i$ and $\mathbf{x}_j$ are support vectors. The dimension of B is $\lambda_s N$, where

$$\lambda_s = \frac{1}{N}\sum_i \Theta(y_i). \tag{1.20}$$

averages

A proper and clean treatment of the determinant would require the introduction of Grassmann variables [2]. This more complicated route will be pursued somewhere else. In this article, we will resort to the following simpler heuristic assumption, which was frequently used for the statistical mechanics of similar problems. We argue that the fluctuations of the (to leading order of the exponent in N) can be neglected, and we can thus average both parts independently. In fact, it is more practical to average over $1/\det(B)$ (again neglecting fluctuations), because this has the representation

$$1/\det(B) = \int \prod_{i=1}^{\lambda_s N} dr_i \prod_{i=1}^{\lambda_s N} \delta\left(\sum_j B_{ij} r_j - 1\right), \tag{1.21}$$

which again is of a similar form as the $\delta-$ distribution part. As an argument, why the fluctuations of the determinant can be neglected, one can use the fact that

$$\lim_{N\to\infty} N^{-1} \langle \ln \det(B) \rangle = -\lim_{N\to\infty} N^{-1} \ln \langle (1/\det(B)) \rangle. \tag{1.22}$$

The term on the left can be calculated from the density of eigenvalues of B [4]. As a result for the average of (1.21) we get

$$\lim_{N\to\infty} -\frac{1}{N} \ln \langle (1/\det(B)) \rangle = -\lambda_s - (1-\lambda_s)\ln(1-\lambda_s). \tag{1.23}$$

For the $\delta-$ distribution part we get

$$\left\langle \prod_{i=1}^{\lambda N} \delta\left(s_i \Theta(-s_i) + \sum_j C_{ij}\alpha_j - 1\right) \right\rangle = \tag{1.24}$$

$$\frac{\exp\left[-\frac{1}{2}\sum_i \frac{(s_i\Theta(-s_i)-1)^2}{2q}\right]}{(2\pi q)^{\lambda N/2}(2\pi Q)^{N/2}}\sqrt{NqQ}\;\frac{\partial}{\partial q}\Omega_N(\sqrt{Nq})$$

where

$$q = \frac{1}{N}\sum_i \alpha_i \tag{1.25}$$

$$Q = \frac{1}{N}\sum_i \alpha_i^2 \tag{1.26}$$

and $\Omega_N(r)$ is the volume of an N dimensional sphere of radius r. These calculations are based on the decomposition

$$\prod_{i=1}^{\lambda N}\delta\left(s_i\Theta(-s_i)+\sum_j C_{ij}\alpha_j - 1\right) = \tag{1.27}$$
$$\int d\mathbf{w}\;\delta\left(\mathbf{w}-\sum_i \frac{\alpha_i\,\mathbf{x}_i}{\sqrt{N}}\right)\prod_{i=1}^{\lambda N}\delta\left(s_i\Theta(-s_i)+\frac{1}{\sqrt{N}}\mathbf{w}\cdot\mathbf{x}_i - 1\right).$$

The average of (1.27) can now be easily calculated from the joint density of the ℓ Gaussian random variables $\frac{1}{\sqrt{N}}\mathbf{w}\cdot\mathbf{x}_i$ and the N dimensional Gaussian vector $\sum_i \frac{\alpha_i\,\mathbf{x}_i}{\sqrt{N}}$.

Since both the conditions (1.25) and (1.26) and (1.20) are of a very simple additive type, the integrals over s_i can be decoupled by introducing their definitions within further δ distributions of the type

$$\delta\left(R-\frac{1}{N}\sum_j f(\alpha_j)\right) = \frac{N}{2\pi}\int d\hat{R}\;\exp\left(iN\hat{R}R - i\hat{R}\sum_j f(\alpha_j)\right). \tag{1.28}$$

decoupling

Note, that here i denotes the imaginary unit. Hence, by using the auxiliary variables $\hat{q}$, $\hat{Q}$, $\hat{\lambda}_s$ and corresponding integrals, the integrations over the s_i factorize and can be carried out. All remains to be done is to perform a 5 dimensional integral which is of the form

