Unicycling Helps Your French: Spontaneous Recovery of Associations by Learning Unrelated Tasks

Inman Harvey and James V. Stone

CSRP 379, May 1995

Cognitive Science Research Paper

Serial No. CSRP 379

The University of Sussex
School of Cognitive and Computing Sciences
Falmer
BRIGHTON BN1 9QH
England, UK

inmanh@cogs.susx.ac.uk
jims@cogs.susx.ac.uk

This version (revised October 1995) is that
accepted by *Neural Computation*.
In Press 1995.

Unicycling Helps Your French: Spontaneous Recovery of Associations by Learning Unrelated Tasks

Inman Harvey
School of Cognitive and Computing Sciences
University of Sussex, Brighton BN1 9QH, England
inmanh@cogs.susx.ac.uk

James V. Stone
School of Biological Sciences
University of Sussex, Brighton BN1 9QH, England
jims@cogs.susx.ac.uk

May 1995 — revision October 1995

Abstract

We demonstrate that imperfect recall of a set of associations can usually be improved by training on a new, *unrelated* set of associations. This spontaneous recovery of associations is a consequence of the high dimensionality of weight spaces, and is therefore not peculiar to any single type of neural net. Accordingly, this work may have implications for spontaneous recovery of memory in the central nervous system.

1 Introduction

A spontaneous recovery effect in connectionist nets was first noted in (Hinton & Sejnowski, 1986), and analysed in (Hinton & Plaut, 1987). A net was first trained on a set of associations, and then its performance on this set was degraded by training on a new set. When retraining was then carried out on a proportion of the original set of associations, performance also improved on the remainder of that set.

In this paper a more general effect is demonstrated. A net is first trained on a set of associations, called task $\mathcal{A}$; and then performance on this task is degraded, either by random perturbations of the connection weights, or as a result of learning a new task $\mathcal{B}$. Performance on $\mathcal{A}$ is then monitored whilst the net is trained on another new task $\mathcal{C}$. The main result of this paper is that in most cases performance on the original task $\mathcal{A}$ initially improves.

The following is a simplistic analogy, which assumes that this effect carries over to human learning of cognitive tasks. If you have a French examination tomorrow, but you have forgotten quite a lot of French, then a short spell of learning some new task, such as unicycling, can be expected to improve your performance in the French examination. Students of French should be warned not to take this fanciful analogy too literally; it requires the implausible assumption that French and unicycling make use of a common subset of neuronal connections.

We will first give an informal argument to explain the underlying geometrical reasons for this effect; follow this with an analysis of how it scales with the dimensionality of weight-space; and then demonstrate it with some examples.

2 High Dimensional Geometry

A number of assumptions will be used here; later we will evaluate their validity.

Learning in connectionist models typically involves a succession of small changes to the connection weights between units. This can be interpreted as the movement of a point W in weight-space, the dimensionality of which is the number of weights. For the present, we assume that training on a particular task $\mathcal{A}$ moves W in a straight line towards a point A, where A represents the weights of a net which performs perfectly on task $\mathcal{A}$; we also assume that distance from A is monotonically related to decrease in performance on task $\mathcal{A}$.

Let A be the position of W after task $\mathcal{A}$ has been learned (see Figure 1). Assume that some 'forgetting' takes place, either through random weight changes, or through some training on a different task, which shifts W to a new point B. The point B lies on the surface of $\mathcal{H}$, a hypersphere of radius $r = |A - B|$ centred on A.

