Learning in sparse attractor networks with inhibition

Si Wu[1] and Thomas Trappenberg[2]

[1] Sussex University, UK,
[2] Dalhousie University, Canada

Abstract. Attractor networks are important models for brain functions on a behavioral and physiological level, but learning on sparse patterns has not been fully explained. Here we show that the inclusion of the activity dependent effect of an inhibitory pool in Hebbian learning can accomplish learning of stable sparse attractors in both, continuous attractor and point attractor neural networks.

1 Introduction

Recurrent attractor neural networks (ANNs) are a fundamental ingredient in many models of brain functions [1, 2]. Probably best known are point attractor neural networks (PANNs) which are trained on random patterns with a Hebbian covariance rule, such as the one popularized by [3]. Another popular type is that of continuous attractor neural networks (CANNs) where the weight matrix is commonly chosen to be of the on-center-off-surround type [2, 4]. Such models have been proposed from basic physiological principles [5] as well as from their ability to describe the dynamics of cognitive functions [6]. While basic Hebbian training has long been described in PANNs [7–9], training on sparse patterns through activity dependent inhibition in PANNs and training in CANNs have not been fully addressed. By sparsity we mean that only a small portion of nodes are active in the network when a single memory pattern is retrieved. In this paper we show that training with inhibition can stabilize sparse network activity in both PANNs and CANNs.

2 Learning in CANNs

We consider a simple CANN model, as used in [2]. The nodes are uniformly distributed in a feature space of range $(-\pi, \pi]$ with periodic conditions. The neuronal states take on binary values of 0 and 1, and the memory patterns, μ, are sparse attractors of localized activity packets (bumps) in the sense that in each pattern $d \ll 2\pi$ consecutive nodes are activated. We further consider that the network holds a continuous family of memory patterns uniformly distributed in the feature space. Conventional studies (e.g.[2]) often assume this form of the recurrent interactions. Here, we derive the recurrent interaction structure based on a properly modified Hebbian learning rule.

We propose a general Hebbian covariance learning rule which is augmented with an additional inhibition constant, C, describing the effect of inhibitory internodes,

$$w_{i,j} = \frac{2\pi}{M} \sum_{m=1}^{M} (\mu_i^m - < \mu_i >)(\mu_j^m - < \mu_j >) - C, \tag{1}$$

where M denotes the number of training patterns, and the average activity of the ith node over all patterns is $< \mu_i >= \sum_m \mu_i^m / M = d/(2\pi)$.

Since the neural field is translation invariant, the interaction between two nodes is determined by their distance in feature space. Without loss of generality, we can thus calculate the interaction between nodes at locations 0 and x. Under the continuous field approximation, $M \to \infty$ with pattern density is $\rho = M/(2\pi)$, the weight is given by

$$w(x) = \int_{-\pi}^{\pi} \mu_0^y \mu_x^y dy - C - d^2/2\pi, \tag{2}$$

with $\mu_x^y = 1$ when $y \leq x \leq y + d$ and $\mu_x^y = 0$ otherwise. We get:

- when $0 < |x| < d$: $w(x) = d - |x| - C - d^2/(2\pi)$
- when $d < |x| < \pi$: $w(x) = -C - d^2/(2\pi)$

Thus, the weight profile describes short-range excitatory and long-range inhibitory interactions as demanded by the center-surround neural field theory [2, 4].

2.1 Stability under the network dynamics

We denote the network states at time t with $\mathbf{S}(t) = \{S_i(t)\}$. Under the continuous field approximation, the network dynamics can be written as:

$$S(x) = \Theta[\int_{-\pi}^{\pi} w(x - z) S(z) dz], \tag{3}$$

where $\Theta(x)$ is a threshold function. We can check the stability of a memory pattern, μ, of a bump at location $[0, d]$, under the network dynamics. This requires:

$$\mu(x) = \Theta[\int_{-\pi}^{\pi} w(x - z) \mu(z) dz. \tag{4}$$

Since the inputs received by nodes in the middle of the bump are always larger than that at the boundaries (due to short-range excitation and long-range inhibition), it is adequate to only check the stability of the boundary points. The recurrent input received by the boundary point, 0, is given by

$$h(0) = \int_0^d w(z) dz = \int_0^d (d - z - C - d^2/(2\pi)) dz = d^2/2 - Cd - d^3/(2\pi). \tag{5}$$

The activity packet (bump) is stable when $h(0) = 0$. From the above equation we see that in case of a pure covariance rule ($C = 0$) the size of the bump can only be $d = \pi$. That is, the memory patterns are not sparse. In order to hold sparse patterns, a inhibition of $C \neq 0$ is required.

In the following, we absorb the constant term from the covariance learning rule (the third term on the right-hand side of eq. 5) in a revised inhibition constant C. Then, the condition for stabilizing sparse patterns of size d in CANNs is $C = d/2$. It is straightforward to check that this is also the sufficient condition: consider the network state

starts from a bump larger than d, then it will shrink due to the recurrent interactions, and if the initial state is smaller than d, it will enlarge.

