

A Hippocampal Model Predicts a Fluctuating Phase Transition when Learning Certain Trace Conditioning Paradigms

Andrew G. Howe

William B Levy, Ph.D

University of Virginia
Department of Neurosurgery
P.O. Box 800420
Charlottesville, VA 22908, USA

Address all correspondence to : Dr. William B Levy
University of Virginia
Department of Neurosurgery
P.O. Box 800420
Charlottesville, VA 22908, USA
Phone : (434) 924-9996
Fax : (434) 982-3829
E-mail : wbl@virginia.edu

Abstract

The hippocampus is needed for at least one kind of trace classical conditioning, the air-puff eye-blink paradigm. A simple model of region CA3 predicts three basic, quantitative observations of the learning behavior of rabbits. One particular quantified prediction is the learnable trace interval. The boundary region of the reliably learnable trace interval represents a phase transition. Within this transition, three behaviorally distinguishable modes are expressed : failure to blink; blink too soon; and occasionally, appropriate predictive blinking. In the region of the phase transition, there is a small sub-interval where the behavioral modes fluctuate rapidly from trial to trial for individual simulations. Such observed fluctuations are an experimental prediction by the model. The discussion also includes a brief conjecture concerning the underlying cause of the phase transition and the fluctuations.

Keywords : CA3, hippocampus, classical conditioning, trace interval

Note About Figures In This File

The figures presented here are in JPEG format, which is not ideal for clarity. Every effort has been made to present the data clearly in the JPEG format used here. We have high quality figures available in extended post script (EPS) format for Figures 3,4,6,7,8 and 9. We will gladly provide them upon request.

Introduction

The hippocampus plays a critical role in storing episodic memories (Scoville & Milner 1957). Somewhat unexpectedly, it also plays the same role in the acquisition and storage of the trace interval in at least one form of trace classical conditioning (TCC) (Solomon et al. 1986). Moreover, for both forms of memory storage, there is the identical observation : normal hippocampal function is needed for learning and initial storage but eventually the hippocampus is no longer required; i.e., long-term storage can be demonstrated after late, but not after early, hippocampal removal (Kim et al. 1995).

For this reason and because this century-old paradigm is so cognitively different from other paradigms used to characterize hippocampal function, e.g., the water maze (Morris 1984), transitive inference (TI) (Dusek & Eichenbaum 1997), and transverse patterning (TP) (Alvarado & Rudy, 1995), we insist that a proper model of hippocampal function must also reproduce the trace conditioning dependency.

A particularly well-documented form of trace conditioning was developed by Gormezano et al. (1983) and used by Solomon et al. (1986). The paradigm consists of two non-overlapping stimuli with a specified amount of stimulus-free time between them (see Figure 1, top panel). Successful learning is defined as the timely delivery of a blink just prior to the onset of an unconditioned stimulus (US). The US follows at a specified time after the offset of a conditioned stimulus (CS). The stimulus-free time between the CS offset and the US onset is called the trace interval. This temporal separation, the "trace interval", is perhaps the most direct embodiment of our hippocampal theory. The paradigm requires the subject to predict the US onset based on the CS, and as such, this paradigm falls directly into the class of problems suited to our computational theory of the hippocampus (Levy 89, 90, 96) as a multi-sensory, sequence encoding and predicting system. That is, the dentate gyrus (DG) and CA3 work together as both a random recoder and a sequence learner/sequence predictor while CA1, the subiculum, and deep layers of entorhinal cortex (EC) map the CA3-discovered associations

back into neocortical/forebrain-based memory. Figure 1 schematically depicts trace conditioning and its representation in the simulated CA3.

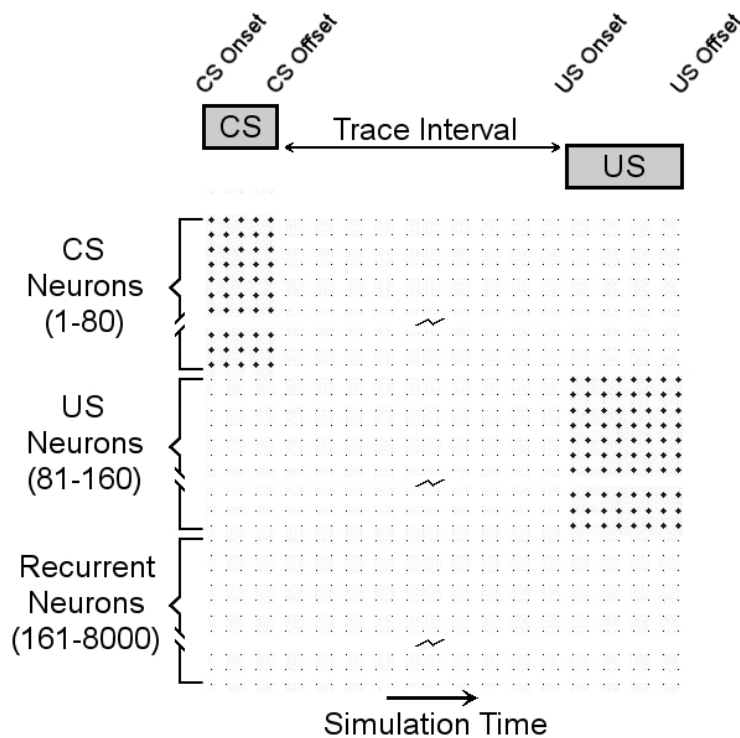


Figure 1 : *Top Panel* - A schematic representation of the trace classical conditioning paradigm. For any training trial, a conditioned stimulus (CS) is first presented for a short, but perceptually significant time. At the offset of the CS, a time period elapses (the Trace Interval), and then via another sensory modality, the unconditioned stimulus (US) is activated for a brief period of time. In order to escape or ameliorate the US (an air puff in Solomon et al. 1986), the hippocampus must provide a prediction of the US slightly before the onset of the US.

Bottom Panel – The trace paradigm is reproduced in the simulation by external forced firing of particular neurons. Neurons 1-80 are turned on as the neural representation of the CS while neurons 81–160 are turned on to represent the US. Recurrent connectivity in the simulated network is random, so activation of sequential neurons is equivalent to a choice of two random but non-overlapping sets of neurons. Time, including the longevity

of the CS, of the US and of the trace interval, is parameterized relative to the representation of a 100 ms off-rate time constant from the NMDA receptor (see α in methods). Simulations presented here have 20 ms resolution.

Here we explore the dynamics of the learned cell-firing modes as a function of trace interval with a computational model of the hippocampus. Our goal is to make predictions in a domain that exceeds intuitive, psychological conceptualizations. Specifically, simulations of the model predict phase transition-like behaviors just beyond the edge of the reliably learnable trace interval. This cell-firing phase transition implies a behavioral phase transition for the learned blink response.

Methods – Simulating the Hippocampus CA3 Region

The Model

The hippocampal complex and its related structures are involved in encoding sequences of cortical patterns (for references, see Levy 1989; Levy 1994). At the heart of this recoding (Levy 1989) is the combination of a CA3 region with sparse, recurrent connections and a temporally asymmetric, associative synaptic modification rule (Levy & Steward 1983). These microscopic properties lead to the major function of the CA3 region : context-dependent sequence learning (Levy 1989, Levy 1996).

The model balances biological accuracy and simplicity. Neurons are simple McCulloch-Pitts devices with binary output of one or zero on each computational cycle (see equation 2).

Postsynaptic summation and axonal impulse transmission occur in one computational cycle. Connections between primary (pyramidal) neurons are excitatory, sparse and random (ten percent connectivity). No self-connections exist in the simulations presented, although when self-connections that are both random and sparse exist, the results are unchanged. The input layer merges the entorhinal cortex (EC) and dentate gyrus (DG) inputs into a single input class, as if convergence of excitation occurs between these two anatomically distinct afferent systems. In particular, activation of an input, $x_j(t) = I$, is guaranteed to fire its unique postsynaptic target, $z_j(t) = I$. Figure 2 illustrates the model schematically.

