LETTER Communicated by Randy Gallistel

A Scale-Invariant Internal Representation of Time

Karthik H. Shankar
shankark@bu.edu
Marc W. Howard
marc777@bu.edu
Center for Memory and Brain, Boston University, Boston, MA 02215, U.S.A.

We propose a principled way to construct an internal representation of the temporal stimulus history leading up to the present moment. A set of leaky integrators performs a Laplace transform on the stimulus function, and a linear operator approximates the inversion of the Laplace transform. The result is a representation of stimulus history that retains information about the temporal sequence of stimuli. This procedure naturally represents more recent stimuli more accurately than less recent stimuli; the decrement in accuracy is precisely scale invariant. This procedure also yields time cells that fire at specific latencies following the stimulus with a scale-invariant temporal spread. Combined with a simple associative memory, this representation gives rise to a moment-to-moment prediction that is also scale invariant in time. We propose that this scale-invariant representation of temporal stimulus history could serve as an underlying representation accessible to higher-level behavioral and cognitive mechanisms. In order to illustrate the potential utility of this scale-invariant representation in a variety of fields, we sketch applications using minimal performance functions to problems in classical conditioning, interval timing, scale-invariant learning in autoshaping, and the persistence of the recency effect in episodic memory across timescales.

1 Introduction

Theories of timing and theories of episodic memory have generally been decoupled in the cognitive neuroscience literature. Timing experiments indicate a degraded precision in memory for longer time intervals; this degradation shows similar characteristics across timescales (Gibbon, 1977). Similarly, uncued memory tasks indicate that forgetting takes place over multiple timescales with similar characteristics in all scales (Howard, Youker, & Venkatadass, 2008; Moreton & Ward, 2010). From a theoretical perspective, this is a hint suggesting a common mechanism underlying the memory for time and memory for a stimulus or event across these apparently diverse domains.

Behavioral experiments on perception or memorization of time intervals point to a scale-invariant internal representation of time. For instance, when human subjects are instructed to reproduce the duration of a fixed interval (Rakitin et al., 1998; Ivry & Hazeltine, 1995), the reproduced duration on average matches the fixed duration, and its variability is simply proportional to the fixed duration. The coefficient of variation (CV, defined as the standard deviation divided by the mean) of the reproduced duration is the same for any other fixed duration. More important, the distribution of the reproduced duration is scale invariant with respect to the fixed duration. That is, the response distributions corresponding to different durations overlap when linearly scaled. Such distributions are said to possess the scalar property. Similarly, in animal conditioning experiments (Roberts, 1981; Smith, 1968) where a conditioned stimulus (CS) is reinforced by an unconditioned stimulus (US) with a fixed latency, the distribution of the conditioned response peaks at the appropriate reinforcement latency and is scale invariant with respect to the reinforcement latency. The fact that these distributions are scale invariant with appropriately timed peaks suggests the existence of a reliable internal representation of time that is scale invariant.

Scale invariance is not simply restricted to the timing of response distributions. It is also observed in the rate of associative learning in animal conditioning. The number of CS-US pairings required for the animals to learn the association between the CS and the US is shown to increase when the reinforcement latency is increased and decrease when the time between successive learning trials, the intertrial interval, is increased (Gallistel & Gibbon, 2000). It turns out that the absolute value of the reinforcement latency and the intertrial interval do not matter. As long as their ratio is fixed, the number of learning trials required is a constant. That is, there is no characteristic timescale involved in learning the associations between stimuli (Balsam & Gallistel, 2009). This suggests that the internal representations of the reinforcement latency and the intertrial interval should be scale invariant.

It is well known that memory performance decays with time. It has been argued that the decay follows a scale-invariant power law function (see Wixted, 2004, for a review). We focus here on free recall, a popular paradigm used to probe human episodic memory over laboratory timescales ranging from a few hundred milliseconds to a few thousand seconds. In free recall, subjects are given a list of words and are asked to recall them in any order in which they come to mind. In general, the items from the end of the list are better recalled, pointing to the fact that the more recently learned items are more easily retrieved. This general phenomenon is referred to as the recency effect. It has been repeatedly observed that altering the time delay between the presentation of items or the recall phase (or both), dramatically affects the recency function. While the empirical case for scale invariance in free recall is mixed (Chater & Brown, 2008; Brown, Neath,

& Chater, 2007, but see Nairne, Neath, Serra, & Byun, 1997), it is clear that the recency effect has been observed with similar properties over a range of timescales, from fractions of seconds (Murdock & Okada, 1970) to dozens of minutes (Glenberg et al., 1980; Howard et al., 2008). Having multiple memory stores operating at different timescales (Atkinson & Shiffrin, 1968; Davelaar, Goshen-Gottstein, Ashkenazi, Haarmann, & Usher, 2005) can of course yield recency effects across multiple timescales. But a simpler account with a single memory mechanism can be envisaged based on a scale-invariant representation of time. If the stimulus history is represented in the brain along a scale-invariant internal time axis, we can naturally obtain a scale-free recency effect (Brown et al., 2007).

Although scale invariance seems to underlie many aspects of cognition involving time, it should be noted that not all aspects of timing behavior are scale invariant (see Wearden & Lejeune, 2008, for a review). The crucial point is that the existence of scale invariance in multiple aspects of behavior involving time suggests that a single scale-invariant representation of stimulus history could play a causal role in all of these domains. Violations of scale invariance are easy to reconcile with a scale-invariant representation. For instance, Wearden and Lejeune (2008, p. 571) note that "additional processes could modulate the expression of underlying scalar timing." These processes would presumably depend on the specific task demands in a particular empirical setting. On the other hand, it is extremely difficult to construct a theoretically satisfying account of scale-invariant behavioral data from an underlying representation of time that is not scale invariant. While it is technically possible to construct scale-invariant behavior from timing information with a scale, this would require that the task-dependent additional processes introduce a scale that happens to exactly cancel the scale of the underlying representation of time.

In this letter, we construct a scale-invariant internal representation of time based on a hypothesis about how the recent stimulus history could be represented in the brain (Shankar & Howard, 2010). Two key assumptions go into this construction. First, we posit a population of persistently firing neurons that behave as a filter bank of leaky integrators. These encode the Laplace transform of the stimulus history. Second, we posit a specific connectivity structure involving bands of balanced excitation and inhibition that transcribes the activity of these leaky integrators into activity in another population of time cells. This transcription procedure is based on an elegant approximation to the inverse Laplace transformation (Post, 1930); hence, we refer to it as timing from inverse Laplace transform (TILT). The time cells thus constructed have properties similar to the output of a set of tapped delay lines (Desmond & Moore, 1988) in that they respond to a stimulus following fixed latencies. We propose that such a representation of timing information in the activity of time cells could underlie behavioral scale invariance for timescales ranging from a few hundred milliseconds to a few thousand seconds.

To demonstrate the potential utility of this scale-invariant representation of time and stimulus history, we use a simple Hebbian learning rule to generate predictions based on the match between the current state of the representation and previous states. This match generates moment-to-moment predictions in real time that can account for scale invariance at the behavioral level. In order to make task-specific behavioral predictions, we specify minimal performance functions that map the prediction function onto appropriate behaviorally observable dependent measures. Our focus in this letter is not to faithfully model the details of these behavioral tasks, but to qualitatively illustrate how the properties of the prediction function derived from the scale-invariant representation of time and stimulus history could give rise to scale-invariant behavior in a broad variety of behavioral domains, including classical conditioning, timing behavior, and episodic memory. Quantitative features of the behavioral measures, such as the coefficient of variation of timing responses or the decay rate of memory, depend strongly on the choice of the performance functions. Since our focus is not on the specific performance functions and because they have a large effect on the predicted behavioral results, we do not attempt quantitative fits to the data.

In this letter, we describe the model and analytically establish scale invariance. We then describe several applications to demonstrate the model's applicability to classical conditioning, interval timing, and episodic memory experiments. In particular, we illustrate appropriately timed behavior in simple classical conditioning and human interval timing experiments, show that the encoded timing information in the online predictions can be exploited to account for the lack of a specific scale in learning associations between stimuli, and describe the recency effect across timescales in free recall experiments.

2 Description of the Model

Before describing the mathematical details, we first give an overview of the basic architecture of the model. It is convenient to describe the model in two stages: the timing mechanism TILT and the associative learning mechanism.

Let $\mathbf{f}(\tau)$ be a vector-valued function that denotes the presentation of a stimulus at any time τ. To describe the timing mechanism, we focus on one of its components, as shown in Figure 1a. The stimulus function activates a column of **t** nodes at each moment. The aim is to reconstruct the history of the stimulus function as a spatially distributed pattern of activity across a column of **T** nodes. We take the column of **t** nodes to be a filter bank of leaky integrators. Once each **t** node is activated, its activity persists over time with a distinct decay rate. At each moment, the pattern of activity distributed along the column of **t** nodes is transcribed by the operator $\mathbf{L}_k^{-1}$ to construct the activity in the column of **T** nodes. Mathematically, the activity in the **t** nodes represents the Laplace transform of the presentation history

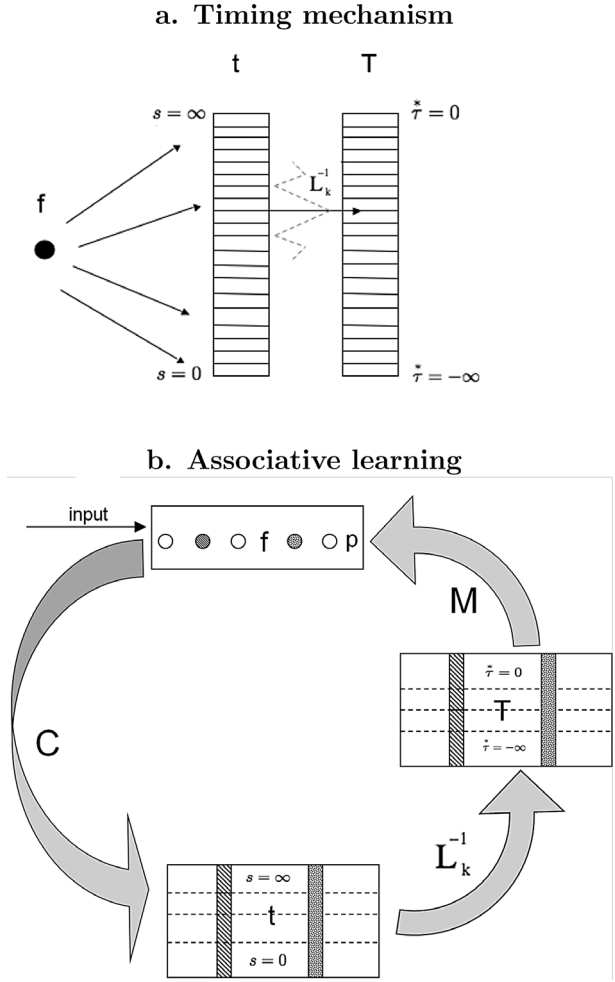

Figure 1: (a) Timing mechanism. The stimulus function activates a **t** column of leaky integrators. Each node in the **t** column has a distinct decay rate s. The activity in the **t** column is mapped onto the column of time cells **T** via the operator $\mathbf{L}_k^{-1}$. At any moment, the activity distributed across the **T** column represents a fuzzy but scale-invariant presentation history of the stimulus. (b) Associative learning. For each node in the stimulus layer **f**, there is a column of nodes in the **t** and **T** layers, as represented by appropriate shading. The **T**-layer activity at each moment is associated in a Hebbian fashion with the **f** layer activity, and these associations are stored in **M**. The associations stored in **M** and the instantaneous **T**-layer activity induce activity in the **f** layer. This internally generated activity in the stimulus layer is interpreted as the prediction **p** for the next moment.

of the stimulus, and the operator $\mathbf{L}_k^{-1}$ approximates the inverse Laplace transformation. The activity distributed along the **T** nodes can be identified as an estimate of the presentation history of the stimulus. The estimate of the stimulus history function is not precise, with errors increasing for times further in the past. The different nodes of the **T** column peak in activity at different latencies following the stimulus presentation. Hence, the **T** nodes can be thought of as time cells. Functionally the activity of the **T** column resembles the output of a set of tapped delay lines, although with important differences that we discuss later.

