Generalization ability of optimal cluster separation networks

1994 J. Phys. A: Math. Gen. 27 2325

(http://iopscience.iop.org/0305-4470/27/7/014)

J. Phys. A: Math. Gen. 27 (1994) 2325–2333. Printed in the UK

Generalization ability of optimal cluster separation networks

A Wendemuth

Department of Physics, Theoretical Physics, Oxford University, 1 Keble Road, Oxford OX1 3NP, UK

Received 26 August 1993, in final form 25 October 1993

Abstract. Optimal separation of two clusters of normalized vectors can be performed in a neural network with adjustable threshold and weights, which is trained to maximum stability. Generalization from arbitrarily selected training clusters to a given bipartitioning of input space is studied. The network's threshold becomes a global optimization (and order) parameter. This causes the generalization ability to increase rapidly with the distance of the cluster separation plane from the origin. Separation is shown to be *stochastic* for small and *deterministic* for large training cluster sizes.

1. Introduction

In any problem of inferring parameters of an information processing device from a set of known example patterns, the key issue is good generalization. That is, a high probability of identifying an unknown pattern correctly. This paper adresses the issue of generalization for the most general of linear threshold classifiers, which include weighted sums of input and a global activation threshold. The examples used for training are required to be linearly separable, such that the network can actually perform the inference with zero training error.

Recently, generalization has been studied [1] in a less general class of neural networks, namely those with zero activation thresholds. These networks were trained with various algorithms (Hebb, pseudoinverse, optimal stability), where the optimally stable network generalized best.

However, the separation of two classes of network output may require a non-zero threshold. For example, this is the case already for such simple tasks as the two-dimensional AND or OR functions in the $(-1,1)$-representation.

Optimal separation of the two output clusters cannot be performed by existing algorithms which simply treat the network threshold as an additional input dimension. When doing so, the algorithm will find *one* solution to the separation problem. However, the separation achieved will not be optimal, since the threshold dimension just becomes another intensive (local) variable. Optimal separation can be achieved only through correct treatment of the threshold as an extensive (global) variable, which is well reflected in the course of this paper.

Recently, suitable algorithms have been proposed [2,3] which treat the threshold correctly, and which meet theoretical stability predictions [4] for large neural networks. Here, the generalization ability of those networks is analysed.

0305-4470/94/072325+09$19.50

2. Problem and solution

The basic problem can be formulated in terms of neural networks as well as in geometric terms. This will show exploitable analogies and allow an interpretation of the results beyond the limits of neural networks.

In neural network terms, a *teacher network* with N-dimensional weight vector B and threshold U is given. A *student network* is going to select randomly αN *examples* (or *questions*) ξ^μ with components $\xi_j^\mu = \pm 1$, $j = 1, \ldots, N$. The examples belong to one of two classes labelled $\tau_\mu = \pm 1$ which the teacher answers according to

$$\tau_\mu = \operatorname{sign}\left(\frac{1}{\sqrt{N}} B \cdot \xi_\mu - U\right). \tag{1}$$

The student *learns* these examples by selecting his network weights J and threshold T such that the examples are stored with maximum stability $\Delta = \Delta_{\text{opt}}$, i.e.

$$\Delta_{\text{opt}} = \min_\mu \left(\frac{J \cdot \tau_\mu \xi_\mu - \sqrt{N} T \tau_\mu}{|J|}\right) = \max_{\{J', T'\}} \min_\mu \left(\frac{J' \cdot \tau_\mu \xi_\mu - \sqrt{N} T' \tau_\mu}{|J'|}\right) \tag{2}$$

which satisfies a correct output ($\xi_\mu^o = \tau_\mu$, $\mu = 1, \ldots, p$):

$$\xi_\mu^o = \operatorname{sign}\left(\frac{1}{\sqrt{N}} J \cdot \xi_\mu - T\right). \tag{3}$$