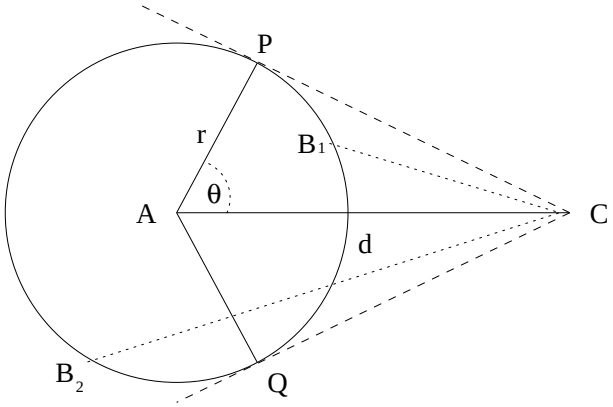

Figure 1: *The circle is a 2-D representation of hypersphere $\mathcal{H}$. Initial movement from a point B on the circumference towards C has two possible consequences: trajectory $B_1 \rightarrow C$ is outside $\mathcal{H}$, whereas $B_2 \rightarrow C$ intersects $\mathcal{H}$.*

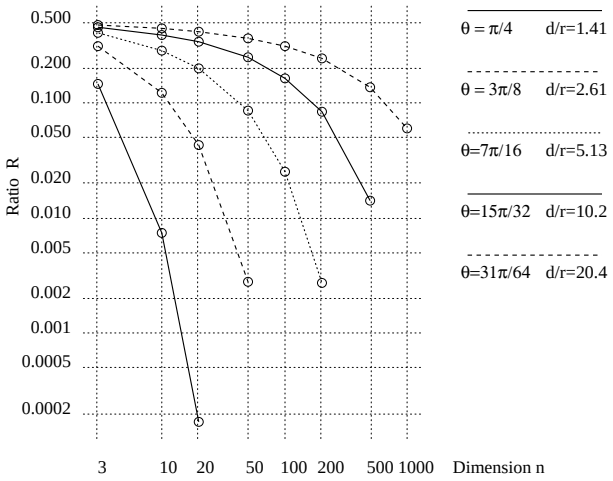

Figure 2: *Graph of ratio $\mathcal{R}_{n,\theta}$ against n. Both axes are logarithmically scaled. Data points are marked by circles on the lines from $\theta = \pi/4$ on left to $31\pi/64$ on the right.*

We then initiate training on a task $\mathcal{C}$ which is unrelated to task $\mathcal{A}$; under our assumptions, training moves W from B towards a point C, which is distance $d = |A - C|$ from A. If the line connecting B to C passes through the volume of $\mathcal{H}$ then the distance $|W - A|$ initially decreases as W moves towards C. In such cases, training on task $\mathcal{C}$ initially causes improvement in performance on task $\mathcal{A}$.

We assume that point A has been chosen from a bounded set of points $\mathcal{S}$, which may have any distribution; that $\mathcal{H}$ is centred on A; that B is chosen from a uniform distribution over the surface of $\mathcal{H}$; and that C is chosen from $\mathcal{S}$ independently of the positions of A or B. What, then, is the probability that line segment BC passes through $\mathcal{H}$? That is, what is the probability that training on task $\mathcal{C}$ generates spontaneous recovery on task $\mathcal{A}$?

If C lies within $\mathcal{H}$ (i.e. if $d < r$) then recovery is guaranteed. For any point C outside $\mathcal{H}$ there is a probability $p \geq 0.5$ of recovery on task $\mathcal{A}$. Figure 2 demonstrates this for a two-dimensional space. The point B may lie anywhere on the circumference of $\mathcal{H}$. The line segment BC only fails to pass through $\mathcal{H}$ if B lies on the smaller arc PQ; where CP and CQ are tangents to the circle, and hence $cos(\theta) = r/d$. Thus $p \geq 0.5$, and $p \rightarrow 0.5$ as $d \rightarrow \infty$.

Consider the extension to a third dimension, while retaining the same values r, d and θ. The probability $q = (1-p)$ that BC *fails* to pass through the sphere $\mathcal{H}$ is equal to the proportion of the surface of $\mathcal{H}$ which lies within a cone defined by PCQ with apex C. This proportion is considerably smaller in 3-D than it is in 2-D. We produce a formula for this proportion for n-dimensions in the next section. We demonstrate analytically what can be seen intuitively, namely that for any given $\theta < \pi/2$, as n increases q tends to zero.