Figure 1b describes an associative learning model that utilizes the representation of time constructed by the timing mechanism. We assume that the stimulus perceived at any moment can be represented as a vector in a high-dimensional vector space, $\mathbf{f}(\tau)$. The dimensionality of this space is assumed to be small relative to the number of possible distinct stimuli but large relative to the number of stimuli that are relevant in the experiments we describe in the application section. The nodes of the **f** layer constitute the basis vectors of this space. The different **f** nodes represent the distinct dimensions of the stimulus space. Any unique stimulus would be represented by a unique set of activations across these nodes; when that stimulus is presented, those nodes would be activated accordingly. Any two distinguishable stimuli will have distinct representations that overlap to a degree that depends on their similarity. In general, the exact representation of a stimulus in the **f** layer would depend on its perceptual properties and the prior experience with that stimulus. We also assume that corresponding to each **f** node, there is a column of **t** nodes and a column of **T** nodes, as described in Figure 1a. All the **t** columns constitute the **t**-layer, a two-dimensional sheet of leaky integrators. Similarly, all the **T** columns constitute the **T**-layer, a two-dimensional sheet of time cells.

At any moment, the activity in the **T**-layer represents the history of the stimuli encountered in the recent past. The stimulus experienced at any moment, the momentary **f**-layer activity, is associated with the momentary **T**-layer activity, which is caused by the stimuli experienced in the recent past. As depicted in Figure 1b, these associations are stored in the operator **M**, which can be interpreted as the synaptic weights between the **T** nodes (presynaptic) and the **f** nodes (postsynaptic). As learning progresses, **M** gathers the statistics underlying the temporal patterns among various stimuli. With sufficient learning, the **T**-layer activity at any moment will automatically stimulate activity in **f**-layer. This internally generated **f**-layer activity can be understood as the real-time prediction $\mathbf{p}(\tau)$ for the stimulus vector likely to be presented at the next moment (Shankar, Jagadisan, & Howard, 2009). The degree to which the **p** generated by the current state of **T** overlaps with a particular stimulus vector is primarily a consequence of the match between the current state of **T** and the state(s) of **T** in which that stimulus vector was encoded. It turns out that after learning, the temporal spread in $\mathbf{p}(\tau)$ reflects the scale invariance in the underlying timing

mechanism. Hence, **p** can give rise to scale-invariant timing behavior when fed into an approriate performance function.

The mathematical details of the timing mechanism and associative learning mechanism follow.

2.1 Timing Mechanism. We parameterize each column of **t** nodes by the variable s that represents the decay rate of each leaky integrator. As a mathematical idealization, we let s range from 0 to ∞, but in reality, we would expect finite, nonzero lower and upper bounds for s. The stimulus function $\mathbf{f}(\tau)$ activates the **t** nodes in the following way:

$$\frac{d\mathbf{t}(\tau, s)}{d\tau} = -s \cdot \mathbf{t}(\tau, s) + \mathbf{f}(\tau). \tag{2.1}$$

This can be integrated to obtain

$$\mathbf{t}(\tau, s) = \int_{-\infty}^{\tau} \mathbf{f}(\tau')\, e^{s(\tau'-\tau)}\, d\tau'. \tag{2.2}$$

The **t** nodes are leaky integrators because they integrate the activity of an **f** node over time with an exponentially decaying weight factor. For continuous and bounded $\mathbf{f}(\tau')$, the function $\mathbf{t}(\tau, s)$ will be finite and infinitely differentiable with respect to s. Figure 2 illustrates the properties of the function $\mathbf{t}(\tau, s)$ for a sample stimulus. Observe from equation 2.2 that $\mathbf{t}(\tau, s)$ is exactly the Laplace transform of the function $\mathbf{f}(\tau')$ for $\tau' < \tau$, when s is identified as the Laplace domain variable restricted to purely real values. That is, at each moment, the entire presentation history of an **f** node up to that moment is encoded across the column of **t** nodes as its Laplace transform.

The nodes in the **T** column in Figure 1a are labeled by the parameter $\overset{*}{\tau}$, which ranges from 0 to $-\infty$. The **T** nodes are in one-to-one correspondence with the **t** nodes, with the mapping $s \rightarrow -k/\overset{*}{\tau}$. At each moment, the operator $\mathbf{L}_{\mathrm{k}}^{-1}$ constructs the activity in the **T** layer from the activity in the **t** layer in the following way.

$$\mathbf{T}(\tau, \overset{*}{\tau}) = \frac{(-1)^k}{k!} s^{k+1} \mathbf{t}^{(k)}(\tau, s) \qquad \text{where } s = -k/\overset{*}{\tau} \tag{2.3}$$

$$\mathbf{T} \equiv \mathbf{L}_{\mathrm{k}}^{-1}[\mathbf{t}].$$

Here k is a positive integer, and $\mathbf{t}^{(k)}(\tau, s)$ is the kth derivative of $\mathbf{t}(\tau, s)$ with respect to s. That is, $\mathbf{L}_{\mathrm{k}}^{-1}$ computes the kth derivative along each **t** column and maps it onto each **T** column. We discuss ways to implement $\mathbf{L}_{\mathrm{k}}^{-1}$ in the next section.

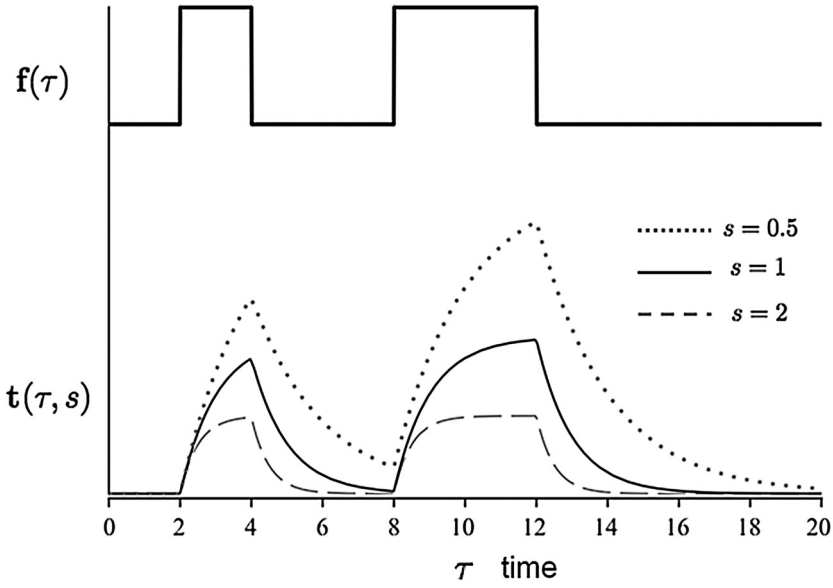

Figure 2: A stimulus is presented twice: for a duration of 2 seconds, followed by a gap of 4 seconds, followed by a subsequent presentation for a duration of 4 seconds. The curve on top represents the stimulus function as a sequence of square pulses of appropriate duration and the three curves below it represent the activity of the **t** nodes with decay rates $s = 0.5$ (dotted), 1 (solid), and 2 (dashed). Note that the activity of the **t** node with larger s (dashed line) saturates quickly, while the activity of the smaller s node (dotted line) takes longer to saturate. Also note that the decay of the smaller s node is sufficiently slow that its activity due to the first stimulus presentation persists at the time of the second presentation; the second presentation adds to the preexisting activity.

At any instant τ, the momentary activity distributed across a column of **T** nodes turns out to be a good approximation to the stimulus history function for all $\tau' < \tau$:

$$\mathbf{T}(\tau, \overset{*}{\tau}) \simeq \mathbf{f}(\tau') \qquad \text{where } \tau' = \tau + \overset{*}{\tau}. \tag{2.4}$$

This happens because $\mathbf{L}_k^{-1}$ approximates the inverse Laplace transformation. In fact, it has been proven (Post, 1930) that in the limit $k \to \infty$, $\mathbf{T}(\tau, \overset{*}{\tau})$ will exactly match $\mathbf{f}(\tau')$ in equation 2.4. In other words, $\mathbf{L}_k^{-1}$ is exactly the inverse Laplace transformation when k goes to infinity. We can now interpret $\overset{*}{\tau}$ as an internal representation of past time and the activity distributed along each **T** column as an approximate representation of the history of the corresponding dimension of the stimulus vector. We refer to this procedure

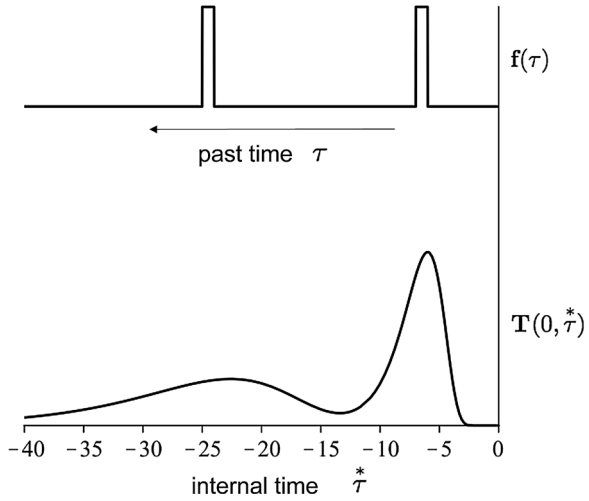

Figure 3: A stimulus is presented twice in the recent past. Taking the present moment to be $\tau = 0$, the momentary activity distributed across the **T** column nodes is plotted. Note that there are two bumps corresponding to two presentations, and the representation of the earlier presentation is more spread out.

of reconstructing the timing information of the stimulus history as timing from inverse Laplace transform (TILT).

To facilitate analytic calculations, we take the parameters s and τ^* to be continuous on an infinite domain. In reality, we would expect their domains to be discrete with finite, nonzero lower and upper bounds. The error that would be introduced due to discretization of s is derived in appendix B. In order to span the timescales relevant to cognition, we propose that τ^* ranges from few hundred milliseconds to few thousand seconds.

Figure 3 describes how the activity of the **T** nodes in a column at any instant represents the stimulus history of the corresponding **f** node. Consider a stimulus presented twice in the recent past: 7 seconds ago and 25 seconds ago. Taking the present moment to be $\tau = 0$, Figure 3 shows the momentary activity spread across the **T** nodes. The pattern of activity across the different τ^* nodes shows two peaks, roughly at $\tau^* = -7$ and $\tau^* = -25$, corresponding to the two stimulus presentations. For this illustration, we have used $k = 12$, and as a consequence, $\mathbf{T}(0, \tau^*)$ does not perfectly match the stimulus function. The match would be more accurate for larger k. The different τ^* nodes are activated at different latencies following the stimulus presentation. This property allows the column of **T** nodes to separately represent multiple presentations of the stimulus in the recent past, as shown in Figure 3. To further illustrate this, let us consider the activity of the **T** nodes

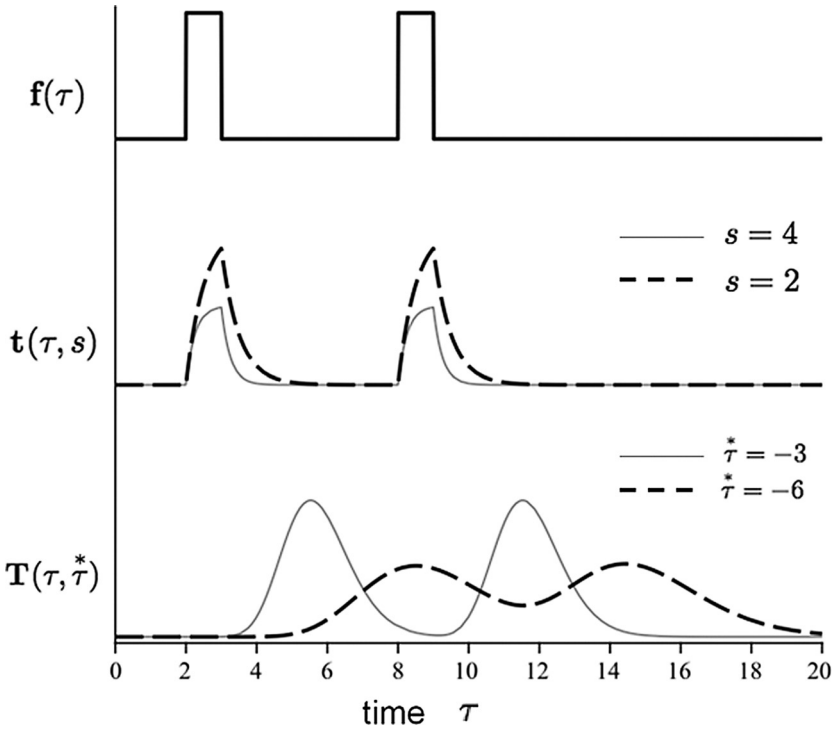

Figure 4: A stimulus is presented twice, and the activity of two nodes in the **t** and **T** columns is shown as a function of time. Note that the activity of the **T** nodes peaks at the appropriate delay after each stimulus presentation.

as a function of real time τ. Figure 4 shows an **f** node activated at two distinct times and the activity of two nodes from the corresponding **t** and **T** columns as a function of real time τ. Note that the **t** nodes peak at the stimulus offset, while the **T** nodes peak at specific latencies following the stimulus offset. The $\overset{*}{\tau} = -3$ node peaks 3 seconds following each stimulus presentation, and the $\overset{*}{\tau} = -6$ peaks 6 seconds after each stimulus presentation. For this reason, the nodes in the **T** layer can be thought of as time cells. Note that the real-time spread in the activity of the $\overset{*}{\tau} = -6$ node is larger than the spread of the $\overset{*}{\tau} = -3$ node. In the next section, we show that the temporal spread in the activity of a node with a specific $\overset{*}{\tau}$ value is proportional to $\overset{*}{\tau}$. For this illustration, we have used $k = 12$. For larger k, the temporal spread in the **T** nodes activity is smaller.