This can be described in geometric terms as well. The two sets of examples belonging to the two classes of output form two *clusters*. The task then is to separate these clusters such that the gap between their convex hulls becomes maximal. The gap size is exactly $2\Delta_{\text{opt}}$, the normal direction from one convex hull to the other is J, and the centre of the gap is at distance T from the origin. The hyperplane which is the centre of the gap is then given by all points ξ satisfying

$$\frac{1}{\sqrt{N}} J \cdot \xi - T = 0. \tag{4}$$

Generalization may also be described in both terms. In a neural network, the generalization ability G is defined as the probability that, after learning, the student's answer S_μ^0 to *any* question S_μ is given correctly ($S_\mu^0 = \tau_\mu$), i.e. in accordance with the teacher. In geometric terms, it is the probability that, after selecting the hyperplane (4), the class τ_μ given by the bipartitioning of input space (1) is identified correctly by equation (3) for any point S_μ. In both formulations, the generalization ability is given by

$$G = \left\langle \Theta\left(\tau_\mu \xi_\mu^0\right)\right\rangle_{\{S_j^\mu = \pm 1\}} = \left\langle \Theta\left(\left(\frac{1}{\sqrt{N}} B \cdot S_\mu - U\right)\left(\frac{1}{\sqrt{N}} J \cdot S_\mu - T\right)\right)\right\rangle_{\{S_j^\mu = \pm 1\}}. \tag{5}$$

This average will be performed now. Since only the *directions* of J and B are important, one can scale $|J|^2 = |B|^2 = N$. Using an integral representation for the argument of the Θ-function gives for large N:

$$\int_{-\infty}^{+\infty} \mathrm{d}x \int_{-\infty}^{+\infty} \mathrm{d}y\, \Theta((x-U)(y-T))$$

$$\times \int_{-\infty}^{+\infty} \mathrm{d}w \int_{-\infty}^{+\infty} \mathrm{d}q\, \mathrm{e}^{(-\mathrm{i}wx - \mathrm{i}qy)} \prod_{j=1}^{N} \left\langle \exp \frac{\mathrm{i} S_j^\mu}{\sqrt{N}} (w B_j + q J_j) \right\rangle_{S_j^\mu = \pm 1}$$

$$= \int_{-\infty}^{+\infty} dx \int_{-\infty}^{+\infty} dy\, \Theta((x-U)(y-T))$$
$$\times \int_{-\infty}^{+\infty} dw \int_{-\infty}^{+\infty} dq\, \exp\left\{-iwx - iqy - \frac{i}{2N}\sum_{j=1}^{N}\left(wB_j + qJ_j\right)^2\right\}$$
$$= G$$
$$= \iint_{(A_+ + A_-)} Dx\, Dy \tag{6}$$

with the Gaussian measure $Dz = 1/\sqrt{2\pi}\, \exp(-z^2/2)$, the integration areas

$$A_\pm = \{x, y | x = \pm U\sqrt{1-r^2} \ldots \infty;\ y = \pm T - rx/\sqrt{1-r^2} \ldots \infty\}$$

and $r = \cos(J, B)$.

Thus $G(U, \alpha)$ can be evaluated once r and T are known. These parameters can be derived from a calculation [4] of the volume V available for a solution (3) in the phase space (J, T) with stability $\Delta > 0$:

$$V = \int_{-\infty}^{+\infty} dT \prod_{j=1}^{N} dJ_j \prod_{\mu} \Theta\left[-\Delta + \left(\frac{1}{\sqrt{N}} J \cdot \xi_\mu - T\right) \operatorname{sign}\left(\frac{1}{\sqrt{N}} B \cdot \xi_\mu - U\right)\right]$$
$$\times \delta\left(|J|^2 - N\right). \tag{7}$$