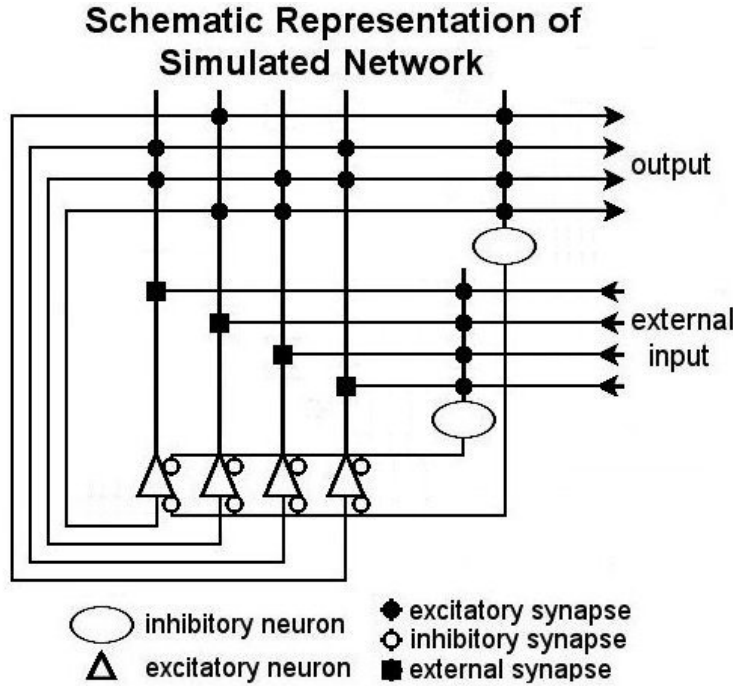


Figure 2 : A schematic representation of our hippocampal model. Excitatory neurons are recurrently connected with excitatory synapses (filled circles). Each neuron receives an external input (filled squares), which forces the cell to fire. Two inhibitory neurons help to maintain neural activity. Inhibitory neurons with external synapses are not force fired; instead the external synapses are summed and output is proportional to this sum. The interneuron input is divisive (see Equation 1), so inhibitory synapses (unfilled circles) appear on the cell body of the excitatory neurons.

In the absence of external excitation, the internal excitation of a neuron, $y_j(t)$, is given by equation 1 :

$$y_j(t) = \frac{\sum_{i=1}^n z_i(t-1)c_{ij}w_{ij}(t-1)}{\sum_{i=1}^n z_i(t-1)c_{ij}w_{ij}(t-1) + K_{fb} \sum_{i=1}^n w_{iI}(t-1)z_i(t-1) + K_{ff} \sum_{i=1}^n x_i(t) + K_0} \quad (1)$$

The sum $\sum_{i=1}^n z_i(t-1)c_{ij}w_{ij}(t-1)$ represents the synaptic excitation for neuron j at time t . The term w_{ij} represents the weight value from neuron i to neuron j , and all synapses were initially set to 0.5. The zero-one binary variable c_{ij} makes explicit the presence or absence of a connection to neuron i from neuron j . Thus, a weight that is initially zero is always zero.

Activity within the simulation is approximately controlled with divisive inhibition, as seen in the denominator of the $y_j(t)$ equation. Actual activity is chaotic and only approximately controlled (Smith et. al. 2000). The values for K_{ff} , K_{fb} and K_0 are set such that random input to a simulation produces activity near the desired activity with oscillations that approximate the minimum obtainable. The constants K_{fb} and K_{ff} scale the influence of the feedback and feedforward inhibitory neurons, respectively. The feedforward interneuron contribution to the denominator of $y_j(t)$ is $K_{ff} \sum_{i=1}^n x_i(t)$, where $x_j(t)$ represents the forced firings injected from the input

layer. The feedback interneuron contribution to the denominator of $y_j(t)$ is $K_{fb} \sum_{i=1}^n w_{iI}(t-1)z_i(t-1)$. The positive weight value applied to the input from an excitatory neuron i to the interneuron I is indicated with w_{iI} (details in Sullivan and Levy 2003). The variable $z_j(t-1)$ is the firing state of neuron i on the last time step. Neuronal output, z_j , is calculated with the following equation :

$$z_j = \begin{cases} 1 & \text{if } y_j \geq \theta, \text{ or } x_j = 1 \\ 0 & \text{otherwise} \end{cases} \quad (2)$$

where θ is a constant threshold value. Throughout this paper, neurons that output a '1' are referred to as “active neurons” or neurons that “fired” and variations on those two terms.

Desired activity is specified as a percentage of active neurons per simulation time step. In order to maintain control of activity as the weights, w_{ij} , change, the pyramidal-interneuron synaptic weights, w_{iI} , on the feedback interneuron are adjusted according to the following rule :

$$w_{iI}(t) = w_{iI}(t-1) + \lambda z_i(t-1) \left[\frac{\sum_{k=1}^n z_k(t-1)}{n} - a \right] \quad (3)$$

where λ is the modification rate constant for pyramidal-interneuron synapses and a is the desired fractional activity. Feedforward pyramidal-interneuron synapses are not adjusted.

Excitatory neuron synapses, w_{ij} , are adjusted according to a biologically inspired post-synaptic associative modification rule with potentiation and depression (Levy & Steward, 1979) and time staggering between pre- and post-synaptic activity (Levy & Steward, 1983).

In order to make simulations more biological, and to produce more behaviorally meaningful predictions that have temporal relevance at the level of behavior, the trace interval spans several simulation time steps (Levy & Sederburg 1997; Rodriguez & Levy 2001). To produce temporal relevance, the model includes a time-dependent NMDA receptor (NMDA-R) off-rate time constant (August & Levy 1999; Levy and Sederberg 1997; Wu & Levy 2005) and assumes that the glutamate-like priming of this receptor decays exponentially (Rodriguez & Levy 2001). Because the NMDA-R off-rate time constant is approximately equal to the duration of the CS (100 ms) in the behavioral experiment, real time can be mapped into the simulation. Specifically, in the simulation, five computational cycles correspond to the CS duration and thus five successive time steps must produce an e-fold decay of the NMDA-R, thereby mapping one computational cycle to 20 ms of real time. The longevity of the CS, US, and the trace interval are thus set relative to each other and these behaviorally relevant events must relate to the NMDA-R off-rate constant (α , see Table 1). In sum, behavioral predictions based on the model gain temporal relevance (e.g. the model has been used to predict the effect of CS longevity, see Wu & Levy 2005). NMDA-R glutamate priming is represented by $\bar{z}_i(t)$, and it is calculated with the following equation :

$$\bar{z}_j(t) = \begin{cases} 1 & \text{if } z_j(t) = 1 \\ \alpha \bar{z}_j(t-1) & \text{otherwise} \end{cases} \quad (4)$$

where α represents the decay time constant of the NMDA receptor (NMDA-R).

The pyramidal-pyramidal synaptic modification rule takes the form :

$$w_{ij}(t+1) = w_{ij}(t) + \mu z_j(t)(\bar{z}_i(t-1) - w_{ij}(t)) \quad (5)$$

i is the input neuron, j is the output neuron, and μ is the synaptic modification rate constant.

The variables in the equations are set to the values listed in Table 1. These microscopic variables express their relevance and importance through mesoscopic variables (see e.g. August and Levy 1999; Smith et al 2000) such as average activity. Many variables of the model have been systematically investigated in Levy et al (2005), and the results of this earlier study shaped the values used in the work presented here. Based on the findings of the above studies, each neuron in the simulated network is randomly and fractionally connected. The synaptic modification rule is also critical (Minai & Levy 1993).

Simulation Parameters		
<i>name</i>	<i>symbol</i>	<i>value</i>
number of neurons	n	8000
desired fractional activity	a	0.05
fractional connectivity	c	0.10
threshold	θ	0.50
feedforward inhibitory constant	K_{ff}	0.0180
feedback inhibitory constant	K_{fb}	0.0512
resting shunting inhibition	K_0	1.0580
NMDA-R off-rate decay constant	α	$e^{(-1/5)}$
excitatory synaptic modification constant	μ	0.01
inhibitory synaptic modification constant	λ	0.50

Table 1 : Parameters for the simulations.

Randomization

There are two forms of randomization present in these simulations :

- 1) the connectivity between neurons
- 2) the initial activity vector prior to the first time step of each trial

Based on the randomizations, three types of comparisons appear in the results section :

- 1) different connectivities, fixed sequence of initial activity vectors
- 2) fixed random connectivity, different sequences of initial activity vectors
- 3) fixed connectivity, fixed sequence of initial activity vectors, different trace intervals

Each simulation of the model uses a pseudo-randomly generated connectivity matrix with fixed fan-in. At the beginning of each simulated training trial, a random, binary-valued vector representing the state of neuronal activity prior to the first simulation time step is generated. These are the only two sources of randomness in the simulations described here. Results for simulations discussed here only differ in randomness across one of these dimensions, not both simultaneously.

Training & Testing

A training trial is composed of a sequence of inputs as visually depicted in Figure 1. Time in the simulation is scaled relative to the NMDA receptor off-rate constant, α . The duration of the CS was 100 ms and the duration of the US was 160 ms (as in Rodriguez & Levy 2001; McEchron &

Disterhoft 1997). First, neurons 1-80 are externally activated for 5 time steps to represent the CS. Then, a period without external activation represents the trace interval. Finally, neurons 81-160 are activated for 8 time steps to represent the US. Figure 1 illustrates the paradigm and its representation in the simulation. Sequential neurons are activated as a convenience for analysis of learned encoding and US prediction. The random connectivity of the matrix negates any possible learning advantage the simulation could realize from activation of sequential neurons. During training, synaptic modification occurs for pyramidal-pyramidal synapses and pyramidal-interneuron synapses (equations 3 & 5).