To summarize, the $\mathbf{L}_k^{-1}$ operator takes the Laplace transformed stimulus function on the s axis, inverts it, and projects it back to the internal time $\overset{*}{\tau}$ axis. Because k is finite, the inversion of the Laplace transform is only approximate, and the resulting function has a "temporal smear" when

compared to the original stimulus function. Moreover, when k is fixed, that is, when the $\mathbf{L}_k^{-1}$ operator takes the same order derivative everywhere on the s axis, this smear is scale invariant.

2.2 Implementation of $\mathbf{L}_k^{-1}$. We now describe how the operator $\mathbf{L}_k^{-1}$ can be implemented. Computing the exact inverse Laplace transform would involve evaluating contour integrals on the complex plane. Implementing such computations in the brain would require extremely complicated neural machinery. The primary problem is that to invert the Laplace transform $\mathbf{t}(\tau, s)$ in the standard way, we need complex values of s that are not stored in the $\mathbf{t}$ nodes. A second problem is that to evaluate the inverse Laplace transform at each $\overset{*}{\tau}$, a global integral over the entire s-space would need to be performed. In contrast, computing the approximate inverse Laplace transformation in the form of $\mathbf{L}_k^{-1}$ for finite k requires only the real values of s and a global integral over all s need not be performed. The problem in implementing $\mathbf{L}_k^{-1}$ boils down to computing the kth derivative of the function $\mathbf{t}(\tau, s)$ with respect to s, which is a local operation. This operation can be implemented in a straightforward way.

First, note that although s is treated as a continuous variable for mathematical convenience, we have to acknowledge the discreteness of the s-axis at the cellular scale. Let us discretize and label the $\mathbf{t}$ nodes centered around $s = s_o$.

$$\ldots., s_{-3}, s_{-2}, s_{-1}, s_o, s_1, s_2, s_3, \ldots$$

The nodes are arranged to monotonically increase in s. Let the successive nodes be separated by Δ, that is, $s_i - s_{i-1} = \Delta$. In general, Δ need not be constant along the s-axis, although we will treat it as a constant for convenience. As illustrated in appendix B, we can evaluate the kth derivative of a function at the point s_o as a linear combination of the functional values at the k neighboring points. Hence, the activity of the $\mathbf{T}$ node $\overset{*}{\tau}_o = -k/s_o$ can be constructed simply from a linear combination of the activity of k neighboring nodes around s_o in the $\mathbf{t}$ column. From equations B.11 and B.12 in appendix B, note that there is a factor Δ^{-k} in the coefficients involved in constructing the kth derivative. For small values of Δ, this would be unacceptably large. Fortunately, only the relative activation of the $\mathbf{T}$ nodes has significance, and the exact magnitude of their activity is inconsequential, so we can ignore any overall constant in the connection weights between the $\mathbf{t}$ and $\mathbf{T}$ columns. For simplicity, let us just analyze the situation when k is even. From equation 2.3 and the coefficients required to construct the kth derivative (see equation B.11), the activity of the $\overset{*}{\tau}_o$ node can be written as

$$\mathbf{T}(\tau, \overset{*}{\tau}_o) = s_o^{k+1} \sum_{r=-k/2}^{k/2} \frac{(-1)^{(k/2-r)}}{(k/2-r)!\,(k/2+r)!}\, \mathbf{t}(\tau, s_o + r\Delta). \tag{2.5}$$

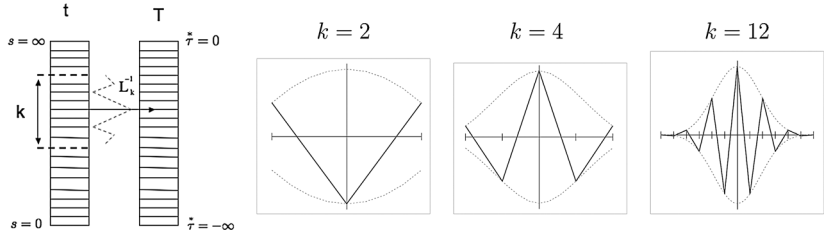

Figure 5: Neural implementation of $\mathbf{L}_k^{-1}$. The left-most panel describes the connectivity between column of **t** nodes and the column of **T** nodes. The nodes in these two columns are mapped in a one-to-one fashion. The activity of any cell in the **T** column depends on the activity of k neighbors in the **t** column. The right panel gives a pictorial representation of the connectivity function between a **T** node and its k near neighbors in the **t** column. The contributions from the neighbors alternate between excitation and inhibition as we move away from the center in either direction, and the magnitude of the contribution falls off with the distance to the neighbor. With $k = 2$, we see an off-center on-surround connectivity. With $k = 4$, we see a Mexican hat–like connectivity, and $k = 12$ shows more elaborate bands of connectivity.

In the left-most panel of Figure 5, we sketch the connections from the **t** column nodes to the $\overset{*}{\tau}_o$ node of the **T** column. The activity in the $\overset{*}{\tau}_o$ node is constructed not just from the s_o node, but from k near neighbors of the s_o node. For three different values of k, we plot these connection weights. The connection weights are positive and negative in an alternating fashion, with magnitude falling off with distance from s_o. For even k, the contributions from either side of s_o are symmetric. Interestingly, these connection weights take the form of binomial coefficients in the expansion of $(1 - x)^k$. A crucial property of these coefficients is that they sum up to 0, implementing balanced inhibition and excitation in the inputs to each **T** node. If Δ is not constant, the connection coefficients would be different from those in equation 2.5; in particular, they would be asymmetric around s_o.

In general, there are good reasons to think Δ should vary across the s-axis. First, note that for small values of s, we need an appropriately small Δ in order to accommodate at least $k/2$ neighbors between 0 and s:

$$\left(\frac{k}{2}\right)\Delta < s \;\Rightarrow\; |\Delta\overset{*}{\tau}| < 2. \tag{2.6}$$

Taking Δ to be a constant will immediately impose an upper bound $\overset{*}{\tau}_{\max}$ on the range of $\overset{*}{\tau}$, while taking Δ to vary inversely with $|\overset{*}{\tau}|$ so that $|\Delta\overset{*}{\tau}|$ remains constant does not impose any upper or lower bound on the range of $\overset{*}{\tau}$. Second, if we take the one-to-one relationship between the **t** nodes and

the **T** nodes seriously, then the distribution of Δ along the s-axis implicitly specifies the distribution of $\overset{*}{\tau}$ values:

$$\Delta \equiv \delta s = \delta(-k/\overset{*}{\tau}) = k\overset{*}{\tau}^{-2}\,\delta\overset{*}{\tau}. \tag{2.7}$$

Here $\delta\overset{*}{\tau}$ is the separation between the **T** nodes along the $\overset{*}{\tau}$ axis. If $g(\overset{*}{\tau})$ is the number density of **T** nodes along the $\overset{*}{\tau}$ axis, then it will be inversely related to $\delta\overset{*}{\tau}$. Hence,

$$g(\overset{*}{\tau}) \sim 1/\Delta|\overset{*}{\tau}|^2. \tag{2.8}$$

That is, the function controlling the distribution of Δ as a function of s also controls the distribution of $\overset{*}{\tau}$. Let us consider three simple possibilities: the choice of a constant Δ leads to $g(\overset{*}{\tau}) \sim |\overset{*}{\tau}|^{-2}$; the choice of $\Delta \sim |\overset{*}{\tau}|^{-1}$ leads to $g(\overset{*}{\tau}) \sim |\overset{*}{\tau}|^{-1}$; or the choice of $\Delta \sim |\overset{*}{\tau}|^{-2}$ leads to $g(\overset{*}{\tau}) \sim 1$.

Note that specifying the distribution of Δ also specifies the number of nodes required in a **T** column to represent the relevant timescales. The calculation is particularly straightforward for the case when $\Delta \sim |\overset{*}{\tau}|^{-1}$. In this case, the separation between neighboring **T** nodes $\delta\overset{*}{\tau} = \alpha\overset{*}{\tau}$ (see equation 2.7), where α is some constant that controls the resolution of the representation in the $\overset{*}{\tau}$-axis. If the $\overset{*}{\tau}$-axis is bounded between $\overset{*}{\tau}_{\min}$ and $\overset{*}{\tau}_{\max}$, describing the lower and upper limit of timescales to be represented, then the nodes can be arranged in the following way:

$$\overset{*}{\tau}_{\min}, \overset{*}{\tau}_{\min}(1+\alpha), \overset{*}{\tau}_{\min}(1+\alpha)^2, \ldots, \overset{*}{\tau}_{\min}(1+\alpha)^{n-1} = \overset{*}{\tau}_{\max}.$$

The number of nodes needed to represent the relevant timescale $\overset{*}{\tau}_{\min}$ to $\overset{*}{\tau}_{\max}$ is given by

$$n = 1 + \frac{\log(\overset{*}{\tau}_{\max}/\overset{*}{\tau}_{\min})}{\log(1+\alpha)}. \tag{2.9}$$

The constant α denotes the separation between neighboring nodes around $|\overset{*}{\tau}| = 1$. Choosing $\alpha = 0.1$, $|\overset{*}{\tau}_{\min}| = 0.5$ and $|\overset{*}{\tau}_{\max}| = 5000$, we obtain $n < 100$. That is, fewer than 100 nodes are required in a **T** column to represent the timescales ranging from 500 milliseconds to 5000 seconds. This estimate is very much dependent on our choice of $g(\overset{*}{\tau})$. With $g(\overset{*}{\tau}) \sim 1$, the required number of nodes would be much larger. If we choose $g(\overset{*}{\tau}) \sim |\overset{*}{\tau}|^{-2}$, which follows from choosing Δ to be constant along the s-axis, then the number of nodes required would be considerably less than that given by equation 2.9.

While the discretization of the s-axis provides a simple implementation of $\mathbf{L}_{\mathrm{k}}^{-1}$, it is not error free. As shown in appendix B, the difference between the discretized kth derivative and the functionally continuous kth derivative is of the order $O(k\Delta^2)$. When k is large, $\mathbf{L}_{\mathrm{k}}^{-1}$ is closer to the exact inverse Laplace

transform, and the **T** nodes will store a more accurate representation of the stimulus history. The downside of having a large k is twofold. First, the complexity of the connectivity function (see Figure 5) reduces the physical plausibility. Second, Δ has to be sufficiently small to keep the error in computing the kth derivative small. When Δ is small, a large number of nodes is required to span the s-axis, thus increasing the neural resource requirements.

It is important to point out that the alternating excitatory and inhibitory connections between the **t** and **T** nodes is simply a description of a functional relationship. One could imagine different ways to neurally implement this functional connectivity. For $k = 2$, we can imagine a simple mechanism wherein the **t** nodes excite the **T** nodes in a one-to-one fashion and the **T** nodes laterally inhibit their immediate neighbors in a column. We might have to invoke more complicated mechanisms for higher values of k. Since the **t** nodes cannot simultaneously be both excitatory and inhibitory, we can imagine an intermediate set of nodes (interneurons) activated by the **t** nodes, such that the **t** nodes simply provide the excitatory input to the **T** nodes while the interneurons provide the inhibitory input to the **T** nodes. The exact mechanism that leads to such alternating excitatory-inhibitory connectivity has to be constrained based on neurophysiological considerations.