For an extensive number of learning patterns, $\ln(V)$ becomes self-averaging. Since the space of weights is connected and convex [3], $\ln(V)$ can be computed by a replica-symmetric calculation, where the saddle point is obtained with respect to replica-symmetric order parameters

$$q = \overline{\frac{1}{N}\sum_{j=1}^{N}\sum_{a \neq b} J_j^a J_j^b} \qquad r = \overline{\frac{1}{N}\sum_{j=1}^{N} J_j^a B_j} \qquad T = \overline{T_a} \tag{8}$$

the averages being taken with respect to replicas a, b. Note that T becomes a global parameter optimized at the saddle point. This is in contrast to simply treating the threshold as an additional ('augmented') input dimension. Maximum stability now corresponds to $V \to 0$. After standard integrations (e.g. see [1] for techniques) one obtains the conditions for G as

$$-\frac{r}{\alpha} = \iint_A Dz\, Dw \left(\Delta + (T - wr)\operatorname{sign}(w-U) - z\sqrt{1-r^2}\right)$$
$$\times \left[-w \operatorname{sign}(w-U) + \frac{zr}{\sqrt{1-r^2}}\right] \tag{9}$$

$$\frac{1-r^2}{\alpha} = \iint_A Dz\, Dw \left(\Delta + (T - wr)\operatorname{sign}(w-U) - z\sqrt{1-r^2}\right)^2 \tag{10}$$

$$0 = \iint_A Dz\, Dw \left(\Delta + (T - wr)\operatorname{sign}(w-U) - z\sqrt{1-r^2}\right) \operatorname{sign}(w-U) \tag{11}$$

with the integration area

$$A = \{z, w | \Delta + (T - wr)\operatorname{sign}(w-U) - z\sqrt{1-r^2} > 0\}.$$

Equations (6) and (9)–(11) can be solved numerically for r, T and Δ as functions of α and U. The generalization ability resulting from these parameters is shown in figure 1.

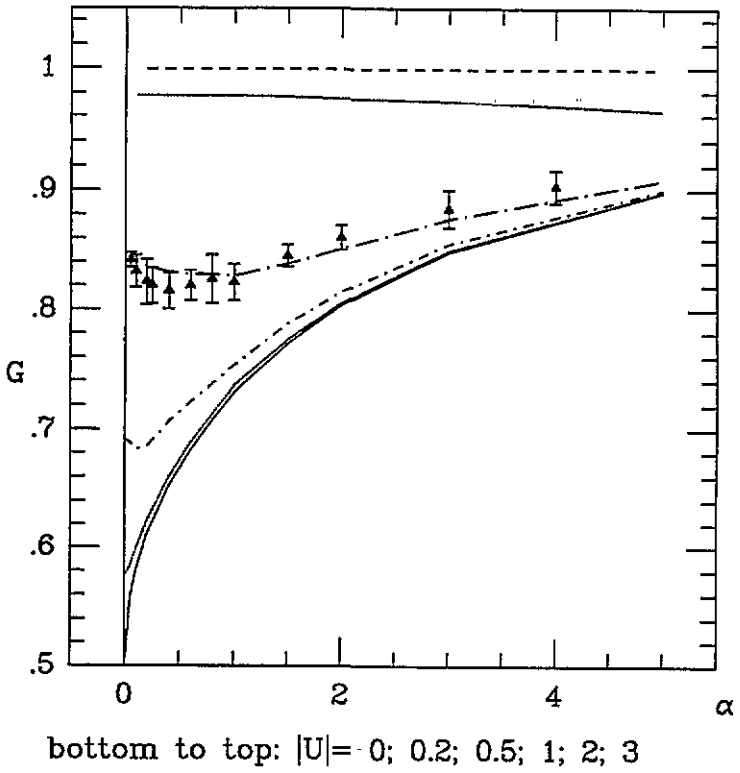

Figure 1. Generalization ability as a function of capacity for various teacher thresholds U. Generalization converges to 1 for $U \to \infty$. For $U = 1$, simulation results are also given for a network with 100 neurons, averaged over 40 runs.