3 Analysis

Let $\mathcal{S}(n, r, \theta)$ be the surface 'area' of the segment of a hypersphere of radius r in n-dimensions, subtended by a (hyper-) cone of half-angle θ; this segment is not a surface *area*, but rather a surface hypervolume of dimensionality $(n-1)$. The surface 'area' of the complete hypersphere is $\mathcal{S}(n, r, \pi)$. For some constant k_n, $\mathcal{S}(n, r, \pi) = k_n r^{n-1}$. We can use this to calculate $\mathcal{S}(n, r, \theta)$ by integration:

$$\begin{aligned} \mathcal{S}(n, r, \theta) &= \int_{\alpha=0}^{\alpha=\theta} \mathcal{S}(n-1, r\,sin(\alpha), \pi)\; r\; d\alpha \\ &= k_{n-1}\; r^{n-1} \int_0^{\theta} sin^{n-2}(\alpha)\; d\alpha \end{aligned}$$

We require the ratio $\mathcal{R}_{n,\theta}$ of $\mathcal{S}(n, r, \theta)$ to $\mathcal{S}(n, r, \pi)$.

$$\mathcal{R}_{n,\theta} = \frac{\int_0^{\theta} sin^{n-2}(\alpha)\, d\alpha}{\int_0^{\pi} sin^{n-2}(\alpha)\, d\alpha}$$

This ratio $\mathcal{R}_{n,\theta}$ gives the probability that the line segment BC (in Figure 1, generalised here to n-dimensions) *fails* to pass through the hypersphere, and is therefore equal to q.

In Figure 2 we plot the ratio $\mathcal{R}_{n,\theta}$ against the dimensionality n, for values of θ from $\pi/4$ to $31\pi/64$. These values of θ are associated with corresponding values of d/r (see Figure 2) between 1.41 and 20.4. For a given value of d/r, as the dimensionality n increases, the ratio $\mathcal{R}_{n,\theta}$ tends to zero. For large n, it is almost certain that the line segment BC passes through the hypersphere $\mathcal{H}$.

This implies that initially the point W moves from B closer to A. Hence performance improves, at least temporarily, on task $\mathcal{A}$.

Returning to the assumptions stated earlier, we can now examine their validity. First, an irregular error surface ensures that training does not, in general, move W in a straight line. Second, if perturbation from A to B is achieved by training on a task $\mathcal{B}$ then B is chosen from a distribution over $\mathcal{H}$ which may not be uniform. Third, perfect performance on task $\mathcal{C}$ may be associated not with one point C, but with many points which are equivalent in that they each provide a similar mapping from input to output. W may move towards the nearest of many Cs, which is therefore not chosen from $\mathcal{S}$ independently of A. This may alter the probability that W passes through $\mathcal{H}$. Fourth, if B lies on a hypersphere of dimensionality $m < n$ then the probability that spontaneous recovery occurs may be reduced.

Despite these considerations, evidence of the effects predicted above can be observed.

4 Experimental results

In two sets of experiments a net was initially trained using back-propagation on a task $\mathcal{A}$. The net had 10 input, 100 hidden, and 10 output units. The hidden and output units had logistic activation functions; weights were initialised to random values in the range $[0.3, -0.3]$. Task $\mathcal{A}$ was defined by 100 pairs of binary vectors which were chosen (without replacement) from the set of 2^{10} vectors. The members of each pair were used to train the net using batch update for 1300 training epochs, with a learning rate $\eta = 0.02$ and momentum $\alpha = 0.9$[1]. After initial training on task $\mathcal{A}$, the weights were perturbed from A by one of two different methods.