2.3 Comparison to Time Cells Constructed from Tapped Delay Lines. It can be shown (see appendix A) that the evolving pattern of activity in the different $\overset{*}{\tau}$ nodes in each **T** column satisfies the following differential equation:

$$\left(\frac{\overset{*}{\tau}}{k}\right)\frac{\partial^2 \mathbf{T}}{\partial \overset{*}{\tau}\partial \tau} - \frac{\partial \mathbf{T}}{\partial \overset{*}{\tau}} + \left(\frac{k+1}{k}\right)\frac{\partial \mathbf{T}}{\partial \tau} = 0. \tag{2.10}$$

This equation, along with the boundary condition $\mathbf{T}(\tau, 0) = \mathbf{f}(\tau)$, is sufficient to completely determine the evolving pattern of activity in the **T** nodes. Note that equation 2.10 does not make reference to **t** or to $\mathbf{L}_k^{-1}$. This leaves open the possibility of constructing the representation **T** through alternative means.

The nodes of **T** behave like time cells, increasing their activity a characteristic time after the occurrence of the stimulus that caused the activation. Tapped delay lines are an alternative means to construct time cells. In tapped delay lines, a chain of connected nodes provides input to one another. Each node responds and triggers the next node in the chain after a characteristic delay following the input it receives. As a consequence, an input at one end of the chain gradually traverses through each node to the other end. As the input traverses through the chain, the delays get accumulated, and consequently each node in the chain responds at different latencies following the

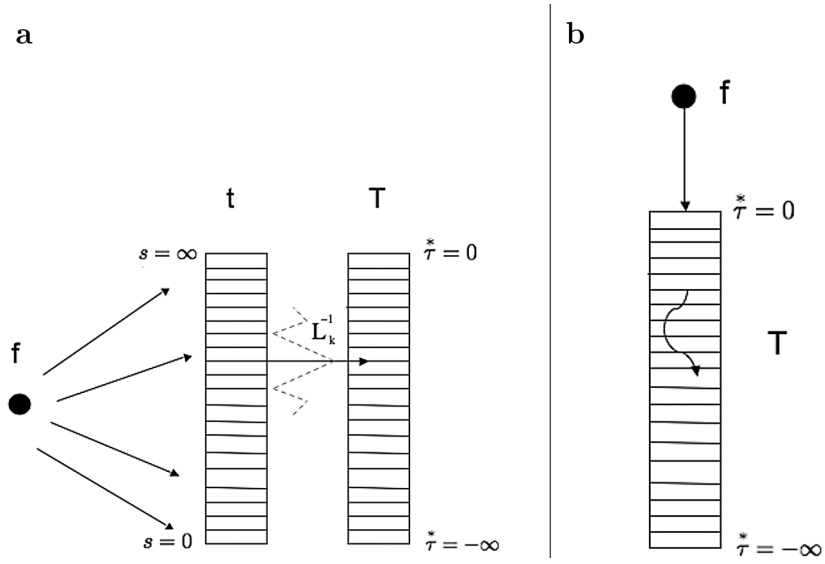

Figure 6: (a) The time cells in the **T** column are constructed from the activity of the leaky integrators in the **t** column through the alternating excitatory-inhibitory connections of the $\mathbf{L}_k^{-1}$ operator. (b) Alternatively, ignore the **t** column and simply consider the stimulus to directly activate the top node in the **T** column. If the **T** column nodes are chained together in an appropriately constructed tapped delay line, then the activity could "trickle down" the **T** column according to equation 2.10.

input. The nodes in a tapped delay line would exhibit similar properties as time cells.

Hence, as described in Figure 6b, we could alternatively view the **T** column as a tapped delay line where the input at any moment $\mathbf{f}(\tau)$ is given to the $\overset{*}{\tau} = 0$ node and the activity gradually progresses to the other $\overset{*}{\tau}$ nodes in a way that enables it to obey equation 2.10. That is, the **T** column activity can be constructed without the leaky integrators in the **t** column if the successive $\overset{*}{\tau}$ nodes of the **T** column are appropriately chained such that equation 2.10 could be implemented.

The procedure of constructing the time cells from the leaky integrators through the operator $\mathbf{L}_k^{-1}$ yields some qualitatively different features from those that would result if it were constructed using tapped delay lines. Whereas the range of $\overset{*}{\tau}$ values that can be achieved with TILT depends on the distribution of time constants of the leaky integrators, the range of $\overset{*}{\tau}$ values that could be achieved with a tapped delay line scheme depends on the range of characteristic latencies the cells exhibit, combined with the architecture of the network. The units in a tapped delay line architecture

would respond to an impulse of input after some characteristic delay. The upper limit on the value of $\overset{*}{\tau}$ that can be achieved using TILT is determined by the largest time constant. The upper limit on $\overset{*}{\tau}$ that can be generated using tapped delay lines is determined by the longest characteristic delay multiplied by the number of links in the chain. If the largest characteristic delay is much shorter than the largest value of $\overset{*}{\tau}$ to be represented, this places strong constraints on a tapped delay line architecture. Under these circumstances, a tapped delay line scheme would require that the spacing of $\overset{*}{\tau}$ values be constant. That is, for a tapped delay line scheme in which the largest characteristic delay is short, $g(\overset{*}{\tau}) = 1$. As discussed above, this would require a relatively large amount of resources. Of course, this would not be a problem if it were possible to construct units with characteristic delays on the order of the largest value of $\overset{*}{\tau}$.

In addition to differences due to capacity constraints, the two approaches to constructing time cells differ in the connectivity between cells. When using tapped delay lines, if the largest characteristic delay is much shorter than the largest value of $\overset{*}{\tau}$, there is strong dependence between the units representing smaller values of $\overset{*}{\tau}$ and larger values of $\overset{*}{\tau}$. As a consequence, disrupting one of the nodes in the chain would tend to disrupt the time cells responding at all subsequent latencies. If there are many links in the chain required to reach the largest value of $\overset{*}{\tau}$, then the system would be very sensitive to failures at small values of $\overset{*}{\tau}$. In contrast, in TILT, there is no direct causal connection between the activity of the **T** nodes representing different values of $\overset{*}{\tau}$. The sequential firing as a function of $\overset{*}{\tau}$ is a consequence of a common dependence on the state of **t** rather than a causal connection between the **T** nodes with small values of $\overset{*}{\tau}$ and large values of $\overset{*}{\tau}$. However, disruption of a **t** node would disrupt the k near neighbors in **T**. This disruption is similar regardless of the value s that is disrupted and is not particularly sensitive to values of s corresponding to small values of $\overset{*}{\tau}$.

2.4 Associative Learning. Up to this point, we have focused on the properties of a single dimension of the stimulus space, corresponding to a single column of **t** and a single column of **T**. Here we describe a simple associative memory that enables **T** to serve as a cue for behavior. This treatment differs from the development up to this point because it requires us to consider all dimensions of the stimulus space and the entire **t** layer and the **T** layer. This is so because we will have to sum over the entire **T** layer to construct a prediction in the stimulus space.

Consider Figure 1b. There are many **f** nodes, grouped together as the **f** layer. Each **f** node provides input to a **t** column; the **t** columns are grouped together as the **t** layer. Each **f** node also corresponds to a **T** column; the **T** columns are grouped together as the **T** layer. Each node in the **f** layer corresponds to a dimension of the stimulus space. Each distinguishable stimulus will have a unique representation in the **f** layer. That is, each

stimulus corresponds to a unique pattern of activation across the nodes of the **f** layer. The exact representation of a stimulus in the **f** layer will depend on its perceptual properties. Perceptually similar stimuli, like a picture of an apple and a picture of a peach, would have highly overlapping representations, while perceptually dissimilar stimuli, like a picture of an apple and the sound of a bird chirping, would have less overlapping representations. In general, with more **f** nodes, more distinguishable stimuli can be represented. Because our interest in this letter is in describing properties of timing and not stimulus discrimination or stimulus generalization, we will consider an idealization where an experimentally relevant stimulus is represented by a unique node in the **f** layer. In the simple experiments we are considering here (e.g., delay conditioning between a tone and a shock), this seems to be a perfectly reasonable assumption.

At any instant, the **T**-layer activity holds the approximate timing information about the presentation history of all stimuli presented up to that instant. A simple way to use this information to learn the temporal relationships between stimuli is to introduce associative learning between the **T** and **f** layers. We define the operator **M** to hold the synaptic weights connecting the **T**-layer nodes to the **f**-layer nodes (see Figure 1b). The weights change at each instant in a Hebbian fashion. It is convenient to represent **M** as an outer product association between the **f** and **T** layers. We shall use the bra-ket notation for this purpose, where $|\cdot\rangle$ represents a tensor, $\langle\cdot|$ represents its transpose, $\langle\cdot|\cdot\rangle$ represents the inner-product of two equi-ranked tensors, and $|\cdot\rangle\langle\cdot|$ denotes the outer product of two tensors:

$$\mathbf{M}(\tau) = \int_{-\infty}^{\tau} |\mathbf{f}(\tau')\rangle\langle\mathbf{T}(\tau')|\, d\tau'. \tag{2.11}$$

At any moment τ, $|\mathbf{f}(\tau)\rangle$ is of rank 1 and $\langle\mathbf{T}(\tau)|$ of rank 2 because the **T** layer consists of the dimensions corresponding to the stimulus space and $\overset{*}{\tau}$. Hence, **M** is a tensor of rank 3. In an indicial notation, the components of **M** at any time τ can be written as $M_i^{j\overset{*}{\tau}}(\tau)$, where i indexes the stimulus dimension in **f** layer and j indexes the stimulus dimension in the **T** layer.

The simple Hebbian learning rule we use here states that every time a stimulus is presented, the synaptic weight between any given **T** node and the associated **f** node is increased by an amount equal to the instantaneous activity of that **T** node. With learning, the activity in the **T** layer can induce activity in the **f** layer through the associations learned in **M**. We shall refer to this internally generated activity in the stimulus layer as **p**, the prediction for the imminent future. Coarsely, this can be interpreted as "what comes to mind next." In the bra-ket notation, this is represented as

$$|\mathbf{p}(\tau)\rangle = \mathbf{M}(\tau)|\mathbf{T}(\tau)\rangle. \tag{2.12}$$

To explicate this in the indicial notation, let us denote the activity of the $\overset{*}{\tau}$ node in the jth column of the **T** layer by $\mathrm{T}^j(\tau, \overset{*}{\tau})$ at any time τ. The activity induced by $\mathrm{T}^j(\tau, \overset{*}{\tau})$ in the ith node in the **f** layer through the synaptic weights **M** will be

$$\mathcal{P}_i^{j\overset{*}{\tau}}(\tau) = \mathrm{M}_i^{j\overset{*}{\tau}}(\tau)\,\mathrm{T}^j(\tau, \overset{*}{\tau}). \tag{2.13}$$

The overall activity induced in the ith node of the **f** layer by the **T** layer is simply taken to be the sum of the individual contributions from each **T** node:

$$\mathrm{p}_i = \sum_j \int \mathcal{P}_i^{j\overset{*}{\tau}}(\tau)\, g(\overset{*}{\tau})\, d\overset{*}{\tau}. \tag{2.14}$$

Since we treat $\overset{*}{\tau}$ as a continuous variable, the contributions from different $\overset{*}{\tau}$ nodes are added up by integrating them along with the number density of nodes $g(\overset{*}{\tau})$. The function $g(\overset{*}{\tau})$ simply denotes the number of nodes in the **T** column that fall within a unit interval on the $\overset{*}{\tau}$-axis. Taken together, the components p_i form the prediction vector **p** defined in equation 2.12.