A test trial presents only the CS (no external US activation) and the synaptic modification rules (equations 3 & 5) are deactivated. Trace intervals from 200-1400 ms (inclusive) were examined because hippocampally lesioned rabbits can learn trace intervals as brief as 300 ms (Moyer et al. 1990) and it is known that learning performance in the model is poor from 0-200 ms (Rodriguez & Levy 2001; Levy et al. 2005). The trace interval was increased in increments of 20 ms.

Behavioral Decoding

It is important to recall that the hippocampus does not directly receive sensory input, nor does it directly drive a conditioned response. Consistent with these facts, the explicit hypothesis (e.g. Levy 1989, 1990, 1996) is that the hippocampus produces a prediction of US onset which is based on the CS and its offset. In order for such a prediction to be useful for the organism, the US prediction must precede the US onset so that a timely response can be delivered (defined by Solomon et al. 1986). For a simulation attempting a learnable interval, neurons that receive the US input become active shortly before the presentation of the US itself after a sufficient number training trials. For our present purposes, successful prediction is defined as at least 30% of US neurons active on at least one simulation cycle in the 60-200 ms of simulation time of time before the US onset. An early blink in animals is defined as at least 30% of US neurons active on at least one time step 200 ms prior to the US onset for simulations. We call this mode “predict too soon”. When US neuron activity for both these periods is below 30% for every simulation time step, we consider this equivalent to the animal not blinking before US onset. It is called a “failure to predict.” Actual activity and timing of activity for US neurons is fairly continuous, so the strict criteria introduce some arbitrariness into the behavioral evaluation.

Visualization

Raster diagrams (Figures 4 & 7) visualize the activity of individual neurons across a single trial. Each point represents the output of one neuron on one time step and active neurons are displayed with heavy points. Neuron index is plotted on the y-axis and simulation time step is plotted on the x-axis. The raster diagrams in this paper display three distinctly different groups of neurons. Neurons 1-80 receive the CS and are activated by external forced firing during testing and training. Neurons 81-160 receive the US and are activated by external forced firing during training alone. Neurons 81-160 are particularly interesting to us because these are the neurons which will provide the prediction. Neurons 161-8000 are activated through recurrent connections alone, and they are responsible for forming the code that bridges the trace interval (Levy et al. 2005). This enormous number of neurons receives a special visualization technique (see Levy et al. 2005) to facilitate interpretation of their activity.

First, we reorder these neurons based on the time step for which each one is first active. Neurons that do not fire at all appear at the end of the list. That is, the earlier a neuron fires in simulation time, the smaller its reordered index number. From this reordering, we sample 80

representative neurons and plot these as neurons 161-240 on the raster diagram. This means that the reordered neurons represent a scaled down snapshot of the entire simulation. For example, the proportion of neurons which never become active are accurately represented by the proportion of neurons which do not fire in the reordered sample. To compare firing pattern growth across trials, we select a single trial as the basis for reordering across all trials (as in Figure 7). Diagrams in Figure 4 have been reordered on the trial depicted, while diagrams in Figure 7 have been reordered on trial 188.

The US neurons (81-160) are analyzed for prediction of the US on the assumption that neurons receiving the US will become active in time to predict the US onset for other brain regions. A pair of small vertical lines on the raster diagrams indicates each criterion time. Timing criteria were set arbitrarily. The leftmost pair indicates the earliest acceptable successful prediction (200 ms before US onset). The middle pair indicate the failure boundary for prediction (60 ms before US onset) – activation on or after this time point is too late for successful avoidance of the US. The rightmost pair indicate the US onset.

Contour diagrams (Figure 6) illustrate average US neuronal activity as a function of trace interval (y-axis) and simulation time step within a trial (x-axis). The US onset is fixed on the x-axis, to facilitate easy interpretation of simulation behavior. US onset is marked with a vertical line. The leftmost diagonal line represents CS onset. CS offset is indicated by the diagonal line to the right. Each row is the average activity of the US neurons for 10 different simulations.

Average US neuron activity diagrams (Figure 3, Figure 8 – bottom row) show the average US neuron activity of one particular simulation across trials and within-trial time. Three lines exist on the figures to mark important time steps. The leftmost line indicates the earliest acceptable successful prediction (200 ms before US onset). The middle line indicates the failure boundary for prediction (60 ms before US onset). The right most line indicates the US onset.

Results

Learning implies a change in the input and output behavior of a system as a function of experience. In the case here, the quantitative experience is the number of paired CS-US training trials. After enough training, rabbits suddenly and accurately, begin to predict the US onset (McEcherson & Disterhoft 1997). Rodriguez & Levy (2001) noted that simulations reproduce the abruptness of the training-induced behavioral change; a representative simulation rapidly transitions from failing to predict US onset to robust and timely prediction of this onset. Moreover, a prior investigation of CS and US longevity noted a phase transition-like behavior in predictive mode dependent on trace-interval length (Wu & Levy 2005). The hippocampal model also reproduces the two behavioral failure modes, and the learnable versus unlearnable trace intervals (Levy et al. 2005 – Figure 5).

Detailed simulations (Levy et al. 2005) show that a single metaphor explains the abruptness of the training-induced behavioral change - an appropriate US prediction only arises when a bridging sequence of reliable neuron firing connects the CS coding neurons to the US coding neurons. This bridge of connected subsets of neurons creates a sequence through state space. The abruptness in learning occurs because the bridge is built gradually from both ends, a span at a time. Until the last span is in place, no US prediction will occur – i.e. no traffic flows across the bridge. An example of the abrupt onset of CS-induced US firing is illustrated in Figure 3, via test trial firing data collected after each training trial. For 95 trials, no US neurons are CS activated. Then over the next three to five trials, >30% of the US neurons start firing.

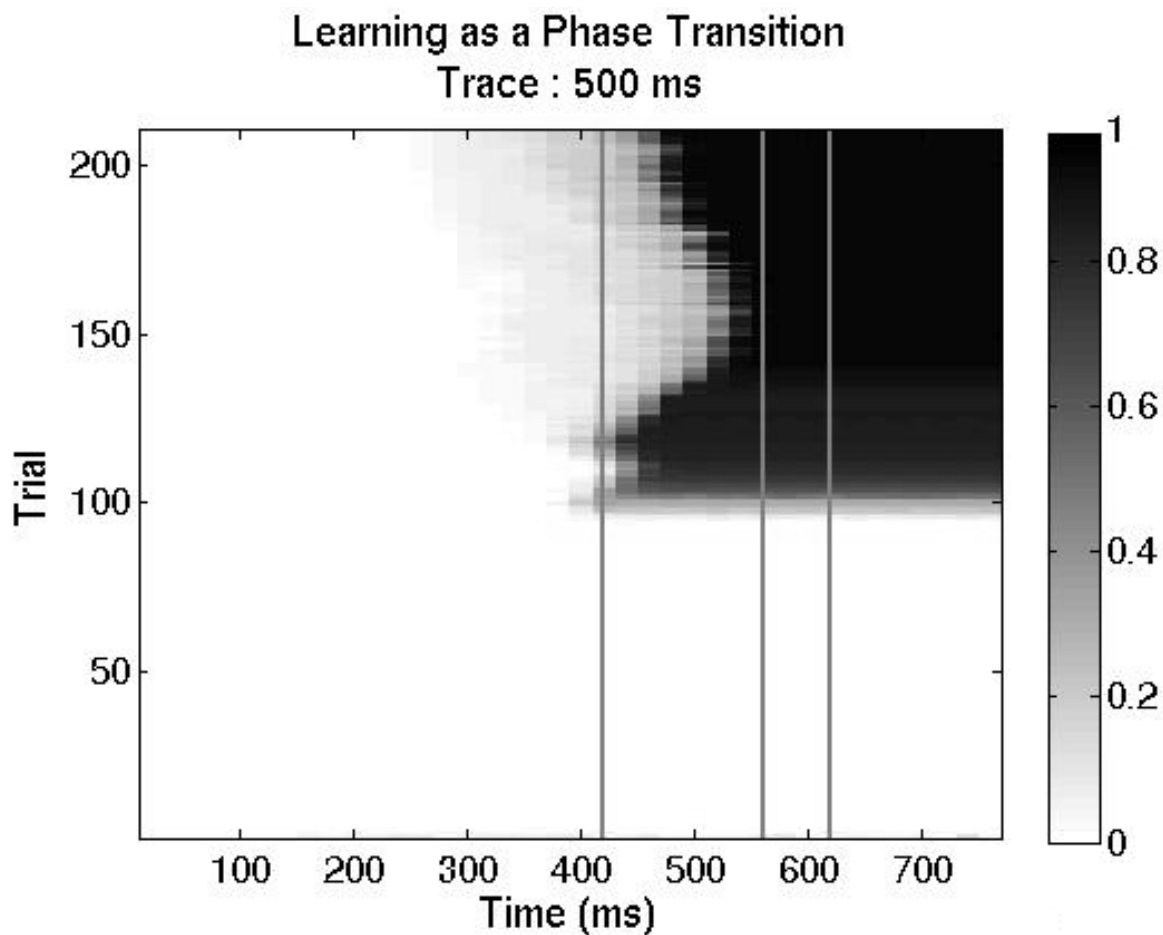


Figure 3 : Predicting the US-onset is an abrupt change in behavior. No US neurons predict the US for 95 trials. Then, within 5 additional trials, more than 30% of US neurons produce a timely prediction. The right line marks the time of US onset; the middle line the first time of unacceptably late prediction; the left line the earliest time for successful prediction. Average US neuron activity across trials (y-axis) and simulation time (x-axis) is shown.