Experiment 1: Perturbing Weights by New Training

After training on $\mathcal{A}$, the net was trained for 400 epochs on 5 new[2] vector pairs in order to perturb the weights away from A to a point B. Finally, the net was trained on a further 5 new vector pairs (task $\mathcal{C}$) for 50 epochs. During training on $\mathcal{C}$, the performance of the net on task $\mathcal{A}$ was monitored. As predicted by the analysis above, performance on task $\mathcal{A}$ usually showed a transient improvement as training on task $\mathcal{C}$ progressed. This procedure was repeated 380 times using a single $\mathcal{A}$, 20 $\mathcal{B}$'s and 20 $\mathcal{C}$'s[3]. Figure 3 shows how many runs improved or declined in performance on task $\mathcal{A}$ at the end of each epoch (together with a few runs which showed no change, given the precision used in calculations). A proportion 241/380 (63.4%) showed incidental relearning on $\mathcal{A}$ after the first epoch, but this dropped to less than 50% after the 5th epoch.

Experiment 2: Perturbing Weights Randomly

After training on task $\mathcal{A}$ as above, the weights of the net were perturbed by adding uniform noise. In Experiment 1, the distance $|A - B|$ had a mean of about 7. In order to make perturbations of comparable magnitude W was perturbed from A to B by adding a random vector of length 7. As described in Experiment 1 this was repeated 380 times. The proportion of the runs which showed incidental relearning is given in Figure 3; this was 248/380 (65.3%) after the first epoch, and remained above 50% for the 50 epochs.

[1] This was designed to be comparable to the situation described in (Hinton & Plaut, 1987).

[2] Here "new" implies that the none of these vectors exist in any previous training set.

[3] B and C were chosen without replacement from 20 sets of 5 vector pairs, giving $20 \times 19 = 380$ different possibilities.

