Convexity, Internal Representations and the Statistical Mechanics of Neural Networks

Manfred Opper

Peter Kuhlmann

and

Andreas Mietzner

Institut für Theoretische Physik III, Universität Würzburg

Am Hubland, D-97074 Würzburg

Abstract

We present an approach to the statistical mechanics of feedforward neural networks which is based on counting realizable internal representations by utilizing convexity properties of the weight space. For a toy model, our method yields storage capacities based on an annealed approximation, which are in close agreement to one step replica symmetry breaking results obtained from a standard approach. For a single layer perceptron, a combinatorial result for the number of realizable output combinations is recovered and generalized to fixed stabilities.

PACS numbers: 87.10, 05.90

The statistical mechanics approach to learning in neural networks is largely based on a famous method developed by E. Gardner [1]. Averaging over all networks, which are compatible with training data, one is able to estimate the properties of a typically trained network. The basic quantity to be calculated in this approach is the volume of the subspace of network couplings which correctly implement the data. In the language of statistical mechanics, this subspace is called *phase space*. The calculation is usually performed by the so called replica method. In its simplest version, the replica symmetric assumption (RS) is made, which seems to be justified for many interesting situations, like the problem of learning a realizable rule from examples (for a review, see e.g. [2, 3, 4]). In the latter case, the logarithm of the phase space volume has a natural meaning. After the learning network has perceived a finite amount of data, this quantity can be understood as a measure of the uncertainty about the unknown teacher network [4, 5].

The situation was less satisfactory for the problem of calculating the capacity of a network. This is defined as the number of random examples for which the typical phase space volume vanishes. While the assumption of replica symmetry yields correct results for the capacity of single layer perceptrons, it severely fails for many simple multilayer machines [6, 7, 8]. Here the RS - capacities violate exact bounds [9] and the more complicated machinery of replica symmetry breaking (RSB) has to be applied.

Hence, it is important to think about alternative analytical approaches to this problem. Recent studies [10, 11, 12] gave new criteria for calculating capacities which are based on an analysis of the number of ways a set of training examples can be stored by using different *internal representations* in the hidden layer. This method reveals new and interesting multifractal properties of the space of couplings, but one has to pay the price of having a more complicated analytical approach with a double set of replicas.

In this letter, we propose an alternative and direct approach to the calculation of the total number of internal representations which may already give good results in a simple annealed approximation, i.e. without replicas. As the new idea we utilize the fact that the phase space of network couplings for fixed internal representations is often a convex set for which many properties can be calculated exactly without using the replica method.

Sofar, we have applied this idea to a toy model which is a variant of the model studied by [10]. In their work, αN input patterns $\boldsymbol{\xi}^\mu$ (being N component vectors) with their outputs S^μ are correctly stored by couplings $\mathbf{J}$, when the internal fields $\Delta_\mu = N^{-\frac{1}{2}} S^\mu \mathbf{J} \cdot \boldsymbol{\xi}^\mu$ belong to one of two specific intervals. As a simplified model, we choose in this paper the symmetric conditions $\Delta_\mu \leq -1$ or $\Delta_\mu \geq 1$ for $\mu = 1, \ldots, \alpha N$, to be satisfied by the vector of couplings $\mathbf{J}$. We constrain the length of $\mathbf{J}$ by $N^{-1}\mathbf{J}^2 \leq \frac{1}{\gamma^2}$, where γ is an adjustable parameter. As in [10], the storage task can be realized in many different ways because the phase space splits into $2^{\alpha N}$ disjoint cells with volumes $V(\{\sigma\})$ (many of them empty!) where the label $\{\sigma\} = (\sigma_1, \ldots \sigma_{\alpha N})$ of the *internal representation* is defined by the signs σ_μ of Δ_μ. There is no direct interpretation of this model as the supervised learning problem for a concrete network, because the storage conditions turn out to be independent of the outputs S^μ. They will be set to $S^\mu = 1$ in the following. However, since each phase space cell equals the weight space of a *single layer perceptron* with outputs $\{\sigma\}$, the model is equivalent to a problem of *unsupervised* learning for the perceptron. Only those outputs $\{\sigma\}$ lead to nonempty cells for which the clouds of patterns with different outputs are separated by a margin (the *stability*) $\min_\mu \sigma_\mu \Delta_\mu / ||\mathbf{J}|| \geq \gamma$. The capacity $\alpha_c(\gamma)$ gives the maximum number of patterns which can be separated by a stability γ. See [13] for a replica treatment of this problem.