Here we are interested in documenting the effect of the trace interval on this learning-based phase transition. Specifically, we illustrate another kind of phase transition fully dependent on the first phase transition, or from another viewpoint, a parameter that provides two more types of learned phase transitions. As the trace interval lengthens, two new outcomes appear, only one of which might be stable in the face of extensive training. Such phases were seen in Levy et al. (2005) and in Rodriguez and Levy (2001) but not studied in any detail. In the region of trace intervals that first yield these new phases, there is noticeable instability. To develop this result, we begin with Figure 4, which illustrates three prototypical, learned cell-firing modes.

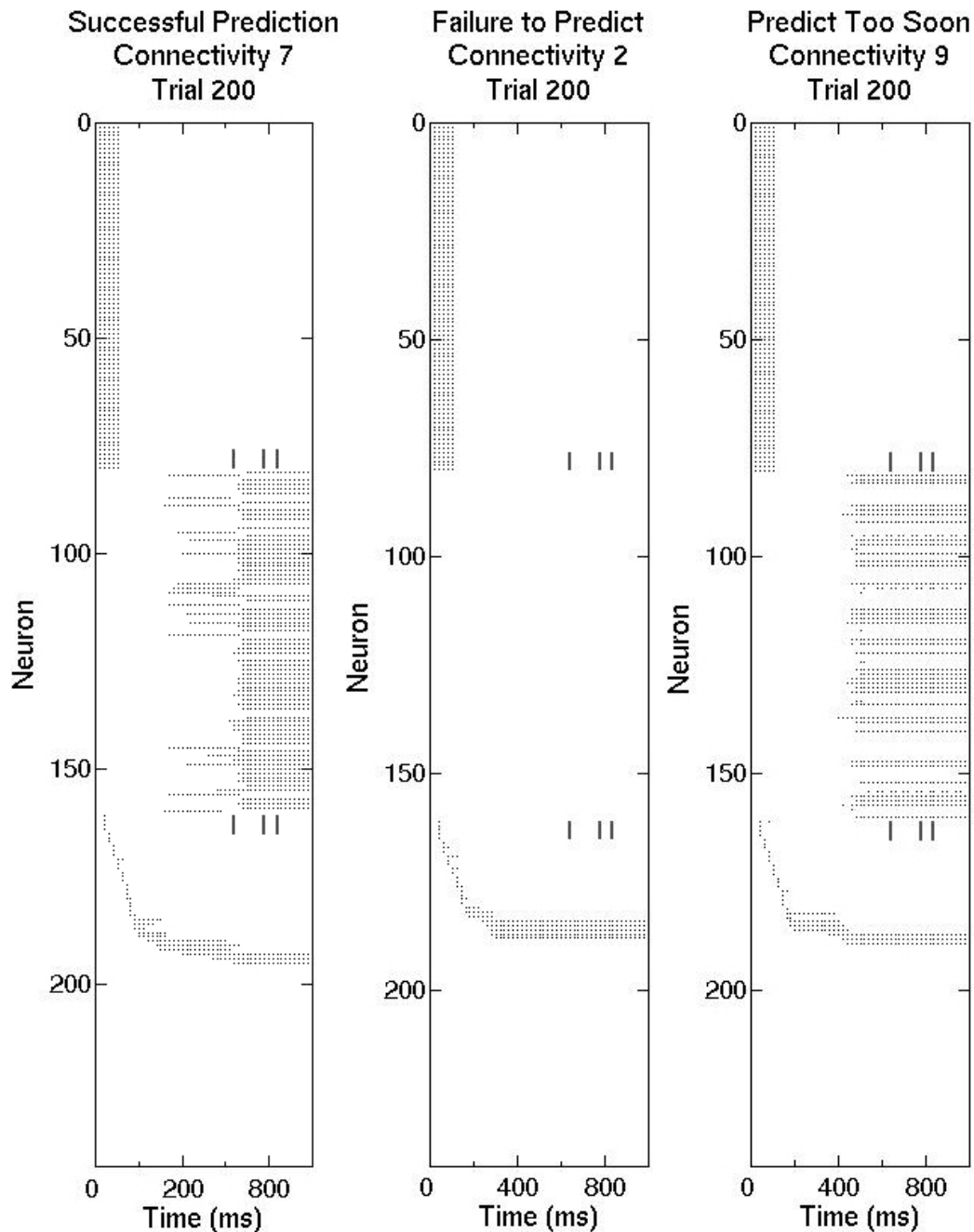


Figure 4 : The 720 ms trace interval produces three different learned firing modes. The modes are named “successful prediction” (left), “failure to predict” (middle) and “predict too soon” (right). The leftmost raster diagram depicts timely, predictive activation of US neurons. The middle raster diagram (connectivity 2) illustrates a failure-to-predict behavior where the system does not predict the US at all, despite the existence of a sequence of recurrent cell firings that span the test trial. The right-most raster diagram illustrates a predict-too-soon response, where many US neurons are active too early before the US onset. Each panel presents an individual simulation with different neuron-to-neuron connectivity. Vertical hash marks delimit the criterion response regions (see Methods for details). Neurons 161-240 on the diagram are reordered and sampled so as to represent the temporal firing patterns of the 7840 neurons which do not directly receive external input (see Methods for details). The data were collected from the test trials following the two hundredth training trial.

Three Learned Behavior Modes

The behavioral modes are defined by the firing of US neurons (see Methods for details). The leftmost panel of Figure 4 illustrates successful prediction; specifically a majority of the US neurons become active at an appropriate time before US onset and a few US neurons are active too soon. The example in Figure 4 barely meets the success criterion because several neurons activate long before US onset. However, not enough neurons activate to trigger a criterion response.

The second panel in Figure 4 illustrates the failure-to-predict mode. In this example, no US neurons are active, and as a result, the simulation does not produce any prediction of the US onset. Despite the fact that no US neurons fire, a sequence of recurrent neuron firing forms a bridge across the trace interval.

The third firing mode, premature US onset prediction or predict-too-soon, is illustrated in the rightmost panel of Figure 4. More than 30% of the US neurons become active before the earliest acceptable prediction time. Such a response is analogous to the short-latency, non-adaptive response from a rabbit reported by Solomon et al. (1986).

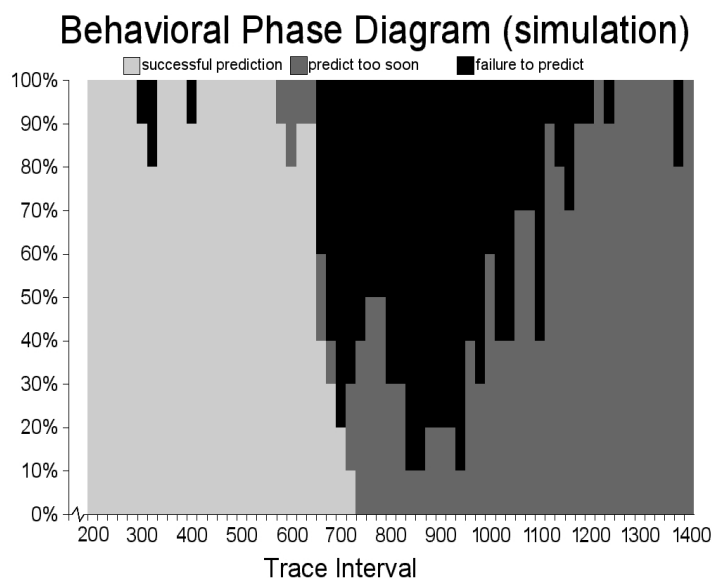
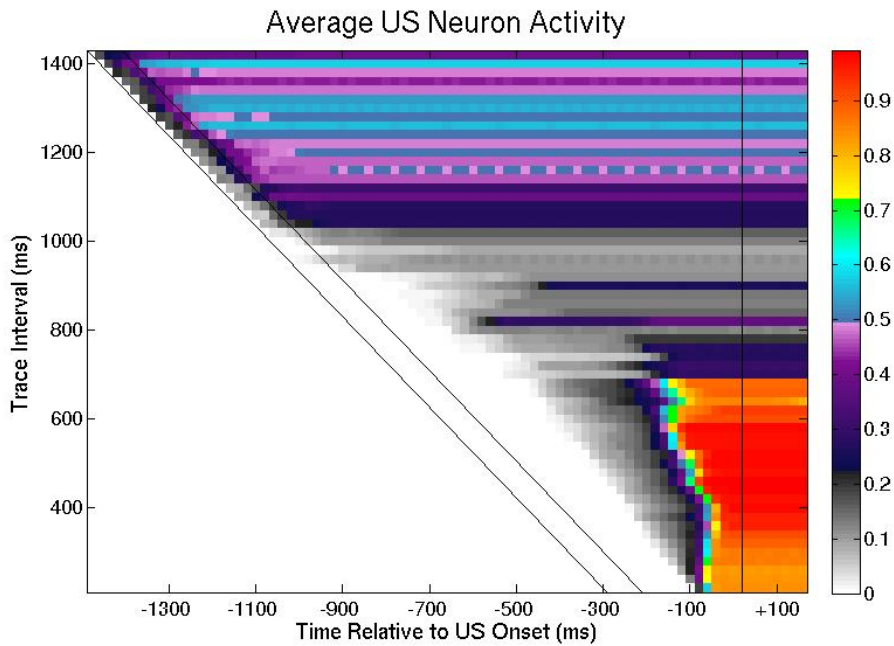