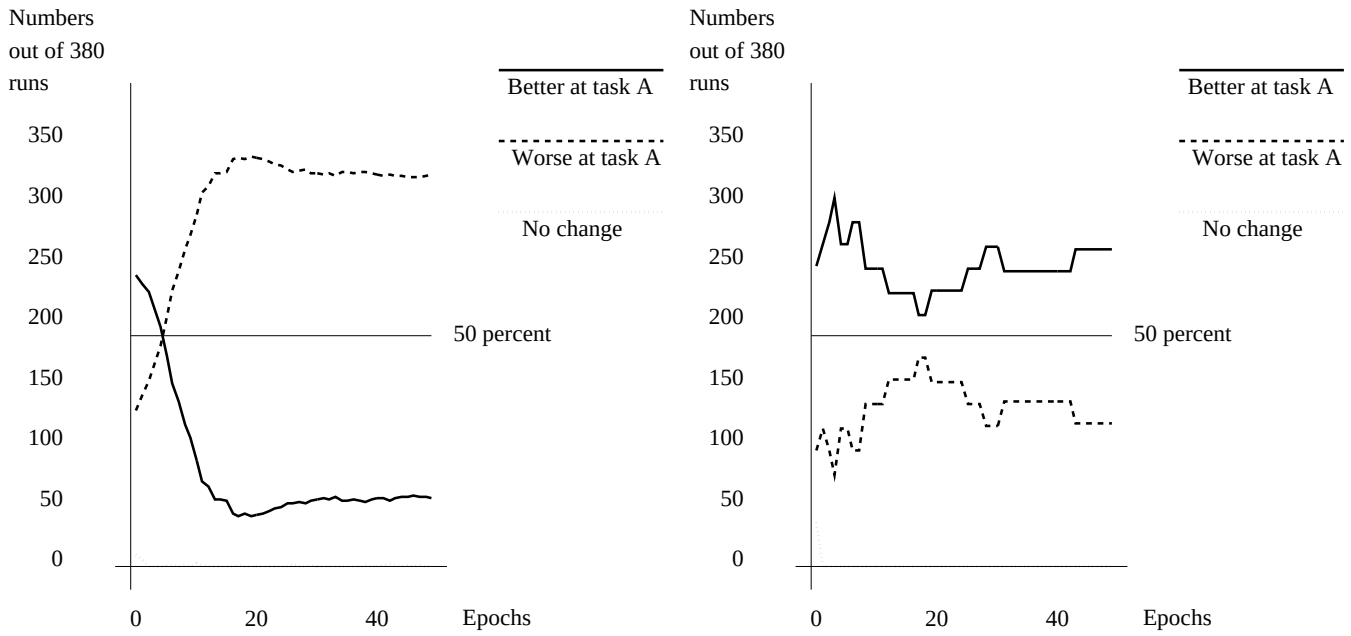

Figure 3: *Graphs of 380 runs, showing numbers improving in performance on $\mathcal{A}$, during 50 epochs of training on $\mathcal{C}$. On the left, experiment 1, the weight vector W was perturbed from A to B by training on task $\mathcal{B}$. On the right, experiment 2, W was perturbed by a randomly chosen vector of length 7 from A to B.*

5 Discussion

A new effect has been demonstrated, such that performance on some task $\mathcal{A}$ improves initially (from a degraded level of performance), when training is started on an *unrelated* task $\mathcal{C}$. This has been analysed in terms of the geometrical properties of high-dimensional spaces. In applying this to weight-spaces, we rely on simplistic assumptions about the way training on a task relates to movement through weight-space. The effect can be observed even if these assumptions are violated, as demonstrated by experiment.

The graphs show evidence of spontaneous recovery. The effect can be seen to be short-lived in the first case, in which perturbation was achieved by retraining on $\mathcal{B}$, and sustained in the second case, in which perturbation was random. Only in the latter case can we expect B to have been chosen from a unbiased distribution over the surface of $\mathcal{H}$, unrelated to the position of C. The graphs indicate only the probabilities of improvement, without reference to the magnitudes of these effects in individual runs.

This recovery effect may be relevant to counter-intuitive phenomena described in (Parisi, Nolfi, & Cecconi, 1992; Nolfi, Elman, & Parisi, 1994), and may also contribute to the effect described in (Hinton & Sejnowski, 1986). It has been suggested[4] that the effect may be related to James-Stein shrinkage (Efron & Morris, 1977; James & Stein, 1961). That is, reducing the variance of (net) outputs reduces the squared error at the expense of introducing a bias. It may be that training on $\mathcal{C}$ incidentally induces shrinkage.

The observed effect is weaker than that predicted from the geometrical argument given above, presumably due to the simplistic nature of the assumptions used therein. However, the effect is robust inasmuch as it does not depend on the learning algorithm, nor on the type of net used. For this reason, the effect may have implications for spontaneous recovery of memory in the central nervous system.

Acknowledgments

Funding for the authors has been provided by the E.P.S.R.C. and the Joint Council Initiative. We thank the referees for useful comments.

[4] G. E. Hinton, personal communication.

References

Efron, B., & Morris, C. (1977). Stein's paradox in statistics. *Scientific American, 236*(5), 119–127.

Hinton, G., & Plaut, D. (1987). Using fast weights to deblur old memories. In *Proceedings of the 7th Annual Conference of the Cognitive Science Society, Seattle.*

Hinton, G., & Sejnowski, T. (1986). Learning and relearning in Boltzmann machines. In Rumelhart, D., McClelland, J., & the PDP Research Group (Eds.), *Parallel Distributed Processing: Explorations in the Microstructure of Cognition. Volume 1: Foundations*, pp. 282–317. MIT Press/Bradford Books, Cambridge MA.

James, W., & Stein, C. (1961). Estimation with quadratic loss. In *Proc. of the Fourth Berkeley Symposium on Mathematical Statistics and Probability 1*, pp. 361–380 Berkeley, California. University of California Press.

Nolfi, S., Elman, J., & Parisi, D. (1994). Learning and evolution in neural networks. *Adaptive Behavior, 3*(1), 5–28.

Parisi, D., Nolfi, S., & Cecconi, F. (1992). Learning, behavior and evolution. In Varela, F. J., & Bourgine, P. (Eds.), *Toward a Practice of Autonomous Systems: Proceedings of the First European Conference on Artificial Life*, pp. 207–216. MIT Press/Bradford Books, Cambridge, MA.