In the approach of [10], it is possible to estimate the entire spectrum of cell sizes. To be precise, the number $\Omega(w)$ of cells with size $w = N^{-1} \ln V(\{\sigma\})$ is found from a free energy like function

$$g(r) = -\frac{1}{Nr} \langle\langle \ln \sum_{\{\sigma\}} V^r(\{\sigma\}) \rangle\rangle \tag{1}$$

by the relations $w_r = \frac{\partial(rg(r))}{\partial r}$ and $N^{-1} \ln \Omega(w_r) = -\frac{\partial g(r)}{\partial(1/r)}$. The capacity is reached when the log of the total number of cells $\ln \mathcal{M} = \lim_{r\to 0} \ln \Omega(r)$ vanishes.

Our new approach is based on the representation $\mathcal{M} = \sum_{\{\sigma\}} Z(\{\sigma\})$ where the decision variable $Z(\{\sigma\})$ counts a 1 if (for fixed $\{\sigma\}$) the linear system of inequalities

$$\sigma_\mu \Delta_\mu = N^{-\frac{1}{2}} \sigma^\mu \mathbf{J} \cdot \boldsymbol{\xi}^\mu \geq 1; \qquad \mu = 1, \ldots, \alpha N, \tag{2}$$

has a solution and 0 else. Sofar, our new method is restricted to the total number of cells, because it avoids the introduction of phase space volumes at all (together with the replica index r).

In order to restrict the number of solution vectors of (2) to maximally *one*, we additionally require that $\mathbf{J}$ has *minimal norm* $||\mathbf{J}||$. This $\mathbf{J}$ is the solution of a convex minimization problem with linear constraints, to which the so called *Kuhn–Tucker theory* (KT) (see e.g. [14]) applies. The KT theory has been used previously [15, 16] in the statistical mechanics of neural networks in combination with the Cavity approach [17]. Our present use of the KT theory is different. Since we expect that our approach might have wider applications in the statistical mechanics of disordered systems, we will briefly review the basic ideas in a somewhat broader context.

The aim is to construct an indicator variable $Z \in \{0, 1\}$ which tells us if there is a solution to a minimization problem for a strictly convex function $f(\mathbf{J})$ when $\mathbf{J}$ is constrained by a set of inequality conditions $g_\mu(\mathbf{J}) \leq 0$ for $\mu = 1, \ldots, \alpha N$. Whenever the functions g_μ are *convex*, there can only be a *single* vector $\mathbf{J}$ which minimizes f. Necessary and sufficient conditions for a minimum are known as the so called *Kuhn–Tucker* conditions. They are based on the Lagrange function $L = f(\mathbf{J}) + \sum_{\mu=1}^{\alpha N} x_\mu g_\mu(\mathbf{J})$ with Lagrange multipliers x_μ. The conditions for a minimum of f are found to be

$$\begin{aligned} \frac{\partial L}{\partial J_j} = 0; \qquad & \frac{\partial L}{\partial x_\mu} = g_\mu(\mathbf{J}) \leq 0; \\ x_\mu \geq 0; \qquad & \sum_{\mu=1}^{\alpha N} x_\mu g_\mu(\mathbf{J}) = 0, \end{aligned} \tag{3}$$