Figure 5 : Criterion-based learned performance is sensitive to the trace interval. Between 200 ms and 640 ms, successful prediction obtains in almost all simulations. There is a precipitous decline in successful prediction for trace intervals 640 – 720 ms. The simulations cannot reliably predict US onset for intervals longer than 720 ms. Both failure modes, predict-too-soon and failure-to-predict, coexist at the onset of the decline (660 ms) until about 1200 ms. Starting at 960 ms, the predict-too-soon failure mode becomes increasingly prominent. The ten simulated networks in this figure differ in terms of connectivity, but they are trained with identical sequences of initial states. See methods for performance criteria. The data was collected from a test trial after 200 training trials.

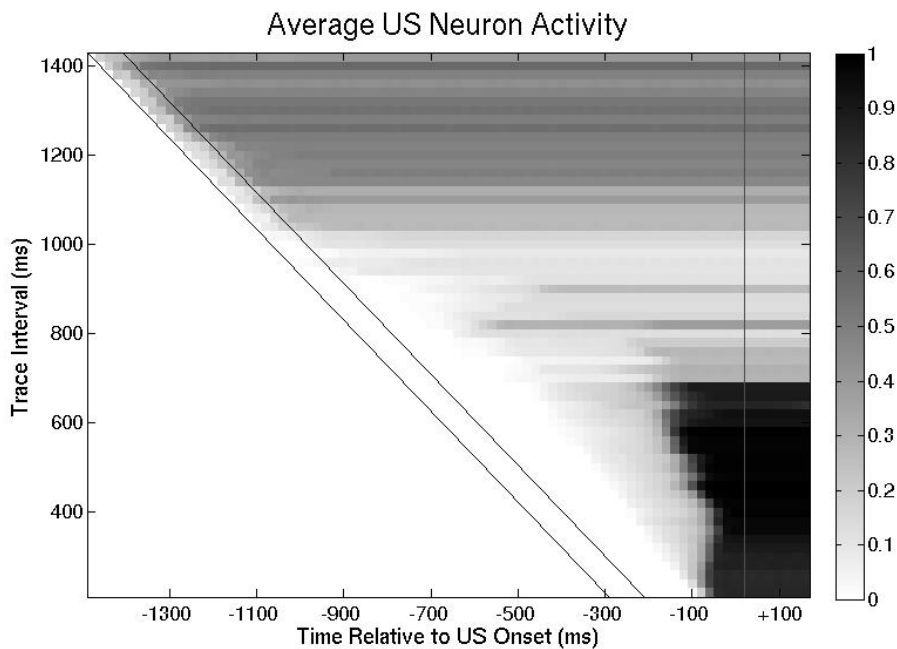
The Phase Transition in Learned Behavior

Trace interval influences the US prediction modes that result from training. Figure 5 illustrates how the US prediction modes change as a function of trace interval. Ten different networks were trained for each trace length. The successful prediction phase begins at 200 ms and extends to 640 ms. For this range, almost all simulations (>95%) produce successful US prediction. Trace intervals longer than 640 ms initiate a region of phase transition in US prediction. Depending on connectivity alone, simulations for trace intervals of 660 ms, 680 ms and 720 ms each produce the three different behavior modes at the conclusion of training trial 200. Beyond 720 ms, few simulations can produce a timely prediction. The failure-to-predict mode dominates from 720 ms until around 960 ms. Then, as trace interval duration exceeds 960 ms, simulations express the predict-too-soon mode with steadily increasing frequency. For

trace intervals of 1200 ms and longer, the behavioral mode is almost always predict-too-soon (>95%).



[above - color version, this version is misleading when printed in gray scale]



[above - gray scale version]

Figure 6 : From the perspective of average US neuron activity, the phase transition in US-onset prediction occurs from trace intervals of 640 ms to 780 ms. Average fractional US neuron activity is plotted here as a function of time within a simulation (x-axis) and trace interval duration (y-axis). Three major regions of US activation are apparent from the diagram. For trace intervals of 200-640 ms, the US neurons become vigorously active shortly before US onset and during the US itself. Predict-too-soon activity, indicated by US firing that can extend back to the CS, is striking at the longest trace intervals (>

840 ms). Intermediate trace lengths (640 ms – 840 ms) are somewhat ambiguous because the average activity data contains a mixture of firing modes, just as seen in the phase diagram (Figure 5). The sequence of random initial states were identical for each simulation. The US neuron firing data were averaged across 10 different network connectivities on test trial 200 and are identical to the data analyzed for Figure 4. The two diagonal lines mark CS onset (leftmost) and offset (rightmost), while the vertical line marks US onset.

The Phase Transition in Average US Neural Activity

The phase transition controlled by trace interval is evident even without the criteria used in Figures 4 and 5, and average cell firing also indicates that distinct phases exist. Figure 6 displays the average US neuron activity after 200 training trials for 10 simulations per trace interval – the same simulations used for Figure 5.

The large majority (>90%) of US neurons fire appropriately for successful prediction for trace intervals from 200 ms to 640 ms. This range delineates the first phase. Most US neurons (>90%) trained for this range do not begin firing until between 200 ms to 60 ms before US onset. After robust firing (>90% of US neurons) begins, it persists until the end of the test trial. Interestingly, US neurons predict a little earlier, relative to US onset, as the trace interval increases, but the earliest US firing does not change as fast as trace interval itself is lengthened.

Beyond a 640 ms trace interval, average US neuron firing is significantly different. Most neurons are no longer vigorously active. The rise in unsuccessful prediction seen in Figure 5 is based in this decrease in active neurons. For trace intervals beyond 940 ms, US neurons begin firing soon after CS onset. For intervals longer than 1100 ms, around half of the US neurons are active on every time step.

Exchanging the source of randomness, i.e., leaving connectivity constant and randomizing the sequence of initial-activity vectors across simulations, produces similar results. The simulations learn to successfully predict (>95%) for trace intervals between 200 ms and 560 ms. Comparing the categorized behavioral phase diagrams reveals that the phase transition is not as dramatic. For trace intervals between 580 ms and 800 ms, the number of successful predictions progressively fall to zero. All three behavior modes appear for several different intervals in this range depending on the sequence of initial states alone. Simulations do not successfully predict trace intervals longer than 800 ms. As in Figure 5, the predict-too-soon mode dominates for extra long trace intervals. Average US neuron activity is qualitatively indistinguishable from Figure 6.

Instability of Individual Simulations

Evaluation of simulation behavior has thus far focused on a single trial of ten different simulations, but this focus obscures the dynamics of learned US-neuron firing across trials. That is, US-neuron firing on a single trial of a single simulation does not necessarily reflect US-neuron firing on other, nearby trials. In fact the stability, or instability, of learned US-neuron firing is directly related to the phases described in Figures 5 and 6. In addition, several other measures reflect the stability (or instability) of learned US firing and therefore, of the implied behavioral response. This section focuses on the representative cell firing patterns from a single connectivity and illuminates some of the detailed dynamics of individual simulations.

For trace intervals 200-640 ms just as noted earlier, during approximately the first hundred or so training trials, very few US neurons fire (the failure-to-predict mode). Suddenly, with just a few

more training trials, the rapid phase transition in learning occurs and most US neurons (>90%) begin to fire and produce a timely prediction of US onset. After this behavioral phase transition, US neuron firing is remarkably stable through 650 total training trials (see Figure 8, bottom row, for an example). Neuron firing in the whole simulation is also stable from one trial to the next (Figure 8, top row, for an example). Simulations also quickly approach a stable value of (1) total firing (average activity for a trial), (2) asymptotic average weight values, and (3) asymptotic low numbers of neurons that never fire during a training trial (see Figure 9 for an example).