3. Interpretation

It is first of all striking that in networks with threshold, the generalization ability initially *decreases* with α. This effect was not observed in networks without thresholds [1]. Furthermore, generalization is improved by an optimally chosen threshold throughout. In particular, the generalization ability at $\alpha = 0$ is already higher than the 'random' result $G = 0.5$ and increases with the teacher's threshold. These effects demand some detailed interpretation.

The non-monotonous behaviour can be explained from the student's threshold $T(\alpha, U)$ determined from equations (9)–(11) (figure 2). For $U \neq 0$, $T/U \to \infty$ as $\alpha \to 0$. This again underlines the global role of T. For small α, there is very little information at hand for the student about the direction of B. (In fact, $r \to 0$ and $\Delta \to \infty$ as $\alpha \to 0$.) However, since $U \neq 0$, the student will detect already from a few teacher's answers that the output is biased. Thus a *stochastic* best choice is to set *all* outputs to one value according to the teacher's bias, which is achieved by $T/U \to \infty$ and gives (from equation (6)) a generalization

$$G(\alpha \to 0, U) = \left\langle \Theta\left(\left(-\frac{1}{\sqrt{N}} B \cdot S_\mu + U \right) \operatorname{sign}(U) \right) \right\rangle_{\{S_j^\mu = \pm 1\}} = \int_{-\infty}^{|U|} D\,x \equiv \operatorname{erf}(|U|)\,. \quad (12)$$

Note from the scaling of equation (1) that the onset of generalization improvement is already at very low thresholds since $|U| < \sqrt{N}$ is possible.

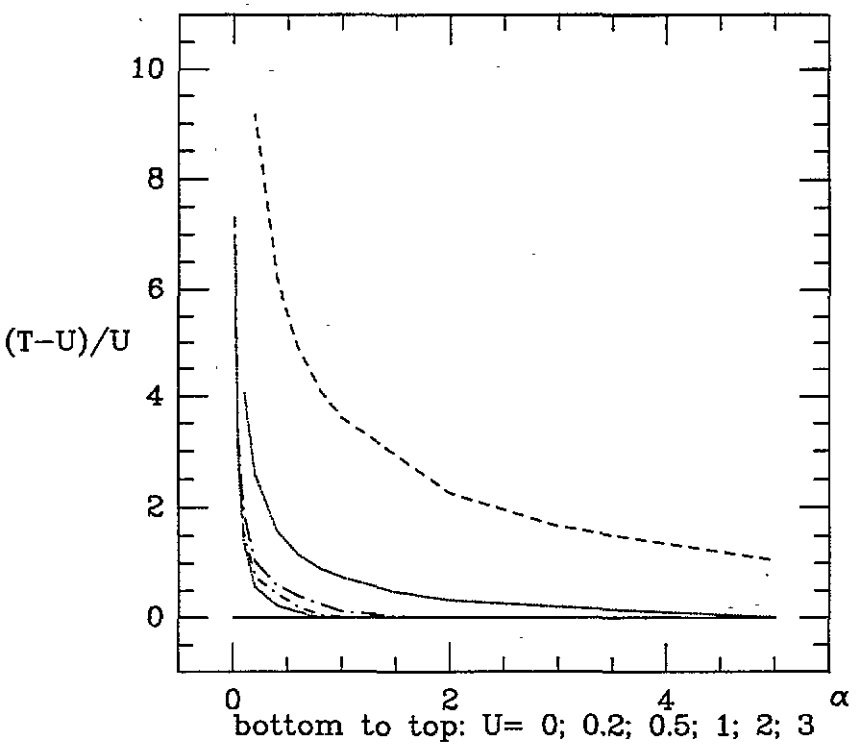

Figure 2. Deviation of student (T) from teacher (U) threshold as a function of capacity for various teacher thresholds U. For $U \neq 0$, T/U becomes infinite for $\alpha \to 0$.

