Synapse removal in discrete neural networks

1995 J. Phys. A: Math. Gen. 28 L25

(http://iopscience.iop.org/0305-4470/28/1/005)

LETTER TO THE EDITOR

Synapse removal in discrete neural networks

R R Viswanathan
Institute of Systems Science, National University of Singapore, Singapore 0511

Received 26 September 1994

Abstract. The effect of removing synapses of particular magnitudes on the retrieval property of perceptron neural networks is determined by means of a space-of-interactions calculation. For a discrete state perceptron, an asymptotic (in number of input neurons) upper bound on the number of errors made in a set of given stored patterns due to synapse removal is computed as a function of the number and size of synapses removed.

There are several ways to quantify fault tolerance in neural networks. In [1–3], the effect of randomly diluting the connections in a neural network on the storage capacity was examined in differing contexts. Another measure of fault tolerance when neurons fail is given by the storage capacity of patterns that continue to be recalled without error, even when some of the neurons in the network fail. This measure of robustness was computed in an earlier paper [4].

In contrast to results of this nature, a different measure of robustness is provided by the proportion of errors made in the recall of already stored patterns when some of the synapses or connections in the network, of some definite magnitudes, are removed from the original network.

The technique of trying to construct minimally-sized neural networks by starting from a larger network, and then trimming it down to size by eliminating redundant synapses and neurons in the network, is quite common among people engaged in the synthesis of neural networks. If there is a large saving in the size of the network achieved by this technique, then a small proportion of errors in the retrieval may be tolerable. In an earlier paper, the error made in the recall of stored patterns when the patterns are real valued and uniformly distributed, and when some of the synapses are removed, was computed by means of a geometric argument [5]. Also, the error resulting due to deletion of the smallest magnitude synapse was estimated in the case of discrete patterns by combinatorial means. The more general case of the deletion of a significant number of the synapses was found to be difficult to tackle combinatorially.

Here we shall use a well known statistical physics technique to compute an upper bound on the proportion of errors resulting due to synapse deletion in the case of discrete-valued patterns, for an arbitrary number of synapse deletions (and for any magnitudes of these synapses).

It should be mentioned at this point that the case of discrete patterns is distinct from that of continuous patterns in terms of the effect of synapse removal. For uniformly distributed continuous patterns, it is not hard to see [5] that the fraction of errors due to synapse removal depends only on the angle between the initial and final vector of weights. In general, this would not be the case for discrete patterns; for instance, if the number of stored patterns to

0305-4470/95/010025+07$19.50

begin with is less than maximal, there may be enough 'room' in weight space to remove weights without affecting the quality of storage. The results below indeed indicate that this expectation is justified.

The basic quantity of interest is the partition function in the space of interactions [6, 7], which we define in what follows. We consider the case when the neuron states are discrete valued and take values ±1. We denote the states of the input layer of neurons by s_i, with $i = 1, \ldots, N$. For convenience, we also assume that there is a single output neuron whose state (±1) we denote by σ. If p patterns (s^μ, σ^μ) are stored by the network, for $\mu = 1, \ldots, p$, then the relations

$$\sigma^\mu = \operatorname{sign}\left(\frac{1}{\sqrt{N}}\sum_{i=1}^{N} w_i s_i^\mu\right) \tag{1}$$

are satisfied. The factor $\frac{1}{\sqrt{N}}$ is introduced for convenient normalization. Patterns are therefore classified by the network according to this rule. Since this functional relation is invariant to an overall scale change of the w's, we choose the spherical normalization

$$\sum_{i=1}^{N} w_i^2 = N \tag{2}$$

for the weights w_i. We shall find it convenient to refer to the coefficient of N, on the right-hand side here, as the strength of the set of weights on the left-hand side; thus, the entire set of weights has strength unity.

Now suppose that some of the synapses w_i, $k = k'N$ in number, are removed. Thus, k' is the fraction of synapses removed. Suppose, furthermore, that this set of weights, which we can take without loss of generality to be the set $(w_1, w_2, \ldots, w_k)$, has strength s, i.e.

$$\sum_{j=1}^{k} w_j^2 = sN. \tag{3}$$

The question we would like to address is then, what proportion of the original number of stored patterns p is misclassified by the trimmed network?

Given a set of patterns classified correctly by the original untrimmed network, some of the patterns would, in general, be classified correctly even by the trimmed network, while others would acquire a non-zero probability of misclassification. However, this probability would, in general, depend on the particular set of stored patterns. Since the assumption of self-averaging is crucial to the space-of-interactions method, we will be concerned here simply with the evaluation of the maximum number of correctly classified patterns p_I that can be stored by the trimmed network; this number *can* be computed by evaluating the quenched average of an appropriate free energy. The ratio of the remaining number of patterns to the original number of patterns stored by the untrimmed network

$$\epsilon_\mathrm{u} = \frac{p - p_\mathrm{I}}{p} \tag{4}$$

therefore constitutes a best-case upper bound on the proportion of errors introduced by the trimming process. It is a best-case bound because we are evaluating the *maximum* number of patterns p_I classified correctly by the trimmed network, which corresponds to the case when the set of k trimmed weights $(w_1, \ldots, w_k)$ with strength s is essentially unique.