$$\langle Z(q,\mathbf{x}_1^{\lambda N})\rangle \propto \int dQ\;d\lambda_s\;d\hat{q}\;d\hat{Q}\;d\hat{\lambda}_s\;\exp[NG(\hat{\lambda}_s,\hat{q},\hat{Q},\lambda_s,Q,q)], \tag{1.29}$$

with

$$G(\hat{\lambda}_s,\hat{q},\hat{Q},\lambda_s,Q,q) =$$
$$-\lambda_s-(1-\lambda_s)\ln(1-\lambda_s)-\hat{\lambda}_s\lambda_s+\frac{1}{2}+\frac{1}{2}\hat{Q}Q+\frac{1}{2}\ln q+\ln\hat{q}-\hat{q}q-\ln(\hat{q}\sqrt{Q})$$
$$+\alpha\ln\left(\exp\left[-\frac{1}{2q}+\frac{\hat{q}^2}{2\hat{Q}}+\hat{\lambda}_s-\frac{1}{2}\ln q-\ln\hat{q}+\ln(\frac{\hat{q}}{\sqrt{\hat{Q}}})\right]\cdot\phi(-\frac{\hat{q}}{\sqrt{\hat{Q}}})+\phi(\frac{1}{\sqrt{q}})\right)$$

with $\phi(x) = \int_x^\infty \frac{dt}{\sqrt{2\pi}}\;e^{-t^2/2}$. To leading order in N (1.29) can be evaluated by the saddlepoint method (note, that the integrations over the 'hat' parameters are

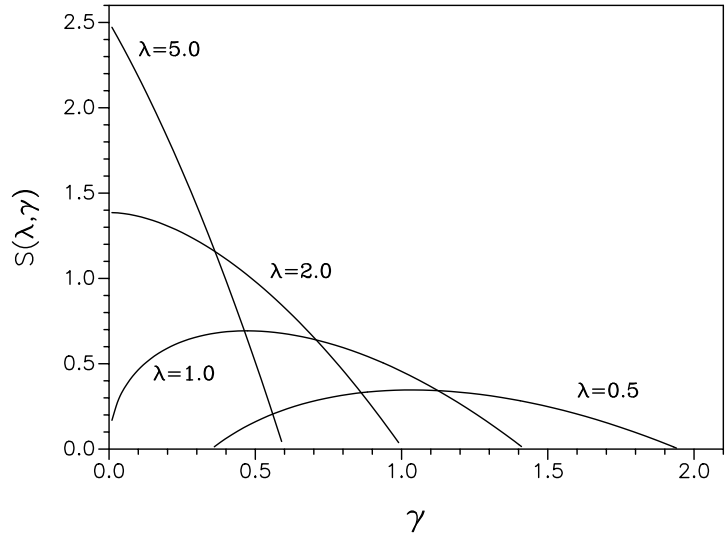

Figure 1.1 log of number of dichotomies for a maximal margin close to γ.

along the imaginary axis). Hence, $\lim_{N\to\infty} \frac{1}{N} \ln\langle Z(q, \mathbf{x}_1^{\lambda N})\rangle$ equals the function $G(\hat{\lambda}_s, \hat{q}, \hat{Q}, \lambda_s, Q, q)$ evaluated at the values of $\hat{\lambda}_s, \hat{q}, \hat{Q}, \lambda_s, Q$ for which the derivatives of G with respect to these parameters equals zero. One finds that these values satify $Q = 1/\hat{Q}$, $\hat{\lambda}_s = \ln(1-\lambda_s)$ and $q\hat{q} = 1-\lambda_s$. Treating $\hat{q}$ and $r = \frac{\hat{q}}{\sqrt{\hat{Q}}}$ as independent variables, and setting $q = 1/\gamma^2$, the annealed entropy is

$$\lim_{N\to\infty} \frac{1}{N} H_{ann}(\lambda N) = \sup_{\gamma>\kappa} S(\lambda, \gamma) \tag{1.30}$$

where S is given by the expression

$$\begin{aligned} S(\lambda,\gamma) &= \lambda \ln 2 + \lim_{N\to\infty} \frac{1}{N} \ln\langle Z(q = 1/\gamma^2, \mathbf{x}_1^{\lambda N})\rangle \\ &= \lambda \ln 2 + \ln(\gamma) + \min_r \left\{ -\ln r + \alpha \ln\left(e^{-\frac{\gamma^2}{2} + \frac{r^2}{2} + \ln r - \ln\gamma} \phi(-r) + \phi(\gamma) \right) \right\} \end{aligned} \tag{1.31}$$

