Analysis of the Effects of Noise on a Model for the Neural Mechanism of Short-Term Active Memory *

J. Devin McAuley and Joseph Stampfli
Department of Computer Science
Depertment of Mathematics
Indiana University
Bloomington, Indiana 47505

June 10, 1993

Abstract

Zipser (1991) showed that the hidden unit activity of a fully-recurrent neural network model, trained on a simple memory task, matched the temporal activity patterns of memory-associated neurons in monkeys performing delayed saccade or delayed match-to-sample tasks. When noise, simulating random fluctuations in neural firing rate, is added to the unit activations of this model, the effect on the memory dynamics is to slow the rate of information loss. In this paper, we show that the dynamics of the iterated sigmoid function, with gain and bias parameters, is qualitatively very similar to the "output" behavior of Zipser's multi-unit model. Analysis of the simpler system provides an explanation for the effect of noise that is missing from the description of the multi-unit model.

*The authors are supported by ONR grant N00014-91-J1261. Please direct all correspondence to mcauley@cs.indiana.edu. Thanks to Sven Anderson, Fred Cummins, Jason Holt, Gary Kidd, Robert Port, Catherine Rogers, James Townsend, and Charles Watson for their constructive comments and respective contributions during the development of this manuscript which has been submitted to *Neural Computation*.

1 Introduction

Over the years, single unit recording studies have consistently suggested that information pertaining to a cued stimulus can be stored temporarily as the tonic activity of a population of neurons (Fuster and Alexander, 1971; Gottlieb, Vaadia, and Abeles, 1989). These firing patterns are similar across modality specific cortical regions, suggesting a general underlying memory mechanism.

Zipser (1991) showed that the activity of hidden units in a fully-recurrent neural network, trained on a very simple memory task, matched the qualitative temporal activity patterns of these memory-associated neurons. He also described a seemingly paradoxical property of this model: simulated random fluctuations in neural firing rate (noise) can slow the rate of information loss.

In the present work, we show that dynamics of a single sigmoid unit mimics the collective response of Zipser's multi-unit model. A mathematical analysis of the one unit model is tractable and provides an explanation for why noise (on average) can improve retention. We believe that this explanation extends to Zipser's multi-unit model.

2 Zipser Model

Zipser's short-term active memory model consists of two linear input units and a number of fully-connected sigmoid units representing short-term memory. The input units, which have connections to all memory nodes, are a binary cue x_c and a real-valued stimulus x_s. The unit equation for all memory nodes is

$$y_i(t+1) = \phi(\sum_j w_{ij} y_j(t) + w_{is} x_s + w_{ic} x_c + \theta_i + X_i(t)) \tag{1}$$

where the sigmoid function is

$$\phi(x) = (1 + e^{-x})^{-1}. \tag{2}$$

The activation of each memory unit i is a weighted sum of the inputs, the activations of all of the memory units, and a bias term θ_i, squashed by the sigmoid shaped function $\phi(x)$. The cue and stimulus inputs and weights are subscripted with c and s, respectively. A Gaussian noise term $X_i(t)$ (with zero mean and standard deviation ν) is added during testing trials only, to simulate random neural excitation.

The training task, shown in Panel A of Figure 1, is to store a cued intensity value in memory for an unspecified duration. During a training sequence, the cue is initialized to 1.0 and a stimulus is selected from the interval $[0, 1]$. The "output" unit of the network is trained using the real-time recurrent learning algorithm of Williams and Zipser (1989) to autoassociate the cued stimulus value for a random number of time steps. The biases θ_i remain fixed at negative values typically in the range $[-1.0, -3.5]$. Between cued stimuli, the cue unit is set to 0.0 and the stimulus unit varies randomly.

The typical response of the network after training is to produce a brief peak to each event of stimulus-plus-cue, to remain at the approximate simulus value, and then to decay slowly (see Panel B of Figure 1). As the interstimulus interval becomes sufficiently large, the activations approach a stable equilibrium, indicating that the network has "forgotten" the stimulus value. The basis for memory in this model is the slow *relaxation* to an attractor following presentation of the stimulus. This is contrasted with Hopfield memory models (Hopfield, 1982) which *use* attractors to store memory items; i.e., input pattern processing converges to a stable activation pattern that is most closely associated with the input.