Furthermore, one can evaluate G at small α if one expands equations (6) and (9)–(11) about the known results $r \to 0$, $\Delta \to \infty$ and $T/U \to \infty$. With the parameter $m = 2\,\mathrm{erf}(|U|) - 1 = 1$ which is identical to the absolute output bias (see equation (18)), one obtains (for derivations see the appendix)

$$T = \frac{1}{\sqrt{\alpha}} \frac{m}{\sqrt{1-m^2}} \operatorname{sign}(U) \tag{13}$$

$$\Delta = \frac{1}{\sqrt{\alpha}\sqrt{1-m^2}} \tag{14}$$

$$r = \sqrt{\alpha}\sqrt{\frac{2}{\pi}} \frac{\mathrm{e}^{-U^2/2}}{\sqrt{1-m^2}} \tag{15}$$

and from equations (13)–(15),

$$G = \frac{1+m}{2} - \frac{1}{\sqrt{2\pi}}\sqrt{\alpha}\sqrt{1-m^2}\,\exp\left\{-\frac{m^2}{2\alpha\,(1-m^2)}\right\}. \tag{16}$$

Equation (16) shows the decrease of G with α, the initial slope being $\partial G/\partial\alpha|_{\alpha=0} = 0$, and it also shows that $|G(\alpha) - G(\alpha = 0)|$ decreases with $|U|$. This is exactly the behaviour observed in the curves of figure 1.

The initial decrease of G with α may be interpreted geometrically. With small but increasing α, the threshold T is forcibly reduced since all examples have to be mapped correctly. This gradually removes the initial stochastic behaviour and therefore lowers G.

The onset of the competing effect of deterministically given J and T (which increase G again) is only to be observed at high α. The reason is that in order to 'bind' the $(N-1)$-dimensional separating plane to its correct position, a high number of examples ($\alpha = \mathcal{O}(1)$) is required. Furthermore, only a number $\alpha_{\text{eff}} N$ examples which are 'close' to the gap between the two clusters are effective for the binding. This is reflected in the well known fact [5] that only a subset of patterns, $2/N \leqslant \alpha_{\text{eff}} \leqslant 1$, determine the separation plane. With increasing U, a higher proportion of the randomly selected examples will be far off the gap position and therefore will not contribute to α_{eff}. This explains that with increasing U, the minimum of $G(\alpha)$ is shifted to higher α (see figure 1).

Similarly, for high α one can perform expansions about $r = 1$, $\Delta = 0$ and $T = U$. A self-consistent calculation yields for $\alpha \gg \mathrm{e}^{U^2/2}$:

$$G = 1 - \frac{1}{\alpha} b^2 \mathrm{e}^{b^2/2} + \frac{1}{\alpha^2} f(U) \simeq 1 - \frac{0.501}{\alpha} \tag{17}$$

where $b \simeq 0.639$ is obtained analytically (see appendix). The derivation of this number is valid for networks and rules with or without threshold. From computer simulations, the factor $\alpha(1-G)$ was claimed [6] to be 0.57, in contrast to the analytic value of 0.501 obtained here.

Note that G does not depend on U to leading order in $1/\alpha$, which clearly shows the *deterministic* character of generalization: for large α, the examples fill the input space quasi-homogenously. Since the direction J is almost correctly determined by the student network (in fact, $1 - r \propto 1/\alpha^2$, see appendix), the allowed variation of T only amounts to $\mathcal{O}(1/\alpha^2)$ in G. The student network is therefore deterministically given by the examples. Furthermore, note that the generalization ability almost reaches the Bayes optimal result of $G = 1 - 0.44/\alpha$ obtained numerically in [6]. Thus Bayes optimal learning will improve generalization only very slightly, but at the cost of sampling a (high) number of possible solutions in the full version space [7].

The theoretical results for G can be verified by simulations. For $U = 1$, αN randomly selected examples were used for the student network's training which was performed with the generalized minimum overlap algorithm described in [3]. The generalization ability was then determined as the average $\Theta(\tau_\mu S_\mu^0)$ over 500 randomly selected questions. The simulation results agree well with the theoretical predictions (figure 1).