We are now ready to write down the partition function in the space of weights. The partition function is simply the volume in weight space that corresponds to the correct storage of input patterns by both the untrimmed and trimmed networks:

$$Z = \int \left(\prod_{i=1}^{N} \mathrm{d}w_i \right) \rho[w] \prod_{\mu=1}^{p_1} \theta\left(\frac{1}{\sqrt{N}} \sigma^{\mu} \sum_{i=1}^{N} w_i s_i^{\mu} \right) \theta\left(\frac{1}{\sqrt{N}} \sigma^{\mu} \sum_{j=k+1}^{N} w_j s_j^{\mu} \right) \tag{5}$$

where

$$\rho[w] = \delta\left(\sum_{i=1}^{N} w_i^2 - N \right) \delta\left(\sum_{j=1}^{k} w_j^2 - sN \right) \tag{6}$$

is a measure which normalizes the full set of weights and fixes the total strength of the removed synapses (recall that we are assuming that synapses w_1 to w_k are removed) to s. The θ functions here ensure that both the untrimmed and the trimmed networks classify the input patterns correctly, so that the relations

$$\sigma^{\mu} = \mathrm{sign}\left(\frac{1}{\sqrt{N}} \sum_{i=1}^{N} w_i s_i^{\mu} \right)$$

and

$$\sigma^{\mu} = \mathrm{sign}\left(\frac{1}{\sqrt{N}} \sum_{j=k+1}^{N} w_j s_j^{\mu} \right)$$

hold simultaneously for all p_1 patterns.

The quenched free energy $f = \langle\langle \log Z \rangle\rangle$, averaged over input s_i^{μ} and output states σ^{μ} for all patterns, is what we want to evaluate. We shall use the replica trick to do this, as is usual. Since we have assumed that the patterns are random, we have

$$\langle\langle s_i^{\mu} s_j^{\nu} \rangle\rangle = \delta_{ij} \delta_{\mu\nu}.$$

With a and b denoting replica indices, it is then necessary to define the following order parameters which arise in the evaluation of the quenched average:

$$q_{ab} = \frac{1}{N} \sum_{i=1}^{N} w_i^a w_i^b \tag{7}$$

and

$$\overline{q}_{ab} = \frac{1}{N} \sum_{j=k+1}^{N} w_j^a w_j^b. \tag{8}$$

We note that q measures the average overlap in weight space between replicas of the whole network, while $\overline{q}$ measures the overlap in weight space between replicas of the trimmed network (with k weights of total strength s removed).

These constraints in weight space are implemented as delta-function constraints; writing integral representations for the delta functions, one each for q_{ab} and $\overline{q}_{ab}$, involves the introduction of further order parameters F_{ab} and K_{ab}, respectively. In addition, when measure (6) is written in terms of integral representations for the respective delta functions, two more order parameters E^a and H^a need to be introduced, respectively.

The integrals can be evaluated in the saddle-point approximation, which is valid for N large. In this limit, with the replica-symmetric ansatz $q_{ab} = q, \overline{q}_{ab} = \overline{q}$, $F_{ab} = F$, $K_{ab} = K$, $E^a = E$ and $H^a = H$, the relevant integrals can be computed by the saddle-point method, and the appropriate saddle-point equations can be written down in relatively simple form.

We note at this point that the value of q is determined by the value of p_1, since the same p_1 patterns are stored by the untrimmed network as well. This computation relating q to the number of stored patterns was performed first in [6], and the equation relating q to $\alpha_1 = p_1/N$ can be written in the form

$$\frac{q}{1-q} = \frac{\alpha_1}{2\pi}\int_{-\infty}^{\infty}\mathrm{D}z\,\frac{\mathrm{e}^{-\tau^2}}{H^2(\tau)} \tag{9}$$

with the variable $\tau = z\sqrt{q/1-q}$ and where Dz is the Gaussian measure $\mathrm{D}z \equiv \mathrm{e}^{-z^2/2}/\sqrt{2\pi}$. Here $H(\tau)$ denotes the function $H(\tau) \equiv \int_{\tau}^{\infty}\mathrm{D}z$.

The saddle-point equations then express the fact that the free energy has zero derivative at the stationary point with respect to the order parameters $\overline{q}$, F, H, E and K. The saddle-point equations for the last four variables here can be used to write the order parameter K in terms of the strength s of the synapse chops and the fractional number k' of synapses chopped:

$$K = \frac{(1-k')(1-s)-k's}{(\overline{q}+s-1)^2}. \tag{10}$$

The saddle-point equation for $\overline{q}$ reads:

$$\frac{K}{2} + \alpha_1\frac{\partial}{\partial\overline{q}}\left(\int_{-\infty}^{\infty}\mathrm{D}z\int_{-\infty}^{\infty}\mathrm{D}t\,\log\left(\int_{\overline{\tau}}^{\infty}\mathrm{D}u\,H(y)\right)\right) = 0 \tag{11}$$

where we have defined

$$\overline{\tau} \equiv z\sqrt{\overline{q}/(1-s-\overline{q})}$$

and

$$y \equiv \frac{1}{\sqrt{s-q+\overline{q}}}\left(t\sqrt{q-\overline{q}} - \left(u\sqrt{1-s-\overline{q}} - z\sqrt{\overline{q}}\right)\right).$$