McAuley, Anderson, and Port (1992) have investigated the behavior of the Zipser model by testing its response on a same-different (roving-level) intensity discrimination task (Durlach and

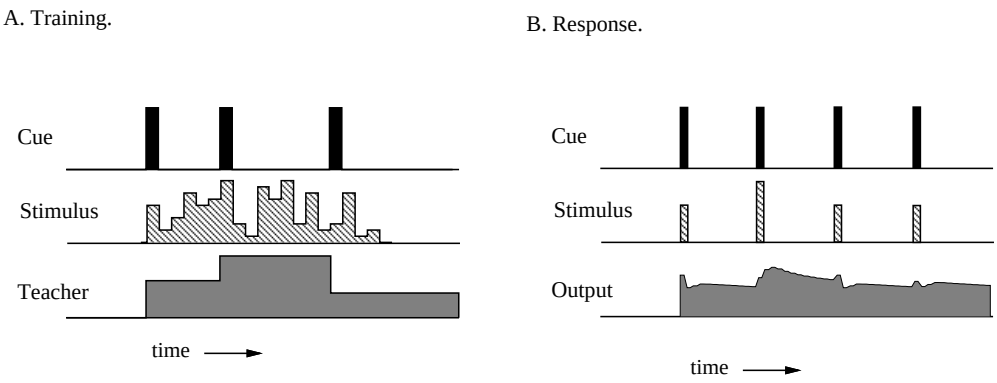

Figure 1: **A.** Cue, stimulus, and teacher values for a hypothetical training sequence. **B.** The output unit peaks in response to each stimulus and cue, and is able to roughly hold the stimulus value by slowly relaxing to an attractor.

Braida, 1969). They replicated Zipser's observation that a noise term $X_i(t)$ added to unit activations during testing trials can improve (on average) the retention of input by slowing the decay to an attractor.

3 Single Unit Model

To better understand this behavior, we study a single unit approximation of this system. Let

$$y(t+1) = \phi(y(t)) = \frac{1}{1+e^{-\gamma(y(t)+\theta)}} \tag{3}$$

where γ is a gain term and θ is a bias term. By initializing $y(0)$ as the "to-be-remembered" stimulus, this equation is a primitive model of short-term memory which we can compare directly to the performance of the Zipser model. Finding equilibria for this system for different values of gain and bias requires numerical techniques such as Newton's method for approximating roots (Atkinson, 1978). However, parameter values for gain and bias which form a boundary between one and two attractor systems can be found explicitly by observing that systems on this boundary have a saddle equilibrium $\overline{x}$ that is tangent to the diagonal line $y(t+1) = y(t)$. At such tangent equilibrium points, $\phi'(\overline{x}) = 1$. This information, combined with the equilibrium equation, can be used to define an expression for the bias as a function of the gain:

$$\theta = -\frac{1 \pm \sqrt{1-4/\gamma}}{2} - \frac{\ln \frac{2}{1\pm\sqrt{1-4/\gamma}} - 1}{\gamma}. \tag{4}$$

This curve in gain-bias space is shown in Figure 2. Points outside the curve configure one attractor models and points inside the curve configure two attractor models. The two attractor systems also have an unstable equilibrium which acts as a threshold. Stimuli $y(0)$ above this threshold converge to an upper attractor while stimuli below this threshold converge to a lower attractor. As the gain increases, the upper and lower attractors approach 1.0 and 0.0 respectively.

4 Comparing Both Models

Figure 3 compares the response of memory, with and without noise, for the Zipser model at time step 10 (top panel) and the single unit model at time step 7 (bottom panel). Stimulus value is

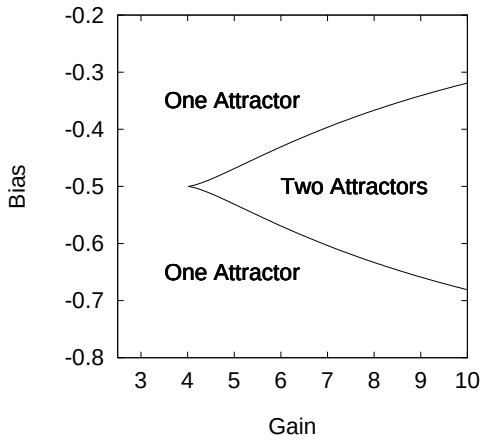

Figure 2: Bifurcations in the dynamics of the single unit sigmoid model as a function of the gain and bias parameters.

plotted along the abscissa and the corresponding memory trace (activation of the output unit at time step n) is plotted along the ordinate. Perfect memory is depicted by the diagonal line (stimulus = trace). Both systems have two attractors. In the single unit case, $\gamma = 6.0$ and $\theta = -0.5$. Below a threshold, memory traces relax towards an attractor located near 0.0. Above this threshold, memory traces relax towards an attractor near 1.0.