To illustrate how **M** enables associative learning, let A be a stimulus that has been encountered only once at time τ_A, and let A activate just one node in the **f** layer. The **T**-layer activity at that moment, $|\mathbf{T}(\tau_A)\rangle$, will be stored in **M** as the synaptic weights connecting the **T** layer to the A node in the **f** layer according to equation 2.11. At a later time τ_o, the **T**-layer activity would have changed to $|\mathbf{T}(\tau_o)\rangle$, which would depend on the duration between τ_A and τ_o, as well as the stimuli presented between these times. To the extent $|\mathbf{T}(\tau_o)\rangle$ overlaps with $|\mathbf{T}(\tau_A)\rangle$, the A node in the **f** layer will be internally activated at the moment τ_o. In this simple situation where A occurred only once, the A component of the prediction **p** is simply the inner product between $|\mathbf{T}(\tau_A)\rangle$ and $|\mathbf{T}(\tau_o)\rangle$:

$$\mathrm{p}_A = \langle \mathbf{T}(\tau_A) | \mathbf{T}(\tau_o) \rangle = \sum_j \int \mathrm{T}^j(\tau_A, \overset{*}{\tau})\, \mathrm{T}^j(\tau_o, \overset{*}{\tau})\, g(\overset{*}{\tau})\, d\overset{*}{\tau}. \tag{2.15}$$

We propose that in simple behavioral tasks, **p** drives the behavior through appropriate behavioral mechanisms. Note that **p** strongly depends on $g(\overset{*}{\tau})$. In all the behavioral applications in this letter, we will choose $g(\overset{*}{\tau}) = 1$. But the obtained behavioral results can equivalently be obtained from any other choice of $g(\overset{*}{\tau})$ if we were to modify the definition of $\mathbf{L}_{\mathrm{k}}^{-1}$ (see equation 2.3) appropriately. To see this, note from equation 2.15 that it is the product of $g(\overset{*}{\tau})$ with $\mathbf{T}(\tau, \overset{*}{\tau})$, which is relevant in computing the prediction **p**. By appropriately absorbing the functional form of $g(\overset{*}{\tau})$ into $\mathbf{T}(\tau, \overset{*}{\tau})$ by modifying the definition of $\mathbf{L}_{\mathrm{k}}^{-1}$, we can obtain the same prediction as that

obtained with $g(\overset{*}{\tau}) = 1$ and $\mathbf{T}(\tau, \overset{*}{\tau})$ defined as equation 2.3. Although the functional form of $\mathbf{T}(\tau, \overset{*}{\tau})$ and $g(\overset{*}{\tau})$ are not separately identifiable mathematically in constructing the behavioral predictions, note that $g(\overset{*}{\tau})$ is a physical quantity. If we could measure the responsiveness of every neuron in the brain during performance of an appropriate task, then the form of $g(\overset{*}{\tau})$ could be directly measured.

3 Emergence of Scale Invariance

As shown by equation 2.4, $\mathbf{T}(\tau, \overset{*}{\tau})$ approximately represents the stimulus presentation history as activity distributed across the different $\overset{*}{\tau}$ nodes. We now demonstrate the scale invariance of this representation. Consider a delta function stimulus that activates a single dimension of the stimulus space at time $\tau = 0$. The activity of the corresponding $\mathbf{t}$ and $\mathbf{T}$ columns at any instant τ following the presentation is given by equations 2.2 and 2.3. The function $\mathbf{t}(\tau, s)$ and its kth derivative are calculated to be

$$\mathbf{f}(\tau) = \delta(\tau - 0) \Rightarrow \mathbf{t}(\tau, s) = e^{-s\tau} \Rightarrow \mathbf{t}^{(k)}(\tau, s) = (-\tau)^k e^{-s\tau}. \tag{3.1}$$

Then we use equation 2.3 to evaluate $\mathbf{T}(\tau, \overset{*}{\tau})$ and express it in two forms:

$$\mathbf{T}(\tau, \overset{*}{\tau}) = \frac{1}{\tau} \frac{(k)^{k+1}}{k!} \left(\frac{-\tau}{\overset{*}{\tau}}\right)^{k+1} e^{k\left(\frac{\tau}{\overset{*}{\tau}}\right)} \tag{3.2}$$

$$= -\frac{1}{\overset{*}{\tau}} \frac{(k)^{k+1}}{k!} \left(\frac{-\tau}{\overset{*}{\tau}}\right)^{k} e^{k\left(\frac{\tau}{\overset{*}{\tau}}\right)}. \tag{3.3}$$

Note that τ is positive because we are interested only in the time following the stimulus presentation, while $\overset{*}{\tau}$ is negative. This makes the ratio $(\tau/\overset{*}{\tau})$ negative. Equations 3.2 and 3.3 are plotted simultaneously with respect to $\overset{*}{\tau}$ and τ in Figure 7. Restricting the attention to one horizontal line in the figure gives the activity of a specific $\overset{*}{\tau}$ node as a function of time, while restricting the attention to a vertical line gives the instantaneous activity of the entire $\mathbf{T}$ column. The scalar property is apparent from the functional form where τ and $\overset{*}{\tau}$ occur as a ratio. Hence, no specific scale is being represented by the $\mathbf{T}$-layer activity. To quantify the underlying scale invariance, observe the following analytical properties.[1]

[1] All integrals here can be calculated using the form $\int_0^\infty y^m e^{-ky} dy = \frac{\Gamma(m+1)}{k^{m+1}}$. For integer values of m, $\Gamma(m+1) = m!$

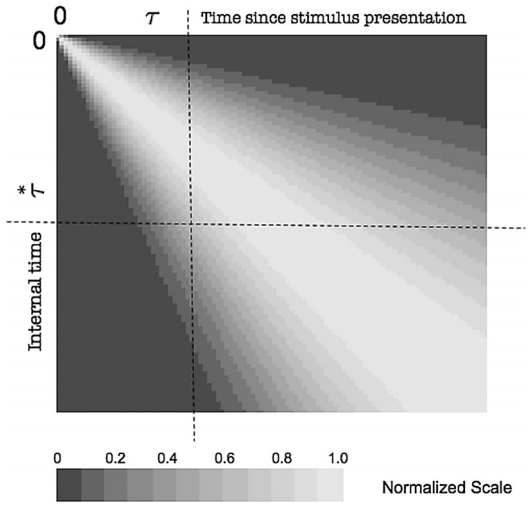

Figure 7: A stimulus is presented at $\tau = 0$. The activity of an entire **T** column is plotted as a function of time. Restricting the focus on the horizontal line reveals the activity of a single node in the **T** column as a function of time, while restricting the focus on the vertical line reveals the entire **T** column activity at any given moment. The activity is normalized about each horizontal line according to the scale shown below the graph. (See the online supplement for a color version of this figure, available at http://www.mitpressjournals.org/doi/suppl/10.1162/NECO_a_00212.)

Let us first look along the vertical line in Figure 7 and consider the distribution of activity over $\overset{*}{\tau}$ at a particular instant τ. The area under the distribution $\mathbf{T}(\tau, \overset{*}{\tau})$ is independent of τ:

$$\int \mathbf{T}(\tau, \overset{*}{\tau})\, d\overset{*}{\tau} = 1. \tag{3.4}$$

The area is 1 because the stimulus was chosen to be a delta function. For any other stimulus function, the area would match that under the stimulus function plotted over time. The peak of the distribution is at

$$\frac{d}{d\overset{*}{\tau}} \mathbf{T}(\tau, \overset{*}{\tau}) = 0 \Rightarrow \overset{*}{\tau} = -\tau[k/(k+1)]. \tag{3.5}$$

When $k > 1$, the mean of the distribution is at

$$\overset{*}{\tau}_{mean} = \int \overset{*}{\tau}\, \mathbf{T}(\tau, \overset{*}{\tau})\, d\overset{*}{\tau} = -\tau[k/(k-1)]. \tag{3.6}$$

When $k > 2$, the variance of this distribution is

$$\int (\overset{*}{\tau} - \overset{*}{\tau}_{mean})^2 \, \mathbf{T}(\tau, \overset{*}{\tau}) \, d\overset{*}{\tau} = \tau^2 [k^2/(k-2)(k-1)^2]. \tag{3.7}$$

The function $\mathbf{T}(\tau, \overset{*}{\tau})$ should not be misunderstood to be a stochastic variable just because its mean and variance are calculated; it is a deterministic function. The coefficient of variation (CV), defined as the ratio of the standard deviation to the mean of the distribution, is then

$$\mathrm{CV} = (k-2)^{-1/2}. \tag{3.8}$$

Note that the CV calculated over $\overset{*}{\tau}$ is a constant for all values of τ. This is a manifestation of the scale invariance of the distribution. Though the above expression is not valid when k is either 1 or 2, the CV can be shown to be independent of τ. As we shift the vertical line in Figure 7 rightward and look at the $\mathbf{T}$ activity at a later time, the distribution becomes wider but the CV remains the same.

Now let us look along the horizontal line in Figure 7 and consider the activity of a specific $\overset{*}{\tau}$ node as a distribution over real time τ. The area under this distribution is also a constant for all values of k and $\overset{*}{\tau}$:

$$\int \mathbf{T}(\tau, \overset{*}{\tau}) \, d\tau = 1. \tag{3.9}$$

Again, the area turns out to be 1 because the stimulus was chosen to be a delta function. The peak of this distribution is at

$$\frac{d}{d\tau} \mathbf{T}(\tau, \overset{*}{\tau}) = 0 \qquad \Rightarrow \qquad \tau = -\overset{*}{\tau}, \tag{3.10}$$

and the mean of the distribution is at

$$\tau_{mean} = \int \tau \, \mathbf{T}(\tau, \overset{*}{\tau}) \, d\tau = -\overset{*}{\tau}[(k+1)/k]. \tag{3.11}$$

The variance of this distribution is

$$\int (\tau - \tau_{mean})^2 \, \mathbf{T}(\tau, \overset{*}{\tau}) \, d\tau = \overset{*}{\tau}^2 [(k+1)/k^2]. \tag{3.12}$$

The coefficient of variation can then be calculated as

$$\mathrm{CV} = (k+1)^{-1/2}. \tag{3.13}$$

Note that the CV calculated over the values of τ is a constant for all values of $\overset{*}{\tau}$, manifesting scale invariance. As we shift the horizontal line in Figure 7 downward, the nodes with a larger (negative) $\overset{*}{\tau}$ respond later with a larger temporal spread, but the CV remains constant.

The scalar property underlying the function $\mathbf{T}(\tau, \overset{*}{\tau})$ is also reflected in the temporal distribution of the prediction activity **p** generated in the **f** layer. This is because **p** is constructed from the inner product of the **T**-layer activities from different instants, as explained in equations 2.12 and 2.15. To see this, consider a stimulus *start* followed by a stimulus *stop* after a time τ_o. For simplicity, let the *start* and *stop* be represented by two distinct **f** nodes. At the moment when *stop* is active in the **f** layer, the **T**-layer activity $\mathbf{T}(\tau_o, \overset{*}{\tau})$ is given by equation 3.2 (with τ_o replacing τ). This will be stored in **M** according to equation 2.11. If the *start* stimulus is later repeated, the **T** activity following it after a time τ will again be $\mathbf{T}(\tau, \overset{*}{\tau})$ in equation 3.2. This will induce an activity in the *stop* node of the **f** layer according to equations 2.12 and 2.15. This will be the *stop* component of the prediction **p**, denoted by $\mathrm{p}_{stop}(\tau)$. Since the only relevant predictive stimulus in this scenario is *start*, we only need to consider the column corresponding to *start* when summing over the columns in equation 2.15:

$$\mathrm{p}_{stop} = \int \mathrm{T}(\tau_o, \overset{*}{\tau})\, \mathrm{T}(\tau, \overset{*}{\tau})\, g(\overset{*}{\tau})\, d\overset{*}{\tau}, \tag{3.14}$$

$$\mathrm{p}_{stop} = \int \frac{1}{\overset{*}{\tau}^2} \frac{(k)^{2k+2}}{(k!)^2} \left(\frac{\tau_o \tau}{\overset{*}{\tau}^2}\right)^k e^{k\left(\frac{\tau_o}{\overset{*}{\tau}}\right)} e^{k\left(\frac{\tau}{\overset{*}{\tau}}\right)}\, g(\overset{*}{\tau})\, d\overset{*}{\tau}. \tag{3.15}$$