for $j = 1, \ldots, N$ and $\mu = 1, \ldots, \alpha N$. The last condition states that positive x_μ imply $g_\mu = 0$. By setting $x_\mu = y_\mu \Theta(y_\mu)$, the last three conditions in (3) can be replaced by the single equation $y_\mu \Theta(-y_\mu) = g_\mu(\mathbf{J})$. Hence, by using these conditions within a δ function, an indicator variable for a solution can be written as

$$Z = \int d[\{y\},\mathbf{J}] \prod_\mu \delta(y_\mu \Theta(-y_\mu) - g_\mu(\mathbf{J})) \cdot D(\mathbf{J},\{y\}) \cdot \prod_j \delta(\frac{\partial L}{\partial J_j}), \tag{4}$$

where $D(\mathbf{J},\{y\})$ is a normalizing determinant.

To apply the KT method to our set of linear inequalities (2), we set $g_\mu(\mathbf{J}) = 1 - \frac{\sigma_\mu}{\sqrt{N}}\mathbf{J}\boldsymbol{\xi}^\mu$ and $f(\mathbf{J}) = N^{-1}\mathbf{J}^2$. We can now count the number of realizable internal representation as a function of $q = N^{-1}\mathbf{J}^2$ and obtain $\mathcal{M} = \int_0^{1/\gamma^2} dq\, \mathcal{N}(q)$ where

$$\begin{aligned}\mathcal{N}(q) = \sum_{\{\sigma\}} \int d[\{y\},\mathbf{J}] \delta(\mathbf{J}^2 - Nq) \prod_\mu \delta(y_\mu \Theta(-y_\mu) + \frac{\sigma^\mu}{\sqrt{N}} \sum_j J_j \xi_j^\mu - 1) \times \\ \prod_j \delta(J_j - \frac{\sigma^\mu}{\sqrt{N}} \sum_\mu x_\mu \xi_j^\mu) \det(B(\{y_\mu\})).\end{aligned} \tag{5}$$

The matrix B is defined by $B_{\mu\nu} = \frac{\sigma_\mu \sigma_\nu}{N} \sum_j \xi_j^\mu \xi_j^\nu \Theta(y_\nu)$ for $\mu \neq \nu$, and $B_{\mu\mu} = 1$. When the inputs $\boldsymbol{\xi}^\mu$ are drawn at random, the calculation of the *typical* value of $\mathcal{N}(q)$ would require a quenched average of the logarithm of $\mathcal{N}(q)$. In order to avoid replicas, we will restrict ourselves to the simpler annealed average, which gives an upper bound to the typical value. Averaging (5) over a spherical density of inputs yields

$$\begin{aligned}\langle\langle \mathcal{N}(q) \rangle\rangle = 2^{\alpha N} \int d[\{y,\hat{y}\},\mathbf{J},\hat{\mathbf{J}}] \exp\left[-\frac{q}{2} \sum_\mu \hat{y}_\mu^2 - \frac{Q}{2} \sum_j \hat{J}_j^2 - N^{-1} \sum_{\mu,j} x_\mu \hat{y}_\mu J_j \hat{J}_j \right] \times \\ \exp\left[i \sum_\mu \hat{y}_\mu (y_\mu \Theta(-y_\mu) - 1) - i \sum_j J_j \hat{J}_j \right] \langle\langle \det(B) \rangle\rangle,\end{aligned} \tag{6}$$

where $q = \frac{1}{N}\sum_j J_j^2$ and $Q = \frac{1}{N}\sum_\mu x_\mu^2 = \frac{1}{N}\sum_\mu y_\mu^2 \Theta(y_\mu)$. The determinant is selfaveraging in the thermodynamic limit, with $\frac{1}{N}\ln\det(B) = -\alpha_{eff} - (1-\alpha_{eff})\ln(1-\alpha_{eff})$, where $\alpha_{eff} N = \alpha \sum_\mu \langle\langle \Theta(y_\mu) \rangle\rangle$.