Long, unlearnable trace intervals (those exceeding 1200 ms) also demonstrate stability. Average activity is well controlled, and simulation-wide firing for adjacent training trials are very similar. The simulations very consistently predict too soon and predictive firing begins during the CS. Simulations for these extra long trace intervals demonstrate learning patterns similar to simulations for the 200 ms to 640 ms range. For approximately the first 120 training trials, US neurons do not fire (fail to predict). Suddenly, prediction starts to occur and predictive firing occurs on nearly every subsequent training trial. Other measures of simulation stability produce trends and values comparable to those for learnable intervals.

Simulations learning trace intervals between 700 ms and 1200 ms are characterized by instability. The predictions produced by US neuron firing vary from trial to trial. Fluctuations and instability are most dramatic and most prevalent immediately after successful prediction begins to decline. For even longer trace intervals, those approaching 1200 ms, simulations gradually become more stable as trace intervals increase.

Figure 7 shows how volatile US neuron firing is across trials, even after a significant amount of training. After 185 training trials, the simulation produces a predict-too-soon response. After 187 trials, it produces successful prediction. On the subsequent trial, 188, it does not provide the smallest prediction of US onset nor any significant activity during the US itself as a completion of the bridging sequence.

Simulation instability rises sharply when the majority of simulations stop successfully predicting US onset. Therefore, instability is also indicative of the phase transition induced by increasing trace interval. Figure 8 shows the learned behavior of one network connectivity trained with one set of initial states on three different trace intervals, 660 ms, 680 ms and 700 ms. As trace interval is increased, each simulation takes longer to reach a stable trial-to-next-trial, cosine comparison (see the left hand column of Figure 8). In terms of predicting US onset, the 660 ms simulation succeeds around 200 trials and maintains this success; the longer 680 ms interval takes about twice the number of trials to reach a stable and successful prediction of US onset; finally, the 700 ms interval is too long for the simulation to learn to correctly predict the onset – even after 650 training trials.

The instability of the simulation is evident in other measures besides cosine similarity of neuron firing and average US neuron activity. Average activity, average weight value, and fraction of non-firing neurons all reflect the instability too. Figure 9 depicts some of these measures using the same simulations as Figure 8.

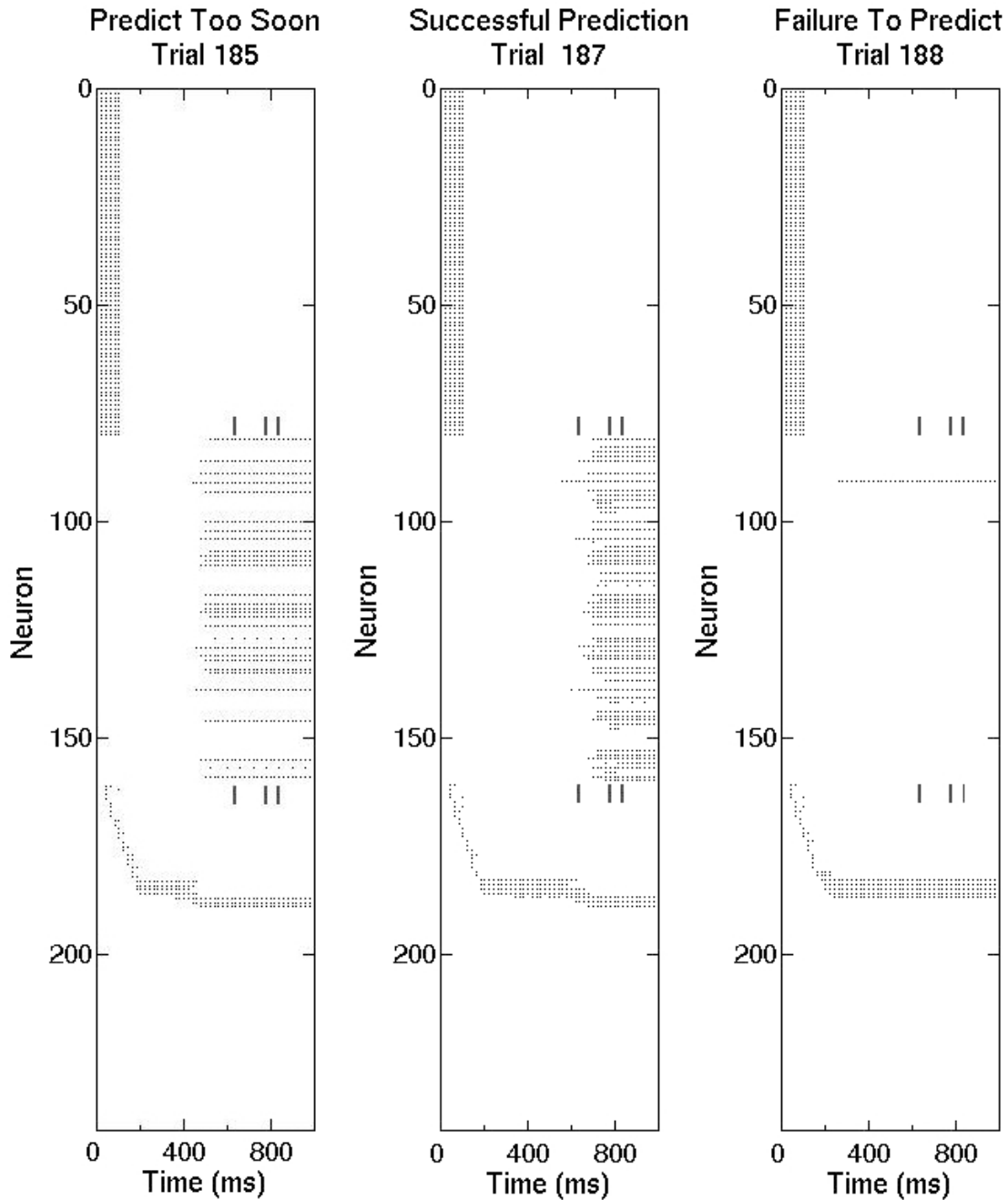


Figure 7 : The US prediction is unstable for trace intervals between 700-1200 ms. In this example from a trace interval of 720 ms, one simulation produces all three learned behavioral modes within four sequential trials. In the leftmost raster diagram (trial 185), sufficient numbers of US neurons fire too early, triggering the early prediction criterion. In the middle raster diagram (trial 187), strong prediction of this US begins within the appropriate interval preceding the US onset. In the rightmost raster diagram (trial 188), the system fails to predict the US except for one lone neuron, which becomes active far too early. Vertical hash marks define the criterion response regions (see Methods for details). Neurons from 161-240 for each diagram are reordered based on trial 188 and sampled to be representative of the 7840 neurons that do not directly receive the external input (see Methods for details).