To examine the effect of noise, we added a Gaussian random variable $X_i(t)$ with a mean of 0.0 and a standard deviation of 0.05 to the output of each unit i on each time step. Memory response was sampled 20 times for each stimulus value after 10 time steps (7 for the single unit model) and then averaged across trials. Each hatch mark indicates the average effect of noise on memory for a fixed stimulus value after a fixed number of time steps. With noise, both models maintain a *better* approximation of a *range* of stimulus values, than without noise. That is, noisy data points in regions of the stimulus space are closer to lying along the diagonal (perfect memory) than the curves showing memory performance without noise. Moreover, the performance of the single unit model is essentially the same as the Zipser's multi-unit model. Although, the memory traces decay faster in the single unit case. In the next section, we analyze the single unit system to explain the effect of noise.

5 Analysis

The average effect of noise is described here as

$$\phi_{\overline{\nu}}(x) = \frac{1}{n}\sum_{i=1}^{n}\phi(x + X_i(t)) \tag{5}$$

where n is the number of trials that are averaged over and $X_i(t)$ is the noise value on trial i. For this analysis, we assume that $X_i(t)$ is a Bernoulli probability function on the discrete set $\{-\nu, \nu\}$; that is, $X_i(t)$ is either ν or $-\nu$ with probability 0.5. There are two cases to consider:

$$\phi_{\nu}(x) = \phi(x + \nu) \tag{6}$$

and

$$\phi_{-\nu}(x) = \phi(x - \nu). \tag{7}$$

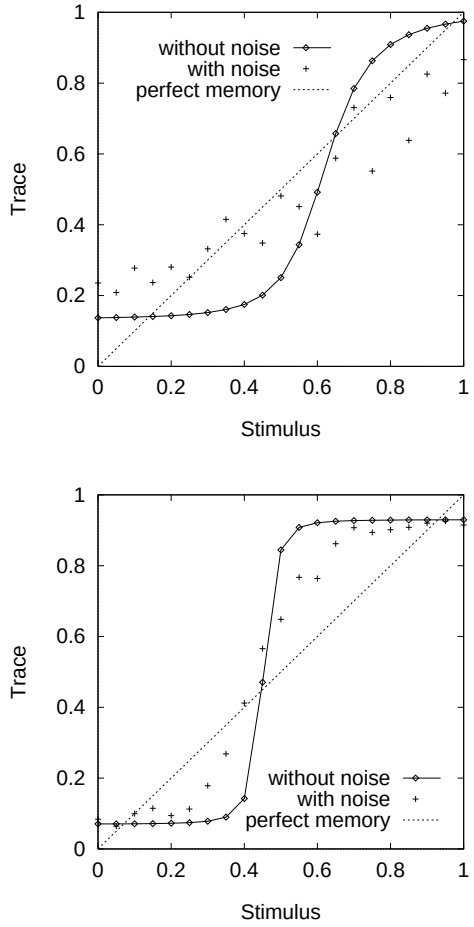

Figure 3: Memory response of the Zipser model at time step 10 (top) compared with the memory response of the single unit sigmoid model at time step 7 (bottom). Each hatch mark shows the noisy response to a fixed stimulus. For comparison, the diagonal line (stimulus = trace) shows perfect memory.

If n is sufficiently large, the number of values $\phi(x+\nu)$ will be approximately equal to the number of values $\phi(x-\nu)$. Consequently, equation (5) can be simplified to

$$\phi_{\overline{\nu}}(x) = \frac{\phi_{\nu}(x) + \phi_{-\nu}(x)}{2}. \tag{8}$$

Suppose that the single unit model $\phi(x)$ is a linear function, then by the principle of superposition averaging will exactly cancel the effect of noise.

$$\phi_{\overline{\nu}}(x) = \phi(x) \tag{9}$$

However, for the single unit sigmoid model, the principles of linearity do not apply and consequently, the effect of noise is not necessarily cancelled by averaging. A summary of all six cases is provided in Table 1. In this table, we let

$$\Delta = \phi(x) - x \tag{10}$$

and

$$\Omega = \phi_{\overline{\nu}}(x) - \phi(x) \tag{11}$$

Case	Δ	Ω	Rate of Information Loss
1	+	+	faster
2	+	−	slower
3	−	+	slower
4	−	−	faster
5	0	+	same
6	0	−	same

Table 1: The nonlinear effect of noise on iterations of the single unit sigmoid model.

where x is the stimulus and $\phi(x)$ is its memory trace after one iteration. The sign of Δ indicates whether successive iterations of $\phi(x)$ are moving towards an attractor that is above or below the initial stimulus x. Positive Δ implies that iterations of $\phi(x)$ are converging towards an attractor that has a value larger than the stimulus. Negative Δ implies the opposite. The sign of Ω indicates the direction noise (on average) pushes iterations of $\phi(x)$ Positive Ω increases $\phi(x)$. Negative Ω decreases $\phi(x)$.