By introducing orderparameters $\hat{q}$, $\hat{Q}$ and λ and defining $\kappa = 1/\sqrt{q}$, the annealed entropy $S(\kappa,\alpha) = \frac{1}{N}\ln\langle\langle\mathcal{N}(q=\frac{1}{\kappa^2})\rangle\rangle$ is obtained in the thermodynamic limit as an extremum of the function

$$\begin{aligned} G(\lambda,\alpha_{eff},\hat{q},\hat{Q},Q,q) = \alpha\ln 2 - \alpha_{eff} - (1-\alpha_{eff})\ln(1-\alpha_{eff}) - \lambda\alpha_{eff} + \\ \frac{1}{2} + \frac{1}{2}\hat{Q}Q + \frac{1}{2}\ln q + \ln\hat{q} - \hat{q}q - \ln(\hat{q}\sqrt{Q}) + \\ \alpha\ln\left(\exp\left[-\frac{1}{2q} + \frac{\hat{q}^2}{2\hat{Q}} + \lambda - \frac{1}{2}\ln q - \ln\hat{q} + \ln(\frac{\hat{q}}{\sqrt{\hat{Q}}})\right]\cdot H(-\frac{\hat{q}}{\sqrt{\hat{Q}}}) + H(\frac{1}{\sqrt{q}})\right) \end{aligned} \qquad (7)$$

with respect to variations of $\lambda, \alpha_{eff}, \hat{q}, \hat{Q}, Q$. Treating $\hat{q}$ and $r = \frac{\hat{q}}{\sqrt{\hat{Q}}}$ as independent variables, one finds $Q = 1/\hat{Q}$, $\lambda = \ln(1-\alpha_{eff})$ and $q\hat{q} = 1-\alpha_{eff}$. Finally, the annealed entropy simplifies to

$$S(\kappa,\alpha) = \alpha\ln 2 + \ln(\kappa) + \mathrm{Extr}_r\left\{-\ln r + \alpha\ln\left(e^{-\frac{\kappa^2}{2}+\frac{r^2}{2}+\ln r-\ln\kappa}H(-r) + H(\kappa)\right)\right\}. \qquad (8)$$

The function $S(\kappa,\alpha)$ is displayed in Figure 1 for four values of α. $S(\kappa,\alpha)$ is positive only below a critical value κ_{max}, which is an *exact* lower bound to the highest stability which can be achieved. Conversely, by solving $S(\kappa_{max}=\gamma,\alpha) = 0$ with repect to α, we get an exact upper bound to the capacity α_c of our model which is shown in Figure 2. Remarkably, it differs only by a fraction of less than 10^{-2} from the single step RSB result derived from Gardner's method in [13].

Within the interpretation of our model as a single layer perceptron with free outputs $\{\sigma\}$, the parameter κ equals the *optimal stability* $\kappa = \max_{\mathbf{J}}\min_\mu \sigma_\mu\Delta_\mu/||\mathbf{J}||$. The value κ_{typ} for which S is maximized, yields an estimate of the *typical* optimal stability, i.e. the one which is achieved in the thermodynamic limit by almost all output configurations. After some algebra the relation $\alpha \int_{-\kappa_{typ}}^{\infty} Dt\ (t+\kappa)^2\ = 1$ is found for $\alpha \leq 2$. This is E. Gardner's result for the typical optimal stability of a perceptron [1], which we now obtain without the replica method. The corresponding entropy is $S(\kappa_{opt}, \alpha) = \alpha \ln 2$, i.e. for $\alpha \leq 2$ *almost all* $2^{\alpha N}$ output configurations of a perceptron are realizable [1, 18]. For $\alpha > 2$, the maximum of the entropy is shifted to $\kappa = 0$ and $S(\kappa = 0, \alpha) = \alpha \ln \alpha - (\alpha - 1) \ln(\alpha - 1)$, which is the correct result for the total number of realizable output configurations, known from the combinatorial approach of Thomas Cover [18]. There it was also shown that this number is nonfluctuating, i.e. it does not depend on the position of the input vectors, explaining why our annealed results are correct at κ_{typ}.