Note that τ and τ_o are positive while $\overset{*}{\tau}$ is negative. We define $x \equiv -\tau_o/\overset{*}{\tau}$ and $y \equiv \tau/\tau_o$. The integral over $\overset{*}{\tau}$ can be converted to an integral over x, since $d\overset{*}{\tau}/\overset{*}{\tau}^2 = dx/\tau_o$:

$$\mathrm{p}_{stop} = \int \frac{1}{\tau_o} \frac{(k)^{2k+2}}{(k!)^2} x^{2k} y^k e^{-kx(1+y)}\, g(\overset{*}{\tau})\, dx. \tag{3.16}$$

For the simple case when $g(\overset{*}{\tau}) = 1$, the integral can be calculated as

$$\begin{aligned} \mathrm{p}_{stop} &= \frac{1}{\tau_o} \frac{(k)^{2k+2}}{(k!)^2} \frac{y^k (2k)!}{(k(1+y))^{2k+1}} \\ &= \frac{1}{\tau_o} \frac{(2k)! k}{(k!)^2} \frac{y^k}{(1+y)^{2k+1}}. \end{aligned} \tag{3.17}$$

Note that the only dependence of time τ is through dependence on y. This immediately leads to scale invariance. It turns out that the area under the

distribution is a constant for all values of τ_o. For the delta function stimulus considered here,

$$\int \mathrm{p}_{stop}(\tau)\, d\tau = 1. \tag{3.18}$$

The peak of the distribution is at

$$\frac{d}{dy}\left(\frac{y^k}{(1+y)^{2k+1}}\right) = 0 \implies y = \frac{k}{k+1} \implies \tau = \tau_o[k/(k+1)]. \tag{3.19}$$

For $k > 1$, the mean of the distribution is at

$$\tau_{mean} = \int \tau\, \mathrm{p}_{stop}(\tau)\, d\tau = \tau_o[(k+1)/(k-1)]. \tag{3.20}$$

For $k > 2$, the variance of the distribution is

$$\int (\tau - \tau_{mean})^2\, \mathrm{p}_{stop}(\tau)\, d\tau = (\tau_o)^2[2k(k+1)/(k-2)(k-1)^2]. \tag{3.21}$$

Again, it should be noted that $\mathrm{p}_{stop}(\tau)$ is not a stochastic variable even though we calculate its mean and variance; it is a deterministic function. The coefficient of variation is a constant for all values of τ_o:

$$\mathrm{CV} = [2k/(k+1)(k-2)]^{1/2}. \tag{3.22}$$

Although the above expression is not valid when k is 1 or 2, the CV can be shown to be independent of τ_o. The timing information contained in the $\mathrm{p}_{stop}(\tau)$ distribution can be utilized by task-dependent behavioral mechanisms to generate well-timed response distributions. The CV of the p_{stop} distribution given by equation 3.22 should not be confused with the CV of the behavioral response distributions, which in general will depend on a variety of factors, including the task demands.

A simplifying assumption used in the above calculations is that $g(\overset{*}{\tau}) = 1$. More generally, when $g(\overset{*}{\tau})$ is a power law function of $\overset{*}{\tau}$, the integral over x in equation 3.16 can be evaluated by factoring out a power of τ_o, and the resulting distribution $\mathrm{p}_{stop}(\tau)$ will be purely a function of y. So we conclude that any power law form for $g(\overset{*}{\tau})$, including of course $g(\overset{*}{\tau}) = 1$, will lead to scale-invariant $\mathrm{p}_{stop}(\tau)$, but the mean, standard deviation, CV, area under the distribution, and the position of the peak will be different from the values calculated above.

4 Effect of Noise

The essence of this model lies in the transcription of the real-time stimulus layer activity into a fuzzy representation of stimulus history in the **T** layer. It is important to have a qualitative understanding of the effect of noise in this transcription process. Note that every step in this transcription is a linear operation, so the effect of noise in each stage can be analyzed in a straightforward way. One crucial advantage of working with a linear system is that the noise and signal can be decoupled and independently analyzed. Let us now consider noise in each component of the system separately and analyze its effect on the **T**-layer activity.

We first show that any noise in the **f** layer will be transmitted into the **T** layer without getting amplified. Then we consider noise in the individual nodes of the **t** layer and show that its effect is amplified in the **T** layer but suppressed in generating the prediction **p**. Finally, we discuss the effect of noise in the connection weights of the $\mathbf{L}_k^{-1}$ operator.

4.1 Noise in the f and t Layers. If the noise in the **t** layer is purely a consequence of the noise in the **f** layer, then it goes into the **T** layer unamplified. To see this, we start with the most general noise form in the **t** layer and note that any noise in the **f** layer has to pass through the **t** layer before being transmitted to the **T** layer. Since each node in the **t** layer has a fixed exponential decay rate, the most general form of **t**-layer activity that includes noise can be obtained by rewriting equation 2.2 in the following way:

$$\mathbf{t}(\tau, s) = \int_{-\infty}^{\tau} \left[\mathbf{f}(\tau') + N(\tau', s)\right] e^{s(\tau'-\tau)}\, d\tau'. \tag{4.1}$$

Here $N(\tau', s)$ is an arbitrary noise function. If the noise in the **t** layer is simply a consequence of noise in the **f** layer, then $N(\tau', s)$ will be independent of s. This is because the noise in the **f** layer will be equally transmitted to all the rows of the **t** layer. In such a situation, the noise across any **t** column will be completely correlated. It is straightforward to understand the effect of such correlated noise in the **T**-layer activity. Since $\mathbf{L}_k^{-1}$ is a linear operator that transforms the activity from the **t** column into activity in the **T** column, the signal and noise will be independently coded into the **T** column. The signal-to-noise ratio in the **T** column will be exactly the same as that in the **f**-layer input. In other words, correlated noise in **f** neither gets amplified nor suppressed as it is transmitted to **T**.

4.2 Uncorrelated Noise in the t Layer. We now consider random uncorrelated noise in individual **t** nodes and compare its effect on the **T** nodes to the effect that would result from an appropriate control signal. The appropriate control signal in this case is the effect of a stimulus that gives rise

to the same amplitude of activation in the **t** node as the noise we are considering. It turns out that such noise is highly amplified relative to a control signal while being transmitted to the **T** layer, but significantly suppressed relative to a control signal in generating **p**.

*4.2.1 Effect of Uncorrelated Noise in the **t** Layer on the **T** Layer.* Uncorrelated noise injected into the **t** layer could be caused by sources that randomly activate the **t** nodes. Computationally understanding the effect of uncorrelated noise amounts to considering the effect of the s-dependence of the function $N(\tau', s)$ in equation 4.3 on **T** and the prediction it generates. The effect of taking the kth derivative of the function $N(\tau', s)$ with respect to s could drastically amplify the noise. To analyze this, let us consider the worst-case scenario, where $N(\tau', s)$ is sharply peaked at $s = s_o$ and $\tau' = 0$, making its kth derivative diverge:

$$N(\tau', s) = \delta(s - s_o)\, \delta(\tau' - 0). \tag{4.2}$$

Following this noise pulse at time 0, the activity of the s_o node will exponentially decay from 1. The effect of this noise pulse on the **T** layer will be maximal at the moment the pulse is delivered. At that moment, the activity from the s_o node of the **t** column will be transferred to the nodes in the **T** column that are neighbors of $\overset{*}{\tau}_o = -k/s_o$. Let us index the neighbors of $\overset{*}{\tau}_o$ by m varying from $-k/2$ to $+k/2$. The activity of these nodes due to the noise pulse in the s_o node can be inferred from equation 2.5 to be

$$\mathrm{T}_{noise} = (s_o + m\Delta)^{k+1} \frac{(-1)^{(k/2-m)}}{(k/2+m)!(k/2-m)!}. \tag{4.3}$$

In order to compare the effect of this noise pulse to an effect generated by a real signal, we should consider an **f**-layer activity that would generate a comparable activity in the s_o node of the **t** column. The appropriate control signal is a delta function stimulus of strength e^k presented at a time $\tau = \overset{*}{\tau}_o$ in the past. This control signal will induce an activity of magnitude 1 in the s_o node at $\tau = 0$, exactly as the noise pulse. But unlike the noise pulse that activates just the s_o node, the control signal also activates the **t** nodes in the neighborhood of s_o. Labeling the neighboring nodes by m, their activity at $\tau = 0$ will be $e^{m(\Delta\overset{*}{\tau}_o)}$. In response to this control signal, the activity of the **T** column nodes in the neighborhood of $\overset{*}{\tau}_o$ due to the signal is

$$\begin{aligned}\mathrm{T}_{signal} &= (s_o + m\Delta)^{k+1} e^{m(\Delta\overset{*}{\tau}_o)} \sum_{r=-k/2}^{+k/2} \frac{(-1)^{(k/2-r)}}{(k/2-r)!(k/2+r)!} e^{r(\Delta\overset{*}{\tau}_o)} \\ &= \frac{(s_o + m\Delta)^{k+1}}{k!} e^{(m-k/2)(\Delta\overset{*}{\tau}_o)} \left[1 - e^{(\Delta\overset{*}{\tau}_o)}\right]^k. \end{aligned} \tag{4.4}$$

The representation of the noise pulse in the **T** layer given by equation 4.3 can be readily compared to the representation of the control signal given by equation 4.4. When $(\Delta\overset{*}{\tau}_o)$ is small, the term $[1 - e^{(\Delta\overset{*}{\tau}_o)}]$ can be approximated by $(\Delta\overset{*}{\tau}_o)$. The ratio of noise to control signal in the **T** layer is approximately given by

$$\left|\frac{\mathrm{T}_{noise}}{\mathrm{T}_{signal}}\right| \simeq \frac{k!}{(k/2!)^2}(\Delta\overset{*}{\tau}_o)^{-k}. \tag{4.5}$$

For very small values of $(\Delta\overset{*}{\tau}_o)$, the effect of noise is very large. This comes as no surprise because we have, in effect, differentiated a singular function. In general, choosing a smooth $N(\tau', s)$ will lead to a smooth representation in the **T** layer, and its size will be directly related to the size of the kth derivative of $N(\tau', s)$.

4.2.2 Effect of Uncorrelated Noise in the ***t*** *Layer on the Prediction.* Although a singular noise pulse in the **t** layer induces a large effect in the **T** layer, it turns out to have a much more modest effect on the prediction **p**. This is because **p** is constructed not just from the activity of the $\overset{*}{\tau}_o$ node, but also from the activity of its neighbors. Since the connection weights from the s_o node to the neighbors of $\overset{*}{\tau}_o$ node are alternatively positive and negative, the activity of the neighbors of $\overset{*}{\tau}_o$ node due to the noise pulse is positive and negative in an alternating fashion, as can be seen from equation 4.3. Since the positive and negative connection weights are well balanced, the cumulative effect of the noise on the neighbors of the $\overset{*}{\tau}_o$ node in generating the prediction will be very small.

To illustrate this, we consider a suitable set of connection weights in **M**, such that when the nodes in the **T** column around $\overset{*}{\tau}_o$ are activated, a prediction will be generated. A simple and reasonable choice is to consider the connection weights in **M** to be that induced by the **T**-layer activity of the control signal at $\tau = 0$. When the noise pulse is injected in the s_o node, the activity induced in the k neighbors of the $\overset{*}{\tau}_o$ node will generate a prediction $\mathrm{P}_{\mathrm{noise}}$ due to the presence of nonzero weights in **M**:

$$\begin{aligned}\mathrm{P}_{noise} = {} & \frac{[1 - e^{\Delta\overset{*}{\tau}_o}]^k}{(k!)^2} \sum_{m=-k/2}^{k/2} (s_o + m\Delta)^{2(k+1)} \\ & \times \frac{(-1)^{(k/2-m)}k!}{(k/2+m)!\,(k/2-m)!} e^{(m-k/2)(\Delta\overset{*}{\tau}_o)}.\end{aligned} \tag{4.6}$$

As a comparison, let us consider the prediction induced by the control signal. As opposed to the noise, the activity induced by the control signal is not confined to just the k neighbors of the $\overset{*}{\tau}_o$ node in the **T** column.

But for a fair comparison, we restrict our focus to the prediction generated exclusively by the activity of the $\overset{*}{\tau}_o$ node and its k neighbors:

$$\mathrm{P}_{signal} = \frac{[1 - e^{\Delta \overset{*}{\tau}_o}]^{2k}}{(k!)^2} \sum_{m=-k/2}^{k/2} (s_o + m\Delta)^{2(k+1)} e^{2(m-k/2)(\Delta \overset{*}{\tau}_o)}. \tag{4.7}$$

Note that the net prediction from all the active **T** nodes will be significantly higher than equation 4.7 because of contributions from nodes beyond $m = -k/2$ to $+k/2$.