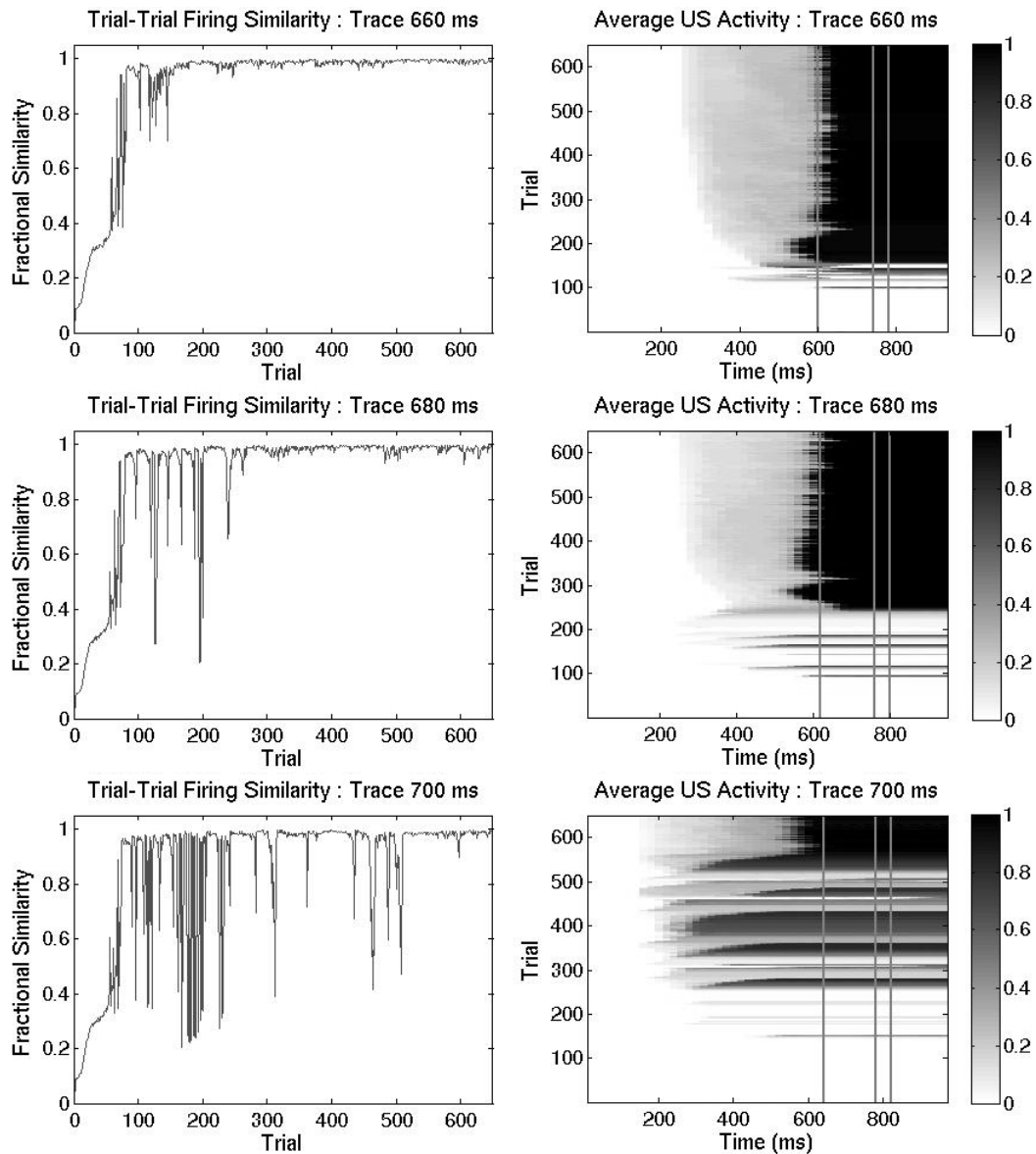


Figure 8 : Neuron firing pattern stability and US neuron prediction are relatively stable for 660 ms and 680 ms, but fluctuate significantly when the trace interval is 700 ms. In the top row, the trace interval is 660 ms; a trace interval of 680 ms is in the center row; and the bottom row illustrates a trace interval of 700 ms. The left column shows the similarity of firing patterns between the current trial and the previous trial. The simulation quickly achieves stable firing patterns for the 660 ms trace interval, but it produces shifting and, sometimes, dramatically different firing patterns from trial to trial for a trace interval of 700 ms. The right column illustrates the simulation's prediction via the average activity of the US neurons. Average fractional US neuron activity is plotted across simulation time (x-axis) and number of training trials (y-axis). For 660 ms, the average firing pattern is relatively stable and produces good predictions after 100 – 200 training trials. At 680 ms, the simulation produces good prediction after several hundred trials while being less stable throughout most earlier training trials. The simulation trained on the 700 ms trace interval is apparently unstable for at least five hundred trials. Its US neurons shift their firing times repeatedly. From a behavioral perspective, it alternates between two distinct modes - predict too soon and failure to predict. In the right column, the leftmost line indicates the earliest acceptable prediction, the middle line indicates the first time that is too late for successful prediction and the rightmost line indicates US onset.

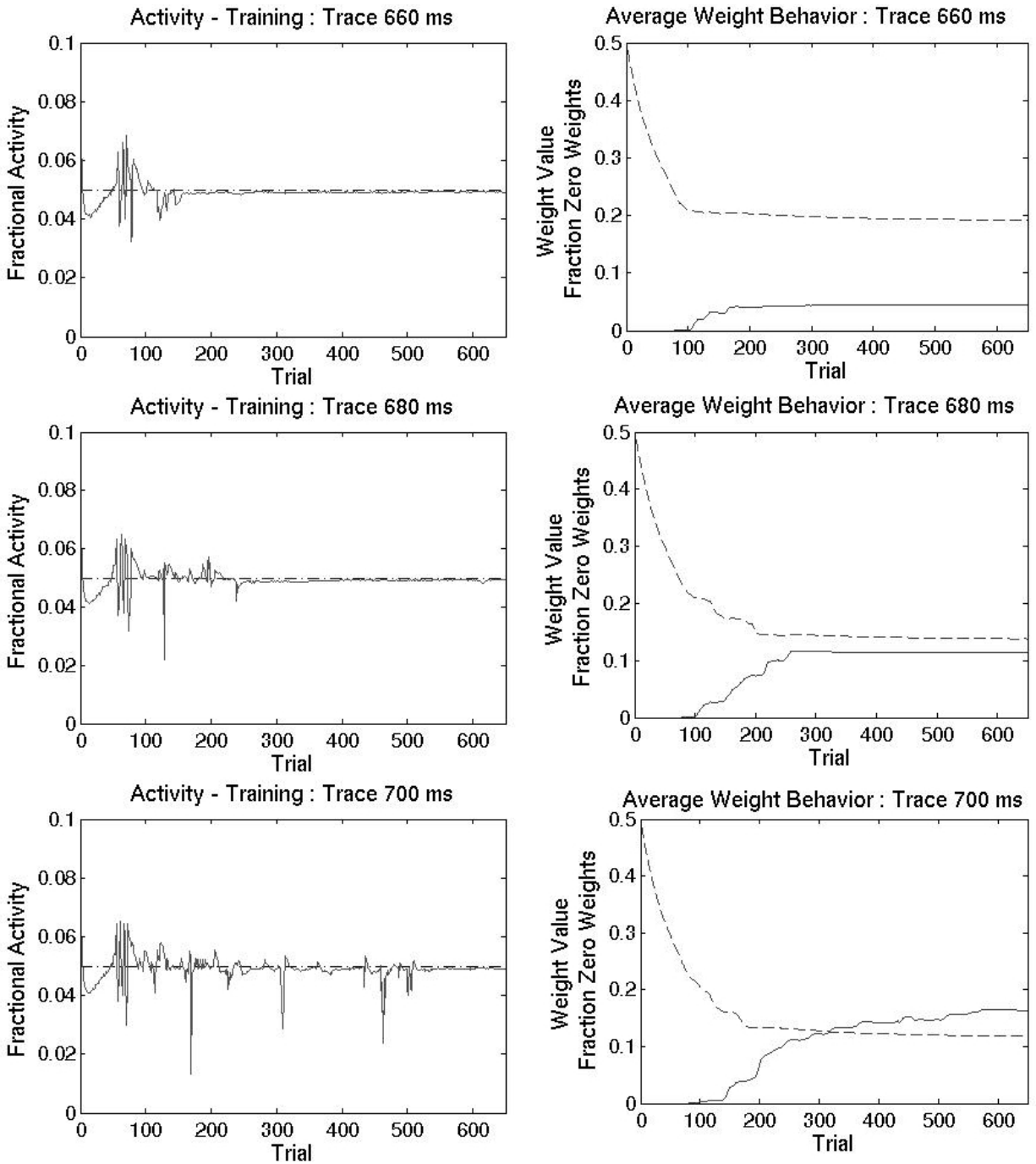


Figure 9 : Selected mesoscopic variables are relatively stable at 660 ms, but fluctuate rapidly when the trace interval is 700 ms. The left column of figures depicts the average fractional activity (solid line) of each training trial. Desired activity is plotted with a dashed line. Fractional activity, one of the simplest measures of simulation dynamics, reflects the instability associated with lengthening the trace interval beyond what can be successfully learned. The right column of figures depicts data summarizing the synaptic weight values. The dashed line indicates the average, non-zero synaptic weight, and the solid line indicates the fraction of total weights which are zero. For a trace interval of 660 ms, the average non-zero weight value is almost twice that of the value for 700 ms. Likewise, the fraction of zero weights rises to a stable level for the 660 ms interval, while it climbs three times as high without as much stability for the 700 ms trace interval.

Discussion

Trace intervals around 720 ± 40 ms produce the most interesting result. At these trace intervals, there will be a rapid fluctuation of behavioral modes because of the trial-to-trial fluctuation of the US-firing patterns. Of interest, these results can be predictions for rabbit eye-blink trace conditioning experiments. Such experimental predictions are encouraged by the earlier quantitative experimental predictions or data fitting; specifically, the model does fit the learnable interval, training trials to learning, and the abruptness of the onset of the learned blink (Levy & Sederberg 1997; Rodriguez & Levy 2001). The results obtained here imply that a single rabbit trained on a trace interval of approximately 720 ms will, after about 150 training trials, unpredictably jump between blinking too soon and failing to blink at all. The rabbit may rarely produce a temporally appropriate blink, blocking the US. However, given the simplicity of the model (McCulloch-Pitts neurons), crude time resolution (20 ms), small number of neurons (8,000 in simulation vs. $\sim 250,000$ in a rat CA3), and the non-biological input (no fluctuations), we must circumscribe the *in vivo* behaviors and electrophysiology that can be anticipated based on the results presented here.

Recall our interpretation of hippocampal function (Levy 1989) : the dentate gyrus-CA3 system recodes neocortical inputs in a manner suitable for making predictions. Thus, the hippocampal output must properly anticipate the US-onset, and based on projections back to neocortex and into the midline striatal system, other brain regions have the option to act on this prediction. The hippocampal prediction for the trace eye-blink paradigm is the US-onset because appropriate anticipation of this onset is what a decision region needs to avoid the unpleasant US. Thus, our interest centers on the within-trial onset of US firing although we can only guess at the delay involved when another brain region uses the CA3-generated US-onset prediction. Thus, there is an arbitrariness to the criterion of the appropriate time and activity level for learned US cell firing to predict US onset.