For each case, we show the average effect of noise on the rate of information loss (memory retention). If Δ and Ω have the same sign then noise degrades the memory trace of stimulus x because it relaxes the system towards an attractor at a faster rate than without noise (cases 1 and 4). If Δ and Ω have opposite signs then noise sustains the memory trace by slowing down the relaxation rate (cases 2 and 3). When $x = -\theta$, it can be shown that $\Omega = 0$ (cases 5 and 6); this is the point at which Ω switches sign. The Δ term changes sign at stable and unstable equilibria. Above and below attractors, Δ is negative and positive respectively. The opposite is true for unstable equilibria.

A point in gain-bias space fixes the number and location of equilibria and hence determines the stimulus ranges for which noise (on average) will improve retention. The six different configurations are enumerated in panel A of Figure 4. For one attractor models (cases 1,2 and 3), noise improves retention of stimuli that are between this attractor and $x = -\theta$. For two attractor models (cases 4, 5, and 6), noise improves retention for stimuli between these attractors, except for the stimulus region bounded by the unstable equilibria and $x = -\theta$.

In panel B, we have fixed the gain and bias of the single unit model at 6 and -0.5 respectively. This model is an instance of case 6, but serves to summarize cases 4 and 5 as well. If we choose to "load" a stimulus value of 0.6 into the memory of this model, then noise should improve retention of this stimulus because the value is between the two attractors. Panel B compares 10 iterations of the functions $\phi(x)$ and $\phi_{\overline{\nu}}(x)$ for stimulus (initial x) $= 0.6$ and $\nu = 0.15$. As expected, the model with noise (dotted line) converges to the upper attractor at a slower rate than the model without noise (solid line).

In panel C, we illustrate the opposite effect. The gain and the bias are fixed at 3.8 and -0.5 respectively. This model is an instance of case 1, but also illustrates the properties of cases 2 and 3. In contrast to panel B, noise added to this model after loading a stimulus value of 0.6 should speed up the trace decay. Moreover, as an instance of case 1 models, noise hurts the retention of values in the entire stimulus range. Panel C compares 10 iterations of $\phi(x)$ and $\phi_{\overline{\nu}}(x)$ for stimulus $= 0.8$ and $\nu = 0.15$. As expected, the model with noise (dotted line) converges to the attractor at a faster rate than the model without noise (solid line).