In the limit when $|\Delta \overset{*}{\tau}_o|$ is very small, explicit summation can be performed retaining just the lowest power of $|\Delta \overset{*}{\tau}_o|$. It turns out that[2]

$$\mathrm{P}_{noise} = \frac{(\Delta \overset{*}{\tau}_o)^{2k}}{(k!)^2} s_o^{2(k+1)} c_k \; + \; \mathcal{O}(|\Delta \overset{*}{\tau}_o|^{2k+1}),$$

$$\mathrm{P}_{signal} = \frac{(\Delta \overset{*}{\tau}_o)^{2k}}{(k!)^2} s_o^{2(k+1)} (k+1) \; + \; \mathcal{O}(|\Delta \overset{*}{\tau}_o|^{2k+1}).$$

Here c_k is just the coefficient of $(\Delta \overset{*}{\tau}_o)^{2k}$ in the Taylor expansion of the right-hand side of equation 4.6. The magnitude of c_k can be shown to be less than 1 for all $k > 2$ and $c_2 = 2.5$. Overall, for $k > 2$, we can constrain the noise level in comparison with the control signal level in the following way:

$$\left| \frac{\mathrm{P}_{noise}}{\mathrm{P}_{signal}} \right| < \frac{1}{k+1}. \tag{4.8}$$

This implies the noise level in the prediction **p** is suppressed for higher values of k.

To summarize, when uncorrelated noise is introduced in the **t** layer, it could induce large effects in the **T** layer, but this noise is drastically suppressed while generating the prediction **p**. This implies that the system is very resistant to noise at the level of behavioral predictions. The result that a pulse of uncorrelated noise in the input is actually suppressed in its effects on the prediction is reassuring, but there are a couple of points that should be noted. While the noise pulse in our derivation is restricted to a single value of s_o, the signal affects all values of s. At first glance, this might appear to be an unfair comparison. But note that as additional noise pulses are included in the k neighbors of s_o, the magnitude of their sum in

[2] It is easy to see that the lowest power of $|\Delta \overset{*}{\tau}_o|$ in P_{signal} is $2k$. But to see that its lowest power in P_{noise} is $2k$, note the following property of binomial coefficients $\sum_{r=0}^{r=k} (-1)^r \binom{k}{r} r^q = 0$ for all $q < k$.

T will tend to go down; the derivation above maximizes the value of the *k*th derivative of the noise.

4.3 Noise in the Connection Weights of $\mathbf{L}_k^{-1}$. A precise analysis of the effect of noise in the connection weights requires a specific form for time-dependent and time-independent noise in the $\mathbf{L}_k^{-1}$ weights. The form of the noise should follow from a specific mapping between the abstract $\mathbf{L}_k^{-1}$ and a physical implementation. A variety of such mappings seems plausible at this time. Rather than committing to one or the other of these and then analyzing its properties in detail, we make some general observations.

Any noise in the connection weights of $\mathbf{L}_k^{-1}$ will affect the **T**-layer activity and the prediction **p**. The excitatory-inhibitory connections between the **t** and **T** layers (see equation 2.5) are finely tuned with the property that the sum of all the connection strengths between any **T** node and k neighboring **t** nodes is precisely 0. Any fluctuation in this balance will proportionally produce an activity in the **T** layer because of the linearity of equation 2.5. Depending on the fluctuation, the function $\mathbf{T}(\tau, \overset{*}{\tau})$ might not peak at a time $\tau = -\overset{*}{\tau}$ (as in equation 3.10) following the stimulus presentation, and its scale invariance might be broken. However, if the connection strengths remain constant over time, the presentation of a stimulus will always lead to the same representation in the **T** layer. Consequently, the prediction **p** generated will consistently peak at an appropriate delay even though scale invariance could be broken. But if the connection weights fluctuate over time, the peak of the **p** distribution could fluctuate, and this fluctuation will be proportional to the fluctuations in the connection weights.

5 Application to Behavioral Experiments

In this section, we demonstrate how the scale-invariant timing information in **p** could be used to support behavioral effects in apparently diverse settings ranging from classical conditioning to episodic memory. The experimental tasks and dependent variables in each of the demonstrations are very different. In each case, we use a minimal behavioral model to construct dependent measures using the information provided by the **p** distribution. In effect, **p** supplies timing information that is utilized by another stage of processing in a way appropriate to solve each problem. Our goal is to suggest that scale invariance observed in diverse cognitive domains could reflect a common source of timing information provided by the scale-invariant representation of the recent history we have described here.

First, we sketch out how **p** can generate well-timed responses at multiple timescales and how the temporal spread in **p** can be mapped on to the temporal spread in behavioral responses in a variety of domains. In all of these demonstrations, we fix $k = 4$, implying that **p** has exactly the same

properties in all of the demonstrations. We compare the model to a fear-conditioning experiment on goldfish (Drew, Couvillon, Zupan, Cooke, & Balsam, 2005) and to an interval-timing experiment on humans (Rakitin et al., 1998). Next, we sketch out how the scale-invariant timing information contained in **p** could cause the lack of scale in the learning rate observed in autoshaping experiments (Gallistel & Gibbon, 2000). Finally, we show that since **p** is constructed from temporal information at many scales, it leads to a recency effect in episodic memory that persists across multiple timescales (Howard et al., 2008). The model's ability to describe the major empirical factors affecting the recency effect is illustrated by comparing its predictions to the results of several free recall experiments.

5.1 Timing in Fear Conditioning. In Pavlovian conditioning, a conditioned stimulus (CS) is paired via some temporal relationship with an unconditioned stimulus (US) during learning. At test, the CS is re-presented and a conditioned response (CR) is observed, reflecting learning about the pairing between the CS and the US. Human and animal subjects can learn a variety of temporal relationships between stimuli and respond in a way that reflects learning those relationships (Gallistel & Gibbon, 2000; Balsam & Gallistel, 2009). In order to illustrate the utility of the prediction **p** in generating timed responses, let us consider an experiment on goldfish by Drew et al. (2005).

During the learning phase, the US (shock) followed the CS (light) after a fixed latency. One group of fish had a latency period of 5 seconds, and another group had a latency period of 15 seconds. On the left side of Figure 8, the time distribution of the CR is plotted with respect to the delay since the onset of CS during the test phase. Note that the peak CR approximately matches the reinforcement delay, even at the earliest learning trials, and it becomes stronger as the number of learning trials increases.

This pattern of results is qualitatively consistent with the predictions of the model. For simplicity, we assume that the onset of the CS is represented by a separate **f** node and focus on the corresponding **T** column. If we take the CS onset to be a delta function as in the previous section, the **T**-layer activity following the CS onset is given by equation 3.2. Let the US occur consistently after a delay τ_o following the CS onset. The **T**-layer activity $\mathbf{T}(\tau_o, \overset{*}{\tau})$ that peaked at the appropriate $\overset{*}{\tau}$ will be recorded in **M** as the synaptic weights connecting the **T** layer and the US node in the **f** layer. Hence at the test phase, when the CS is repeated, the component of the prediction corresponding to the US, p_{us}, automatically inherits the timing properties with the peak at an appropriate delay. The functional form of p_{us} is simply that of p_{stop} in equation 3.17. Moreover, during every learning trial, the synaptic weights in **M** are updated by the exact same amount. Consequently, p_{us} grows larger with learning trials.

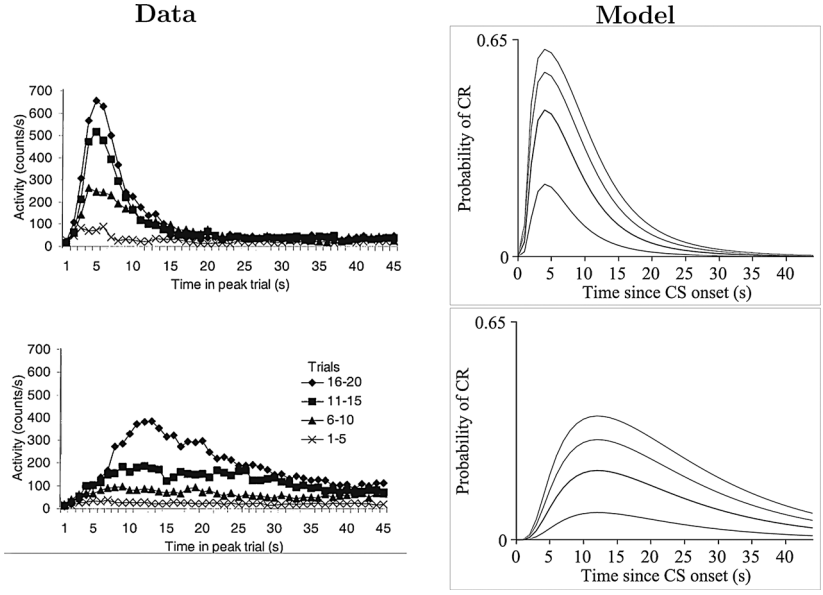

Figure 8: Timing in goldfish. During training, the US (shock) was presented 5 s (top panel) and 15 s (bottom panel) after the onset of the CS (light). The rate of CR is plotted in the left panel as a function of the time after presentation of the CS in the absence of the US. The different curves represent different numbers of learning trials. Notice that the response gets stronger with learning trials. The right panel shows the probability of CR generated from simulations of the model. In these simulations, for simplicity, only the onset of CS is encoded into $\mathbf{t}$. The parameters used in this simulation are $\theta = 0.1$ and $\phi = 1$ (see text for details). (Reproduced from Drew et al., 2005.)

The quantity p_{us} by itself has several properties that render it inappropriate for treating it as a direct measure of response. First, it starts out at 0, where, in general, there is some spontaneous probability of response even prior to learning. Second, with $\mathbf{M}$ calculated using the simple Hebbian learning rule, p_{us} grows without bound as a function of trial. Here we will use a minimal model to map p_{us} onto the probability of a CR, which is what can be behaviorally observed. We calculate the probability of fear response at each moment within each trial as

$$\text{Probability of response} = \frac{p_{us} + \theta}{p_{us} + \theta + \phi}. \tag{5.1}$$

Here, θ and ϕ are free parameters that control the background rate of responding and the scale over which the probability of response saturates.

In the absence of the CS, p_{us} is 0, and the baseline response probability is $\theta/(\theta+\phi)$. We shall take the probability of CR to be simply equation 5.1 with the baseline response probability subtracted from it. Overall, equation 5.1 can be interpreted as the relative activation of the US component with respect to the other components of $\mathbf{p}$. A heuristic interpretation of θ is that it is a measure of the activation of the US component by $\mathbf{T}$ columns that do not correspond to the CS. Similarly, a heuristic interpretation of ϕ is that it is a measure of the activation of $\mathbf{p}$ nodes other than the one corresponding to the US.

In this experiment, the time interval between successive learning trials, the intertrial interval, is on an average 90 seconds for both the 5 seconds reinforcement latency condition and the 15 seconds reinforcement latency condition. Although equation 5.1 does not explicitly consider the effect of the intertrial interval, its effect can be assumed to be represented in the parameters θ and ϕ. The experiment corresponding to the left panel of Figure 8 is simulated, and the probability of a conditioned response is plotted in the right panel. The parameter controlling the accuracy of the temporal representation, k, is fixed across all of the applications in this letter. In these simulations, θ and ϕ are free parameters. As long as ϕ is much bigger than θ, the simulation results make a good qualitative match to the experimental data. It should be noted that although the p_{us} is scale invariant, the probability of response defined by equation 5.1 is not precisely scale invariant because of the nonlinearity introduced through the parameters θ and ϕ. In fact, the data are not exactly scale invariant either; the response distributions of 5 and 15 seconds do not precisely overlap when linearly scaled. Drew et al. (2005) point out that this could be due to intergroup differences in acquisition rate. As noted above, the fact that the intertrial interval is fixed across the experiments would also lead us to expect deviations from scale invariance. Changing the intertrial interval across experiments would be expected to result in changes to θ and ϕ.