Qualitatively, the curves in Fig. 1 resemble those obtained in [10] and [19] for the number of phase space cells as function of the logarithm of their volume $\frac{1}{N} \ln V(\{\sigma\})$. Qualitatively, the larger κ, the more the solution vector $\mathbf{J}$ of the minimization problem can be perturbed by still keeping all patterns stored in the perceptron. Hence, a monotonic relation between κ and volume can be expected, but has sofar not been established. It is interesting to note from Fig.1 that for $\alpha < 1$, there is always a smallest nonzero stability.

It is tempting to speculate about whether the annealed approach to our model may be exact over a broader region of κ. A preliminary replica study for the calculation of $\langle\langle \mathcal{M}^n \rangle\rangle$ shows that the annealed result gives in fact a solution to the saddlepoint equations. However, sofar the stability of this solution has to be checked in future work.

It would be interesting to apply our method to 'real' multilayer networks. It can be shown [20] that for the case of a committee machine the annealed approach would already yield an exact bound to the capacity which was previously obtained by Mitchison and Durbin [9]. We conjecture that from a replica symmetric treatment of our method even better results can be obtained.

Acknowledgement M. Opper is grateful to S. Diederich for suggesting a simplification of the KT conditions.

References

[1] E.Gardner; J.Phys. A 21, 257 (1988).

[2] H. Seung, H. Sompolinsky, and N. Tishby; Physical Review A 45, 6056 (1992).

[3] T. L. H. Watkin, A. Rau and M. Biehl; Rev. Mod. Phys. 65, 499 (1993).

[4] M. Opper and W. Kinzel; *Statistical Mechanics of Generalization*, to appear in: *Physics of Neural Networks*, ed. by J. L. van Hemmen, E. Domany and K. Schulten, published by Springer Verlag.

[5] Manfred Opper; Phys. Rev. E 51, 3613 (1995).

[6] E. Barkai, D. Hansel and H. Sompolinsky; Phys. Rev. A. 45, 4146 (1992).

[7] E. Barkai, D. Hansel and I. Kanter; Phys. Rev. Lett. 65, 2312 (1990)

[8] A. Engel, H. M. Koehler, F. Tschepke, H. Vollmayr and A. Zippelius; Phys. Rev. A 45, 7590 (1992).

[9] G. J. Mitchison and R. M. Durbin; Biol. Cybern. 60, 345 (1989).

[10] R. Monasson, D. O'Kane; Europhys. Lett. 27, 85 (1994).

[11] R. Monasson and R. Zecchina; Phys. Rev. Lett 75, 2432 (1995).

[12] R. Monasson and R. Zecchina; preprint *cond-mat/9601122*.

[13] A. Mietzner, M. Opper and W. Kinzel; J. Phys. A 28, 2785 (1995).

[14] L. Collatz, W. Wetterling. *Optimization Problems*, Springer Verlag, 1975.

[15] W. Kinzel and M. Opper: *Dynamics of Learning*; in: *Physics of Neural Networks*, ed. by J. L. van Hemmen, E. Domany and K. Schulten; published by Springer Verlag, p. 149 (1991).

[16] F. Gerl and U. Krey; J. Phys. A 27, 7353 (1994).

[17] M.Mezard and G.Parisi and M.A.Virasoro. *Spin Glass Theory and Beyond*, Lecture Notes in Physics, 9, World Scientific, 1987.

[18] T.M.Cover; IEEE Trans.El.Comp. 14, 326 (1965).

[19] A. Engel and M. Weigt; *Multifractal Analysis of the Coupling Space of Feed–Forward Neural Networks*, Universität Magdeburg preprint (1995).

[20] Ole Winther, private communication.

Figure Captions:

Fig1: Entropy $S(\kappa, \alpha)$ as a function of the stability κ for $\alpha = 0.5, 1.0, 2.0, 5.0$

Fig2: Storage capacity α_c as a function of γ.