4. Conclusion and outlook

The optimal cluster separation network has been shown to generalize from examples obeying a teacher's rule. Generalization is stochastic for small α and deterministic for large α. The influence of the teacher's threshold U in the stochastic regime has been quantified, in the deterministic regime it has been shown that U does not influence generalization to lowest order in $1/\alpha$. Simulations verify the theoretical results. Optimal cluster separation networks are able to improve generalization considerably by taking into account the teacher's threshold properly.

In geometric terms, the same conclusions hold. The equivalent task there is to find the correct bipartitioning of input space by inferring from αN randomly drawn normalized examples with given classification, where U corresponds to the distance of the bipartitioning hyperplane from the origin.

Since the latter formulation of the problem is more general, there is a wider range of applicability. For example, the problem could be defined in process control, where an allowed range of parameter values for which the system is stable has to be separated as

far as possible from a forbidden range where the system is unstable. By the equivalence of the two problems, this task can be solved in a linear threshold model by a neural network as described here. In this context, the present paper gives the probability G of correctly classifying an unknown situation as forbidden or allowed. Since allowed and forbidden situations usually do not occur with the same frequency, this probability will rise considerably with the difference in frequencies $\langle\tau_\mu\rangle$, or absolute output bias $m = |\langle\tau_\mu\rangle|$. This can be shown under the assumption that the examples are representative for all possible situations, and, as in (12),

$$\langle\tau_\mu\rangle_{\text{examples}} \simeq \langle\tau_\mu\rangle_{\{S_j^\mu\}} = \left\langle -1 + 2\Theta\left(-\frac{1}{\sqrt{N}} B \cdot S_\mu + U\right)\right\rangle_{\{S_j^\mu=\pm 1\}} = 2\,\mathrm{erf}(U) - 1\,. \tag{18}$$

Thus the difference in frequencies, which can be measured from a few examples, can be mapped to a threshold $U^* = \mathrm{erf}^{-1}(\frac{1}{2}(1 + \langle\tau_\mu\rangle))$. In turn, the respective curve for $G(\alpha)$ as shown in figure 1 then predicts the generalization ability as a function of α. If the separation of parameter ranges is to be done with a confidence level of C (0.95, say), i.e. with a probability C of classifying *any* parameter vector correctly, one can use figure 1 to predict that the number of examples that have to be known (i.e. measured) for this purpose is $N \cdot G^{-1}(G = C, U = U^*)$.

Acknowledgments

It is a pleasure to thank D Sherrington and ACC Coolen for inspiring discussions. Also, I would like to acknowledge support by the Science and Engineering Research Council of Great Britain, the Friedrich–Naumann–Stiftung and the European Community under contract no ERB4001GT922302.

Appendix

Expansion for small α

One first performs the z-integration in equations (6), (9), (10) and (11). The result is

$$G = \int_U^\infty Dw\ \mathrm{erf}\left(\frac{wr - T}{\sqrt{1-r^2}}\right) + \int_{-\infty}^U Dw\ \mathrm{erf}\left(\frac{-wr + T}{\sqrt{1-r^2}}\right) \tag{A1}$$

$$\frac{1}{\alpha} = \int_U^\infty Dw\big[\mathrm{erf}\ A_+ + A_+ f(A_+)\big] + \int_{-\infty}^U Dw\big[\mathrm{erf}\ A_- + A_- f(A_-)\big] \tag{A2}$$

$$0 = \int_U^\infty Dw\ f(A_+)\Big[-w\sqrt{1-r^2} + A_+ r\Big] + \int_{-\infty}^U Dw\ f(A_-)\Big[-w\sqrt{1-r^2} + A_- r\Big] \tag{A3}$$

$$\int_U^\infty Dw\ f(A_+) = \int_{-\infty}^U Dw\ f(A_-) \tag{A4}$$

where

$$A_\pm = \frac{\Delta \pm (T - wr)}{\sqrt{1-r^2}} \qquad f(x) = x\ \mathrm{erf}\ x + \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-x^2/2}. \tag{A5}$$