Further assessment of the model's implication for behavioral or electrophysiological experiments is helped by understanding why the model here has its demonstrated properties. From our viewpoint, nothing of this interpretation disqualifies the fluctuation prediction.

First a conjecture mentioned elsewhere (Levy et al. 2005) : The learnable trace interval is limited by the density of randomly discoverable neuron-firing sequences. For trace intervals of 200 ms – 640 ms, there must be many sequences in recurrent neuron state space that can be entered by CS neuron-firing and that can also be terminated by US neuron-firing. However, as the trace interval extends beyond this range, suitable neuron firing sequences (called codes) become rarer and rarer. One reason they become rarer is that a neuron cannot be used in two noncontiguous portions of the sequence, and as trace length increases, the simulation runs out of neurons to recruit for bridging sequence segments. Another reason seems to be a weakening of the available sequential codes. Weakening refers to the net, total connection strength from a set of firing neurons at one time step into the set of firing neurons a time step hence. Evidence for the weaker codes at longer intervals is seen in the right column of Figure 9.

Based on the mode fluctuations, the sequential recurrent codes are apparently competing with an attractor dominated by firing US neurons. Thus, any weakening of sequential recurrent codes allows the end-of-sequence US neuron-firing to invade earlier timesteps by virtue of a backward coding cascade. In part this competition occurs and is modulated by the tendency for backward cascading.

The backward cascade, also referred to as “predictive recall”, was predicted in Levy (1989, e.g., see Figure 14A; 1990; Levy et al 1995 - Figure 6) and first published with simulation results in Prepscius and Levy (1994). Blum and Abbott (1996) described it for an animal interacting with a changing input (taking short cuts). Levy and Sederberg (1997) were the first to document the backward cascade in the trace classical conditioning paradigm and Levy et al. (2005) performed detailed analysis (see Figure 4).

Although the US neurons can backward cascade, they occupy a unique position: they are activated at the end of every training trial, and thus will always be a moderately stable attractor due to potentiation between US neurons and depotentiation of most other inputs. The depotentiation apparently leads to relative stability for intervals in the 200 ms - 640 ms range, where US neurons do not tend to backward cascade too far.

Long trace intervals apparently facilitate backward cascade of US neurons, possibly due to the weakness of the associatively-created sequence of recurrent neurons. Because neurons cannot fire in two distinct segments of a bridging sequence without destroying the bridge, long sequences tend to be weak because they have a greater tendency to try to reuse neurons via associative LTP and concomitant LTD of inactive inputs.

The strength of the weights for shorter, learnable trace intervals indicate that shorter trace intervals do not attempt to incorporate neurons in two distinct parts so often as longer sequences. This critical detail helps to explain the stability, and perhaps the predictive success, by the vast majority of simulations learning short to moderate length trace intervals.

On the other hand, the strength of the US attractor remains relatively constant across trace intervals in terms of synaptic weights to neurons of the attractor. Moreover, because all neurons can be randomly activated at any time point and because US neurons are activated on every training trial, there is also a high potential for weak associative synaptic strengthening from every neuron with connections into the US attractor system of neurons. Thus, we have an idea why shorter intervals are successfully learned and more stable, but longer intervals eventually produce a US collapse, and finally, why in between intervals are associated with fluctuating modes. That is, the weak associations of early-firing neurons into the US-attractor are eventually stronger than the progressively weaker recurrent sequence of neuron firings. In sum, the work here builds from ideas about learning phase transitions found in motor systems (Kelso 1995) and discovers, in the model, other variables affecting phase transition-like behavior. These discoveries, through simulations, go beyond the published experimental observations and predict outcomes of novel experimental parameterizations and measurements. Specifically, the relatively abrupt, learning-based phase transition of the trace eye-blink paradigm will produce a region of phase instability over a particular, small range of trace intervals.

Acknowledgments

This work was supported by a grant from the National Institutes of Health, NS041582.

References

Alvarado M, Rudy J (1995) Rats with damage to the hippocampal formation are impaired on the transverse-patterning problem but not on elemental discriminations. *Behavioral Neuroscience* 109: 204–211.

August DA, Levy WB (1999) Temporal sequence compression by an integrate-and-fire model of hippocampal area CA3. *J Comp Neurosci* 6: 71-90.

- Blum KI, Abbott LF (1996) A model of spatial map formation in the hippocampus of the rat. *Neural Comput* 8: 85-93.
- Dusek JA, Eichenbaum H (1997) The hippocampus and memory for orderly stimulus relations. *Proc Natl Acad Sci* 94: 7109-7114.
- Gormezano I, Kehoe EJ, Marshall BS (1983) Twenty Years of Classical Conditioning Research with the Rabbit. In: AN Epstein, ed. *Progress in Psychobiology and Physiological Psychology*. Academic Press, New York, NY. pp. 197-267.
- Kelso, JAS (1995) *Dynamic patterns: The Self-Organization of Brain and Behavior*. MIT Press, Cambridge, MA.
- Kim JJ, Clark RE, Thompson RF (1995) Hippocampectomy impairs the memory of recently, but not remotely, acquired trace eye-blink conditioned responses. *Behavioral Neuroscience* 109: 195-203.
- Levy WB (1989) A computational approach to hippocampal function. In R Hawkins, G Bower, eds. *Computational Modeling of Learning in Simple Neural Systems*. Academic Press, Orlando, FL. pp. 243-305.
- Levy WB (1990) Maximum entropy prediction in neural networks. *International Joint Conference on Neural Networks I*: 7-10.
- Levy WB (1994) Unification of hippocampal function via computational considerations. *INNS World Congress on Neural Networks IV*: 661-666.
- Levy WB (1996) A sequence predicting CA3 is a flexible associate that learns and uses context to solve hippocampal-like tasks. *Hippocampus* 6: 579-590.
- Levy WB, Sanyal A, Rodriguez P, Sullivan DW, Wu XB (2005) The formation of neural codes in the hippocampus: trace conditioning as a prototypical paradigm for studying the random recoding hypothesis. *Biological Cybernetics* 92: 409-426.
- Levy WB, Sederberg PB (1997) A neural network model of hippocampally mediated trace conditioning. *IEEE International Conference on Neural Networks I*: 372-376.
- Levy WB, Steward O (1979) Synapses as associative memory elements in the hippocampal formation. *Brain Research* 175: 233-245.
- Levy WB, Steward O (1983) Temporal contiguity requirements for long-term associative potentiation/depression in the hippocampus. *Neuroscience* 8: 791-797.
- Levy WB, Wu X, Baxter RA (1995) Unification of hippocampal function via computational/encoding considerations. In: DJ Amit, P del Giudice, B Denby, ET Rolls, A Treves, eds. *Proceedings of the Third Workshop on Neural Networks: from Biology to High Energy Physics*. *Intl. J. Neural Sys* 6, (Supp.). World Scientific Publishing, Singapore. pp. 71-80.
- McEchron MD, Disterhoft JF (1997) Sequence of single neuron changes in CA1 hippocampus

of rabbits during acquisition of trace eyeblink conditioned responses. *J Neurophysiol* 78: 1030-1044.

Minai AA, Levy WB. (1993) Sequence Learning in a Single Trial. *World Congress on Neural Networks (WCNN) Volume 2*: 505-508.

Morris R (1984) Developments of a water-maze procedure for studying spatial learning in the rat. *J Neurosci Methods* 11: 47-60.

Moyer JR Jr., Deyo RA, Disterhoft JF (1990) Hippocampectomy disrupts trace eye-blink conditioning in rabbits. *Behavioral Neuroscience* 104: 243-252.

Prepscius C, Levy WB (1994) Sequence prediction and cognitive mapping by a biologically plausible neural network. *INNS World Congress on Neural Networks IV*: 164-169.

Rodriguez P, Levy WB (2001) A model of hippocampal activity in trace conditioning: Where's the trace? *Behav Neurosci* 115: 1224-1238.

Scoville WB, Milner B (1957) Loss of recent memory after bilateral hippocampal lesions. *J Neurol Neurosurg Psychiatry* 20: 11-21.

Smith AC, Wu XB, Levy W (2000) Controlling activity fluctuations in large, sparsely connected random networks. *Network* 11: 63-81.

Solomon PR, Vander Schaaf ER, Thompson RF, Weisz DJ (1986) Hippocampus and trace conditioning of the rabbit's classically conditioned nictitating membrane response. *Behavioral Neuroscience* 100: 729-744.

Sullivan DW, Levy WB (2003) Synaptic modification of interneuron afferents in a hippocampal CA3 model prevents activity oscillations. *International Joint Conference on Neural Networks (IJCNN) 2003 Proceedings*: 1625-1630.

Wu X, Levy WB (2005) Increasing CS and US longevity increases the learnable trace interval. *Neurocomputing* 65-66: 283-289.