S is $1/N\times$ log of the average number of dichotomies for which the maximal margin is in a small interval around the value γ

1.6 Results and Discussion

The minimization of (1.31) must be done numerically. The resulting function $S(\lambda, \gamma)$ is displayed in 1.1 for four values of λ. We have shown the positive part only, but the function extends to negative values as well. For $\lambda < 2$, the maximum of S is achieved for a margin $\gamma > 0$ which can be found by differentiating (1.31) with

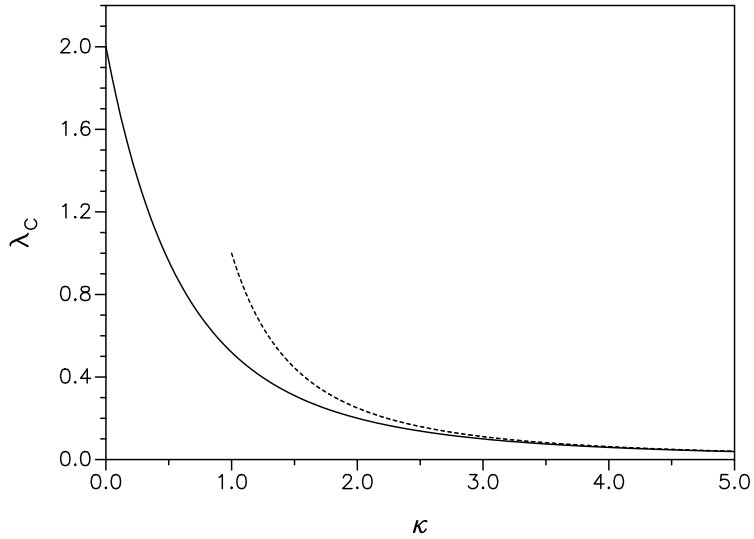

Figure 1.2 Capacity λ_c as a function of the margin κ. The dashed line is $1/\kappa^2$.

respect to γ. This results in the equation

$$\lambda \int_{-\gamma}^{\infty} \frac{dt}{\sqrt{2\pi}}\, e^{-t^2/2}(t+\gamma)^2 \;= 1. \tag{1.32}$$

Solving for γ, we also find that for this λ, $S(\lambda,\gamma) = \alpha \ln 2$. This result means that for $\lambda < 2$, almost all $2^{\lambda N}$ dichotomies will be realized with a margin γ given by (1.32). On the other hand, fixing the margin γ, the value of λ given by (1.32) yields the corresponding λ_c. Relation (1.32) (Fig.(1.2) is a well known result in the statistical mechanics of neural networks, which was first derived by Elizabeth Gardner [3] using a rather different approach based on the method of replicas. As can be seen, the result is in agreement with the suggested scaling $\lambda_c \sim 1/\gamma^2$ for large margins γ. For $\lambda > 2$, the maximum of $S(\lambda,\gamma)$ is shifted to $\gamma = 0$ and we obtain

capacity

$$S(\lambda, \gamma = 0) = \lambda \ln \lambda - (\lambda - 1)\ln(\lambda - 1), \tag{1.33}$$

which gives the correct result (1.7) for the VC entropy with zero margin. In Fig.1.3, we have displayed the annealed VC entropy (1.30) as a function of λ/λ_c for three values of κ. While for small κ, the decrease of the annealed entropy (divided by $\lambda ln2$) is similar to the bound (1.7), the decrease becomes faster with increasing margin κ. In any case, for λ large enough, the annealed entropy also achieves negative values in contrast to the bound (1.7). This should not be too surprising, because it simply means that for too many inputs, there is a nonzero probability, that *none* of the 2^ℓ classifications can be realized with a margin greater than a given κ by hyperplanes through the origin. Our result shows that for sufficiently large margins, the VC complexity of the set of perceptrons is drastically reduced, even stronger than predicted by general bounds. Although it is not trivial to express the generalization error in terms of the annealed entropy (because the data dependent margin is also a