In the limit $T \to \infty$, $\Delta \to \infty$, $r \to 0$, one obtains the result for small α. Then A_+ and A_- reduce to $A_\pm = \Delta \pm T$. Since $A_+ \to \infty$, from equation (A4) one finds that also $A_- \to \infty$ is required since both sides of the equation must approach infinity. The leading terms in

equations (A4) and (A2) then give

$$(\Delta + T)\ \mathrm{erf}(-U) = (\Delta - T)\ \mathrm{erf}(U) \tag{A6}$$

$$\frac{1}{\alpha} = A_+^2\ \mathrm{erf}(-U)\ + A_-^2\ \mathrm{erf}(U) = \left(\Delta^2 + T^2\right) + 2\Delta|T|m \tag{A7}$$

where $m = 2\,\mathrm{erf}(|U|) - 1$ as before. Combining equations (A6) and (A7) results in Δ and T as in equations (13) and (14). Turning to the expansion of equation (A3), one obtains the leading terms

$$0 = \sqrt{1-r^2}\left[(\Delta + T)\int_U^\infty Dw\,(-w)\ + (\Delta - T)\int_{-\infty}^U Dw\,(w)\right] + \frac{r}{\alpha}$$

$$\Longleftrightarrow\ \frac{r}{\sqrt{1-r^2}} = 2\alpha\frac{\mathrm{e}^{+U^2/2}}{\sqrt{2\pi}}\,\Delta\,.$$

Inserting Δ gives the expresssion (15). G can now be obtained after rewriting (6) as

$$G = \mathrm{erf}(|U|) + \left[\int_{-\infty}^{-|U|} - \int_{-|U|}^{\infty}\right] Dw\ \mathrm{erf}\left(\frac{-|T| - wr}{\sqrt{1-r^2}}\right). \tag{A8}$$

The leading terms are

$$G = \mathrm{erf}(|U|) - m\,\mathrm{erf}(-|T|) = \mathrm{erf}(|U|) - \frac{m}{\sqrt{2\pi}}\frac{\mathrm{e}^{-T^2/2}}{|T|} \tag{A9}$$

where the approximation of the error function for large negative arguments has been used. After insertion of T one obtains equation (16).

Expansion for high α

For high α, equations (9), (10) and (11) must be expanded in the limit $r \to 1$, $\Delta \to 0$ and $T \to U$. Guided by numerical results from computer simulations, we will self-consistently find that $a \equiv A_+(w = U)$ and $b \equiv A_-(w = U)$ approach a finite non-zero value. This can be seen by transforming the variable of integration to A_+ (resp. A_-) and rename it z. Writing $R \equiv \sqrt{r^{-2} - 1}$, the three saddle-point equations read

$$\frac{\sqrt{2\pi}}{R\alpha} = \int_{-\infty}^{a} \mathrm{d}z\,\exp\left(\tfrac{1}{2}\right)[R(a-z)\ + U]^2\left[\mathrm{erf}(z) + zf(z)\right]$$

$$+ \int_{-\infty}^{b} \mathrm{d}z\,\exp\left(\tfrac{1}{2}[R(b-z)\ - U]^2\right)\left[\mathrm{erf}(z) + zf(z)\right] \tag{A10}$$

$$\int_{-\infty}^{a} \mathrm{d}z\,\exp\left(\tfrac{1}{2}[R(a-z)\ + U]^2\right)\left[-R^2(a-z) - RU + z\right]f(z)$$

$$= -\int_{-\infty}^{b} \mathrm{d}z\,\exp\left(\tfrac{1}{2}[R(b-z)\ - U]^2\right)\left[-R^2(b-z) + RU + z\right]f(z) \tag{A11}$$

$$\int_{-\infty}^{a} \mathrm{d}z\ \exp\left(\tfrac{1}{2}[R(a-z)\ + U]^2\right)\ f(z) = \int_{-\infty}^{b} \mathrm{d}z\ \exp\left(\tfrac{1}{2}[R(b-z)\ - U]^2\right)\ f(z)\,. \tag{A12}$$