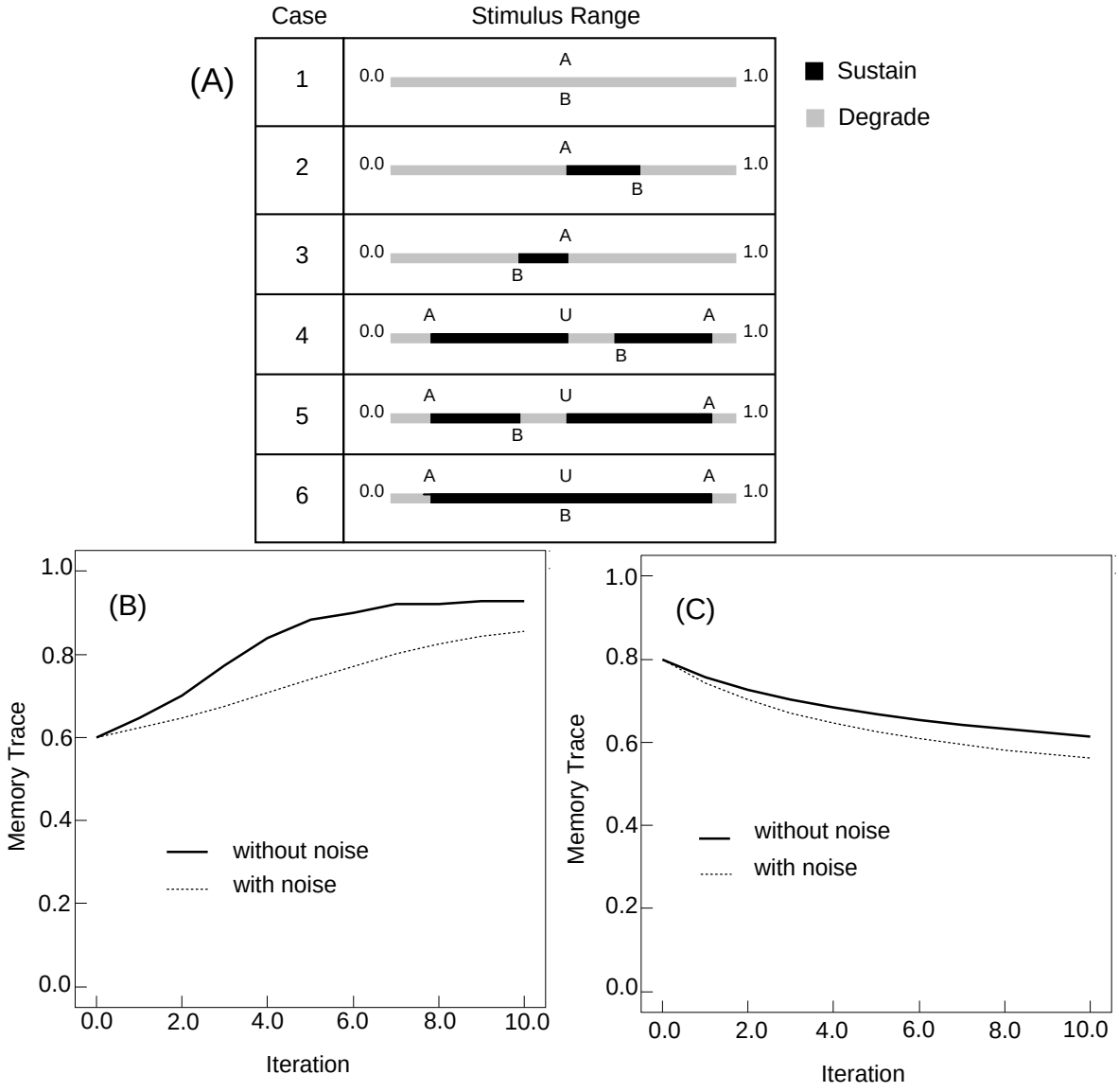

Figure 4: (A) Enumeration of the stimulus intervals for which noise slows the rate of information loss, as a function of memory dynamics: **A** indicates an attractor, **U** indicates an unstable equilibrium, and **B** indicates the point $x = -\theta$. Noise improves retention for stimuli within the dark shaded regions. The model in (B) is an instance of case 6. It compares memory performance with and without noise for a stimulus of 0.6. The model in (C) is an instance of case 1. It compares memory performance with and without noise for a stimulus of 0.8.

6 Conclusions

We have shown that a single sigmoid unit approximates the collective "output" behavior of a many-unit fully-recurrent network for short-term active memory (Zipser, 1991). The stimulus regions for which noise slows the rate of information loss has been shown to vary predictably as a function of the gain and bias parameters. Thus, the surprising effect of noise has a rather straightforward explanation in the nonlinear dynamics of the sigmoid function. For one attractor models (low gain), the stimulus region for which noise improves retention is small or nonexistent (as in case 1). For two attractor models (high gain), the stimulus region for which noise improves retention is much larger (between the two attractors) and continues to increase in size with further increases in gain. However, the memory of large-gain models is inherently poor. This suggests that an "optimal" balance of having good inherent retention and sizable stimulus regions which are helped by noise may lie near bifurcation boundaries between one attractor and two attractor models. We have yet to quantify a measure of "how much" noise improves retention. In general, this research suggests one way that the nervous system may take advantage of noise inherent in the system, rather than be hindered by it as generally assumed, to better represent and process information.

References

Atkinson, K. (1978). *An Introduction to Numerical Analysis*. Wiley and Sons, New York.

Durlach, N. and Braida, L. (1969). Intensity perception. I. preliminary theory of intensity resolution. *Journal of the Acoustical Society of America*, 46(2):372–383.

Fuster, J. and Alexander, G. (1971). Neuron activity related to short-term memory. *Science*, 173:652–654.

Gottlieb, Y., Vaadia, E., and Abeles, M. (1989). Single unit activity in the auditory cortex of a monkey performing a short term memory task. *Experimental Brain Research*, 74:139–148.

Hopfield, J. J. (1982). Neural networks and physical systems with emergent collective computational abilities. In *Proceedings of the National Academy of Sciences*, volume 79, pages 2554–2558. National Academy of Sciences.

McAuley, J. D., Anderson, S., and Port, R. (1992). Sensory discrimination in a short-term trace memory. In *Proceedings of the Fourteenth Annual Meeting of the Cognitive Science Society*, pages 136–140, Hillsdale, NJ. L. Erlbaum Assoc.

Williams, R. and Zipser, D. (1989). A learning algorithm for continually running fully recurrent neural networks. *Neural Computation*, 1(2):270–280.

Zipser, D. (1991). Recurrent network model of the neural mechanism of short-term active memory. *Neural Computation*, 3:179–193.