For training patterns with width d we can modulate the retrieval width $\tilde{d} < d$ with different inhibition values:

- $\tilde{d} < d$: $h(0) = \int_0^{\tilde{d}} (d - z - C)dz = \tilde{d}(d - \tilde{d}/2 - C)$.
  Thus, the bump width, obtained by $h(0) = 0$, is $\tilde{d} = 2(d - C)$. From the requirement $0 < \tilde{d} < d$, this implies $d/2 < C < d$.
- $\tilde{d} > d$: $h(0) = \int_0^{d} (d - z - C)dz - \int_d^{\tilde{d}-d} Cdz = d^2/2 - C\tilde{d}$.
  The bump width is therefore $\tilde{d} = d^2/(2C)$. From the condition $\tilde{d} > d$, we have $C > d/2$.

These analytic solutions are compared to simulations in Figure 1A.

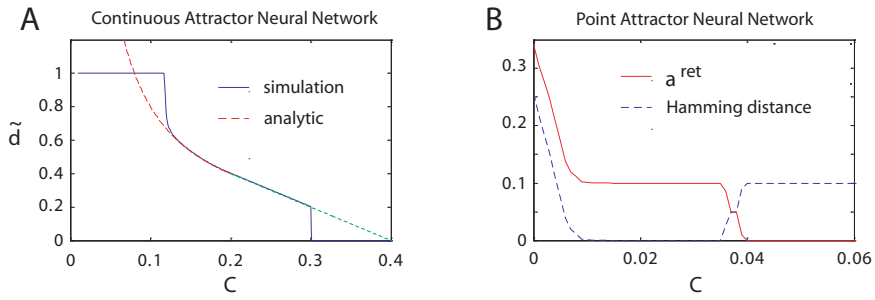

Fig. 1. The effect of inhibition on learning in attractor networks. (A) The solid line represents the simulations of a CANN model and shows the ratio of active nodes after 10 iterations from an initial memory pattern. The results compare well with the analytic results (dashed line) for appropriate values of C. Too small or too large inhibition leads to a loss of the memory states. (B) Average retrieval sparseness, a^{ret}, and Hamming distance between network state and memory pattern for a point attractor network for different inhibition constants, C. Due to the attractor dynamics, a range of inhibition values around the analytic solution can support sparse memory states.

3 Learning in PANNs

Learning sparse representations in the point attractor networks (PANNs) can also be solved with global inhibition. Again, we consider the Hebbian learning rule with global inhibition,

$$w_{i,j} = \frac{1}{\sqrt{M}} \sum_m (\mu_i^m - a)(\mu_j^m - a) - C, \tag{6}$$

where a denotes the sparseness of patterns which is the probability for a node to be active in each pattern (the ratio between the number of active nodes and the total number of nodes). The commonly used network update rule [7] is given by:

$$S_i = \Theta[h_i] = \Theta[\frac{\sqrt{M}}{N} \sum_{j=1}^{N} w_{ij} S_j], \tag{7}$$

so that the activity of the ith node is determined by the sign of the input, h_i. In the limit of many patterns, and under the condition that the network is homogenous, h_i becomes Gaussian distributed with mean and variance given by:

$$< h_i >= -Ca, \quad \sigma^2 =< h_i^2 >= a^3(1-a)^2. \tag{8}$$

Thus, the probability of the ith node to be active is $P(h_i > 0)$ and to be inactive is $P(h_i < 0)$. On the other hand, the probability for the ith node to be active is also equal to a. Thus, under the self-consistent requirement, it must hold that

$$\frac{P(h_i > 0)}{P(h_i < 0)} = \frac{a}{1-a}, \tag{9}$$

from which the relationship between the inhibition, C, and the sparseness, a, can be derived:

$$a = \frac{1}{2} - erf(\frac{Ca}{\sqrt{2}\sigma}). \tag{10}$$

From this condition, we see that when $C = 0$, $a = 0.5$, that is, the retrieved patterns are not sparse. However, patterns with the correct retrieval sparseness $a^{\text{ret}} = a$ can be maintained for a range of inhibition values as shown in Figure 1B.

4 Conclusion

We showed here that sparse attractor networks can be trained with Hebbian learning if inhibition is taken into account, and we calculated how the spareness of retrieved states is related to the inhibition constant.

References

1. Caianello, E.: Outline of thought processes and thinking machines. J. Theor. Biol (1961)
2. Amari, S.: Dynamics of pattern formation in lateral-inhibition type neural fields. Biological Cybernetics **27** (1977) 77–87
3. Hopfield, J.: Neural networks and physical systems with emergent collective emergent computational abilities. Proceeding of the National Academy of Science PNAS **79** (1982) 2554–2558
4. Grossberg, S.: Contour enhancement, short-term memory, and constancies in reverberating neural networks. Studies in Applied Mathematics **52** (1973) 217–257
5. Wilson, H., Cowan, J.: A mathematical theory of the functional dynamics of cortical and thalamic nervous tissue. Kybernetik **13** (1973) 55–80
6. Schöner, G. In: Dynamical Systems Approaches to Cognition. Cambridge University Press (2007)
7. D.J. Amit, H. Gutfreund, H.S.: Storing infinite numbers of patterns in a spin-glass model of neural networks. Physical Review Letters **55** (1985) 1530–1533
8. Zhang, K.: Representation of spatial orientation by the intrinsic dynamics of head-direction cell ensembles: a theory. Journal of Neuroscience **16**(4) (1996) 2112–2126
9. Stringer, S., Trappenberg, T., Rolls, E., Araujo, I.: Self-organising continuous attractor networks and path integration: One-dimensional models of head direction cells. Network: Computation in Neural Systems **13** (2002) 217–242