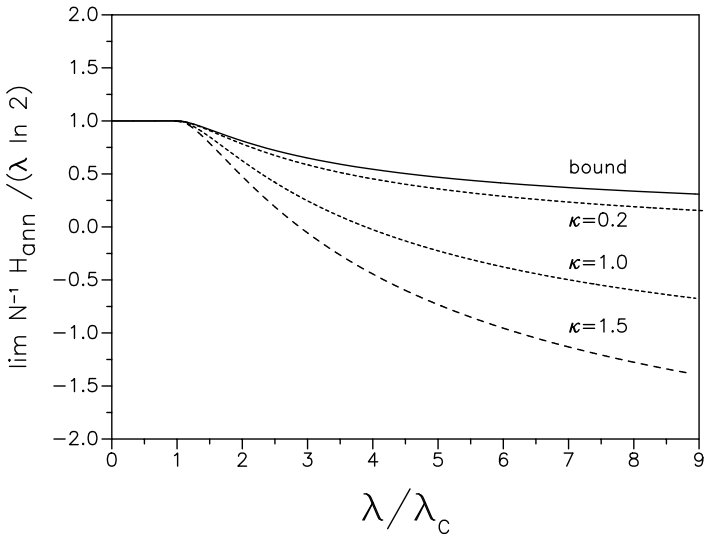

Figure 1.3 Annealed entropy for three different values of margin κ as a function of the number of inputs/capacity. The upper line gives the bound (1.7).

random variable) we expect that our results give a further illustration why margin classifiers and supportvector machines generalize so well, when the achieved margin κ is large. Although the results have been derived for a , one can expect that in the limit $N \to \infty$, by the central limit theorem, any other product distribution for the components of the input vector $\mathbf{x}$ with zero mean and unit variance will lead to the same result. It would be interesting to see if one can proove that our result (by the symmetry of the spherical distribution) may actually give an upper bound on the annealed entropy for any distribution of inputs $\mathbf{x}$ with $||\mathbf{x}||^2 \leq N$. Such a result would be helpful for obtaining sharper bounds on the VC entropy.

Acknowledgements

I would like to thank Peter Kuhlmann and Andreas Mietzner for their pleasant collaboration on [6], on which the present calculation is based. I am also grateful to the referees for their helpful comments.

References

1. N. Alon, S. Ben-David, N. Cesa-Bianchi, and D. Haussler. Scale–sensitive Dimensions, Uniform Convergence, and Learnability. *Journal of the ACM*, 44(4):615–631, 1997.
2. K. Efetov. *Supersymmetry in Disorder and Chaos*. Cambridge University Press, Cambridge, 1997.
3. E. Gardner. The space of interactions in neural networks. *Journal of Physics A*, 21:257–70, 1988.

4. M. Opper. Learning in neural networks: Solvable dynamics. *Europhysics Letters*, 8(4):389–392, 1989.

5. M. Opper and W. Kinzel. Physics of generalization. In E. Domany J.L. van Hemmen and K. Schulten, editors, *Physics of Neural Networks III*. Springer Verlag, New York, 1996.

6. M. Opper, P. Kuhlmann, and A. Mietzner. Convexity, internal representations and the statistical mechanics of neural networks. *Europhysics Letters*, 37(1):31–36, 1997.

7. H. S. Seung, H. Sompolinsky, and N. Tishby. Statistical mechanics of learning from examples. *Physical Review A*, 45(8):6056–6091, 1992.

8. J. Shawe-Taylor, P. Bartlett, R. Williamson, and M. Anthony. A framework for structural risk minimization. In *COLT*, 1996.

9. V. Vapnik. *Estimation of Dependences Based on Empirical Data [in Russian]*. Nauka, Moscow, 1979. (English translation: Springer Verlag, New York, 1982).

10. V. Vapnik. *The Nature of Statistical Learning Theory*. Springer Verlag, New York, 1995.

11. V. Vapnik and A. Chervonenkis. Uniform convergence of frequencies of occurence of events to their probabilities. *Dokl. Akad. Nauk SSSR*, 181:915 – 918, 1968.

12. T. L. H. Watkin, A. Rau, and M. Biehl. The statistical mechanics of learning a rule. *Reviews of Modern Physics*, 65:499–556, 1993.