For $r \to 1$, $R \to 0$, and the R-terms in numerators can be neglected if a, b are finite. Then equation (A12) immediately gives $a = b$, which after inserting into equation (A11) yields

$$0 = \int_{-\infty}^{b} \mathrm{d}z\left(z^2\ \mathrm{erf}(z) + \frac{z}{\sqrt{2\pi}}\mathrm{e}^{-z^2/2}\right) = b^3\ \mathrm{erf}(b) + \frac{1}{\sqrt{2\pi}}(b^2 - 1)\mathrm{e}^{b^2/2}\,. \tag{A13}$$

Equation (A13) is satisfied by the required $b \simeq 0.639$ which self-consistently confirms the above ansatz. With the help of (A11), equation (A10) can be reduced to an expression for R:

$$\frac{\sqrt{2\pi}\mathrm{e}^{+U^2/2}}{2R\alpha} = \int_{-\infty}^{b} \mathrm{d}z\ \mathrm{erf}(z) = b\,\mathrm{erf}(b) + \frac{1}{\sqrt{2\pi}}\mathrm{e}^{-b^2/2} = \frac{\mathrm{e}^{-b^2/2}}{b^2\sqrt{2\pi}} \tag{A14}$$

where, in the last step, equation (A13) was used. This finally gives

$$r \simeq 1 - \tfrac{1}{2}R^2 = 1 - \frac{1}{\alpha^2}\,\frac{1}{2}\pi^2 b^4 \mathrm{e}^{(b^2+U^2)} \simeq 1 - \frac{1.236\,\mathrm{e}^{U^2}}{\alpha^2}. \tag{A15}$$

Note that in the derivation of equation (A14), it is necessary that $\alpha >> \mathrm{e}^{U^2/2}$ which is important for estimations by computer simulations. For G, equation (A1) can be rewritten

$$G = 1 - \int_U^\infty Dw\left(1 - \mathrm{erf}\left(\frac{w - rT}{\sqrt{1-r^2}}\right)\right) - \int_{-U}^\infty Dw\left(1 - \mathrm{erf}\left(\frac{w + rT}{\sqrt{1-r^2}}\right)\right). \tag{A16}$$

For $r \to 1$, a change of variables to $z = (w \mp rT)/\sqrt{1-r^2}$, rewriting r into R and neglecting terms proportional to R in the integrand gives

$$G = 1 - \frac{2}{\sqrt{2\pi}}\mathrm{e}^{-U^2/2}\,R\int_0^\infty \mathrm{d}z(1 - \mathrm{erf}(z)) = 1 - \frac{R\,\mathrm{e}^{-U^2/2}}{\pi} \tag{A17}$$

where from the definitions of a and b the lower integration limit is

$$(0.5(b-a) + UR)/(1+R^2) \to 0.$$

Inserting R from equation (A14) then gives the desired result (17).

References

[1] Opper M *et al* 1990 On the ability of the optimal perceptron to generalise *J. Phys. A: Math. Gen.* **23** L581–6

[2] Rujan P 1993 A fast method for calculating the perceptron with maximal stability *J. Physique I* **3** 277–90

[3] Wendemuth A 1993 Learning optimal threshold and weights for the perceptron of maximum stability *Preprint* Oxford University

[4] Gardner E 1988 *J. Phys. A: Math. Gen.* **21** 257

[5] Fletcher R 1987 *Practical Methods of Optimization* (New York: Wiley)

[6] Opper M and Haussler D 1991 *Phys. Rev. Lett.* **66** 2677

[7] Watkin T, Rau A and Biehl M 1993 The statistical physics of generalization *Rev. Mod. Phys.* **65** 499