The capacity α_1 is maximum when $\overline{q} \to (1-s)$, which means that there is effectively only one way to choose $k = k'N$ weights with strength s (these are the chopped weights); this is the 'best case' referred to earlier. In this limit, equation (11) takes the simplified form

$$(1-k')(1-s) - k's - \frac{1-s}{4}\alpha_1\left(\frac{1}{4\pi}\sqrt{\frac{q+s-1}{1-s}} + \frac{q+s-1}{1-s} + 2\right) = 0. \tag{12}$$

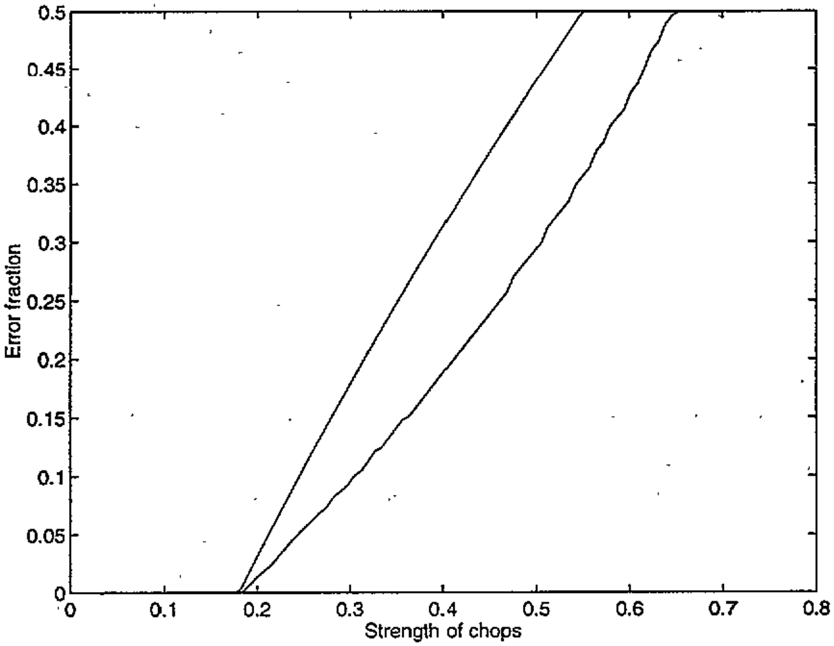

Figure 1. The bound on the error fraction as a function of the strength of chops s. The curve on the left-hand side is for $k' = 0.4$, and that on the right-hand side is for $k' = 0.2$.

Equations (9) and (12) must be solved together to determine α_1 as a function of k' and s. A best-case upper bound on the classification error of the trimmed network when $p_1 < p$ is then provided by

$$\epsilon_u = 1 - \frac{\alpha_1}{\alpha} \tag{13}$$

where $\alpha = p/N$ and p is the initial number of patterns stored by the original network. Note that when all synapses are removed, the output is merely guessed at random, so that one expects 50 percent errors in this case. The value $\epsilon = 0.5$ is therefore a default upper bound.

The bound given by our calculation is not always tight; in fact it exceeds 0.5 for some values of s. In this case, we use the default bound of one half. However, for smaller s values, we expect the bound to be tighter. For small s and k' values, in fact, $\alpha_1 > \alpha$, which means that there is more than enough 'room' in weight space to accomodate the synapse removals without compromising on the storage of all of the original patterns; in this case, the process of synapse removal results in zero error.

A numerical solution of these equations is provided in figure 1, which gives an upper bound on the error produced, due to synapse chopping as a function of the strength s of the chops, for two different values of the fractional number of synapses chopped k'. The 'wiggles' in the figure are due to the finite accuracy of our numerical solution. It is assumed that $p = 3N/2$ patterns are stored to begin with. In contrast to the case of real valued inputs, where it was found that the error only depended on the strength of the chops [5], the bound we have obtained for discrete patterns depends on both the number and strength of the chopped weights.

In conclusion, we have seen here how a best-case upper bound on the errors introduced in patterns stored in a perceptron upon removal of some of the synapses may be estimated by using statistical mechanics. This bound is asymptotic in the size of the network. It would be of further interest to obtain similar results for more general neural architectures as well.

References

[1] Sompolinsky H 1986 *Phys. Rev.* A 34 2571

[2] Bouten M, Engel A, Komoda A and Serneels R 1990 *J. Phys. A: Math. Gen.* **23** 4643

[3] Komoda A, Serneels R, Wong K Y M and Bouten M 1991 *J. Phys. A: Math. Gen.* **24** L743

[4] Viswanathan R R 1994 *Phys. Lett.* **188A** 55

[5] Srinivasan S H, Vinay V and Viswanathan R R to appear in *Proc. Int. Conf. on Neural Information Processing (Seoul, 1994)*

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

[7] Gardner E and Derrida B 1988 *J. Phys. A: Math. Gen.* **21** 271