In addition to using a simplified behavioral model as in equation 5.1, another simplifying assumption used in these simulations is that only the onset of the CS contributes to the prediction of the US. If this assumption were exactly true, it would imply identical CRs for both delay conditioning and trace conditioning as long as the total time interval between the CS onset and US is held constant. To more accurately model the various classical conditioning CS-US pairing paradigms, we should consider not just the CS onset, but also the whole CS duration and the CS offset, and associate each of them with the US. This would of course generate well-timed responses and potentially also distinguish the results from various experimental paradigms, but would require a more detailed behavior model (as opposed to equation 5.1) and more free parameters such as the relative saliencies of the CS-onset, CS-duration, and CS-offset.

5.2 Human Interval Timing. It is well known that when human subjects are asked to reproduce a temporal interval, the errors produced across trials

are scale invariant (Rakitin et al., 1998). This property suggests that the scale-invariant internal representation of time described here could be the source of the judgment in human interval timing. Our goal is not to provide a detailed model of all aspects of interval timing, but simply to illustrate a consistency with the qualitative properties of interval timing data.

At first glance, one property of human interval timing data seems to rule out the representation of internal time developed here. The CVs of the response distributions are typically small enough that it would seem to require unreasonably large values of k to generate a prediction $\mathbf{p}$ with a similarly small CV. For instance, Rakitin et al. (1998) observed CVs less than 0.2. To generate a time distribution of $\mathbf{p}$ with a CV so small, we would require a value of k greater than 50. This would require computing the 50th derivative of $\mathbf{t}(\tau, s)$, which would require exquisitely balanced connections over 50 near neighbors to accomplish. However, the CV of $\mathbf{p}(\tau)$ computed over τ does not directly determine the CV of the response distributions that would be observed. This is because the response in the task cannot be considered a simple readout of the prediction signal, but the output of a potentially complex decision. Here, we sketch a simple model of human interval timing and compare it to the response distributions observed by Rakitin et al. (1998) in their experiment 3. The goal of this excercise is not to construct a detailed model of performance in this task. Rather, it is to demonstrate that a simple response such as a variable threshold acting on the function $\mathbf{p}(\tau)$ can give rise to scale-invariant behavioral response distributions with small CVs. This demonstration relies on the scale invariance of the $\mathbf{p}$ function and not specifically on its mathematical form.

On each training trial in experiment 3 (Rakitin et al., 1998), a stimulus was presented on a computer screen for a duration τ_o, after which the stimulus changed color and then disappeared. The subjects were instructed to remember the duration without explicitly counting. To ensure that the subjects did not count, random distracting digits were presented on top of the stimulus during training trials. After a block of 10 training trials, a block of 80 test trials began, during which the subjects' memory for the interval was evaluated. During test trials, the stimulus was again presented with the distracting digits on the screen. Subjects were instructed to press a button repeatedly for a brief period centered around the target duration τ_o. The target disappeared after the subject stopped pressing the button or after $3\tau_o$ had elapsed, whichever came first. On approximately 25% of the test trials, the stimulus changed color once τ_o elapsed. These trials reinforced the memory for the target duration and provided the subjects feedback about their estimate of τ_o, helping them correct for any systematic errors and more accurately center their responses around the target duration. The left panel of Figure 9 shows the normalized response distributions for three target durations of 8, 12, and 21 seconds. When linearly rescaled, the distributions overlap reasonably well, demonstrating the scalar property.

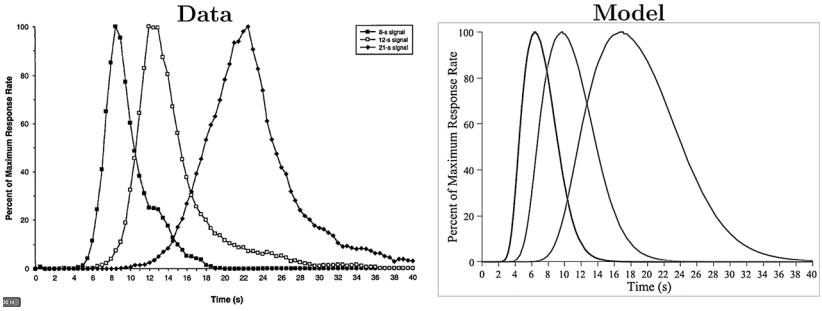

Figure 9: (Left) The data from experiment 3 of Rakitin et al. (1998). Human subjects were instructed to reproduce time intervals of 8, 12, and 21 seconds. The response distributions peak near the appropriate times and obey the scalar property. The coefficient of variation (CV) of the response distributions is about 0.2. (Right) The response distribution generated by the model according to equation 5.2, with $k = 4$ and $\Omega = 0.98$. This simple behavioral model is precisely scale invariant.

The three distributions peak near the appropriate interval (some a bit early, some a bit late), and all three show a pronounced asymmetry.

Clearly these response distributions are much sharper (CV $\simeq 0.2$) than the response distributions from the fear-conditioning experiment (see Figure 8). Recall that the modeled response distributions in Figure 8 are given by equation 5.1, where p_{us} is simply p_{stop} in equation 3.17 with $k = 4$. Because the fear response of the goldfish is very different from the button-press response of human subjects, it is unreasonable to expect the same behavioral performance function (see equation 5.1) to be valid here. We now show that a different (but simple) performance function that reads in p_{stop} with $k = 4$ can yield CVs that match the empirical values found in human interval timing.

In each test trial, the subjects showed a characteristic low-high-low response pattern, as though they start and stop responding based on some decision threshold (Rakitin et al., 1998). Since the subjects perform the task many times, it is reasonable to assume that the subjects are aware of the maximum value attained by p_{stop} in any trial. So let us consider normalized-p_{stop} denoted by $\bar{P}_{stop}$ that reaches a maximum value of 1 within any trial as the quantity accessed by a threshold based decision mechanism. Let the subjects respond uniformly when $\bar{P}_{stop}$ is above a threshold that varies from trial to trial:

$$\bar{P}_{stop} \geq \Omega[1 + v], \tag{5.2}$$

where Ω is the threshold and v is a random variable denoting the variability of the threshold across trials. Because the $\bar{P}_{stop}$ is scale invariant, the time within a trial at which it crosses any value of the threshold will be linearly scaled to the target interval. Since we assume a constant responding when $\bar{P}_{stop}$ crosses the threshold, the response times will also be linearly scaled to the target interval. Integrating over any probability distribution for thresholds retains the property of scale invariance. Hence, the threshold can be any random variable, but the responses will be scale invariant. Although the shape of the response distribution will depend on the threshold distribution, the property of scale invariance does not.

Here we have chosen the distribution of v across trials to be gaussian with zero mean and a standard deviation σ. Moreover since $\bar{P}_{stop}$ varies between 0 and 1 in each trial, we have to restrict the value of v between -1 and $(\Omega^{-1} - 1)$. For each testing trial, a random value of v is chosen, and the response interval is computed based on equation 5.2. The right panel of Figure 9 shows the response distribution averaged over many testing trials with $\Omega = 0.98$ and the standard deviation $\sigma = 0.20$. The CV of the generated response distribution is comparable to that of the data and is significantly smaller than the CV of p_{stop} with $k = 4$. This illustrates that even with a small value of k, a suitable task-dependent performance function can be constructed from p_{stop} to yield scale invariance and a small value of CV.

There are two crucial features of the response distributions that should be noted. First, note that the model yields slightly asymmetric response distributions that are qualitatively consistent with the data. Both the data and the model response distributions are positively skewed. Second, note that the positions of the peaks of the response distributions from the model are consistently earlier than the target durations (at 80% of the target duration), consistent with equation 3.19. Even though we can attain a small CV with a small value of k, the position of the peak of the distribution cannot be pushed closer to the target duration without increasing k or choosing a more elaborate function to control behavior using p_{stop}. It is also possible that the location of the peak is a consequence of learning that takes place due to feedback during the testing trials. The learning rule we are using here (see equation 2.11) may be too simplistic to account for this experiment. It seems natural that the information gained about the errors in the prediction during the feedback trials would be used by the subjects to reposition the peak accurately, while this information is completely ignored by the learning rule. If we knew with certainty that the peak of the distribution was well timed on the initial learning trials, this would constrain k to be high. Conversely, if we knew that the peak was early during the initial learning trials, this would similarly constrain k to be small. Unfortunately, the data from experiment 3 of Rakitin et al. (1998) report only the response distributions averaged over blocks of 80 test trials. In contrast, in experiment 1 of Treisman (1963), where blocks of only eight test trials were averaged, a

significant lengthening of the reproduced time interval was observed in the later blocks. That is, the estimate of a time interval gradually grows with practice on that interval. We could potentially shift the early peak generated by the model gradually toward the accurate spot by replacing the simple Hebbian learning rule with an error-driven learning rule.

The temporal information used in the simulations above is deterministic. On every trial, the activity of the **f**, **t**, and **T** layers follows the exact same pattern, and so does the prediction p_{stop}. The variability in the response distribution is achieved purely through the threshold variability in the decision process (see equation 5.2), and not through any intrinsic variability in the timing mechanism. Of course, there should also be some variability in the timing mechanism that partly contributes to behavioral variability across trials, and so it is more reasonable to consider $\mathbf{p}(\tau)$ as a measure of central tendency at a moment τ within a trial. However, some types of noise in the timing mechanism can violate scale invariance in the variability across trials. By choosing the timing mechanism to be deterministic and the decision threshold to be noisy, we do not need tailored noise to generate scale invariance in variability across trials. This is because the decision threshold is imposed on the prediction **p**, which has a scale-invariant smear. As long as the decision mechanism simply acts on the normalized prediction, any kind of noise in the decision threshold will automatically lead to scale invariance in the variability across trials. More specifically, for any choice of random variable v in equation 5.2, no matter how complex, the variability across trials will also be scale invariant.

5.3 Scale Invariance in Learning Rate. Scale invariance in classical conditioning is observed not only in the timing of response distributions, but also in the rate of response acquisition (the learning rate). In the earlier illustration with the data of Drew et al. (2005) (see Figure 8), note that it takes more learning trials in the 15 seconds condition to generate equally strong responses as in the 5 seconds condition. At first glance it appears that the learning rate is slower for longer CS-US latencies. But it turns out that we would not expect this to be the situation if the intertrial interval in the 15 seconds reinforcement condition is also appropriately scaled up. In general, the number of learning trials needed to elicit a basic-level response turns out to be independent of the scale of the reinforcement latency when the intertrial interval is appropriately adjusted to the same scale (Gallistel & Gibbon, 2000). It has been pointed out (Balsam & Gallistel, 2009) that the learning rate is simply proportional to the informativeness of the encoded predictive relationship of the CS to the US. This point of view can be precisely formalized in the model by computing the informativeness of the temporal distribution of the prediction **p** generated by the model. We now show that the scale invariance of **p** will lead to scale invariance in its informativeness and consequently lead to a scale-invariant learning rate.

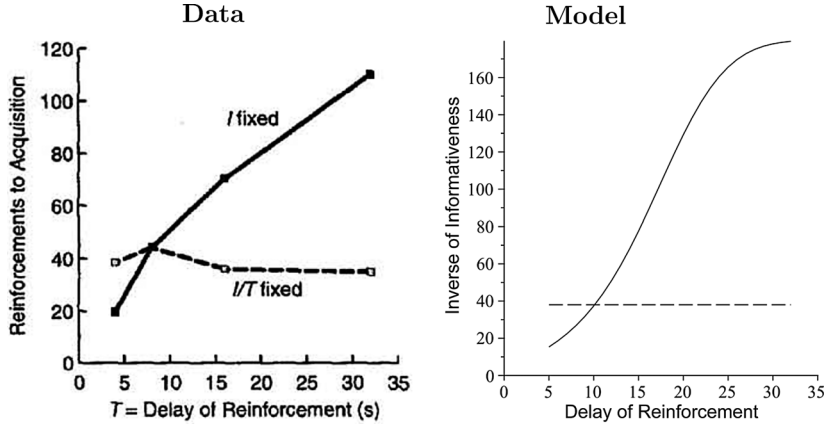

Figure 10: (Left) The number of trials required for pigeons to acquire CR in autoshaping experiment (Gallistel & Gibbon, 2000). The required number of reinforced learning trials depends on only the ratio of the intertrial interval to the delay of reinforcement. The information contained in the temporal distribution of the prediction $\mathbf{p}$ generated by the model is computed from equation 5.12. (Right) $\mathrm{H}_{\mathrm{com}}^{-1}$ is plotted with a factor of 6, so as to be readily comparable with the left panel. In both panels, the solid curve corresponds to the intertrial interval $\mathcal{I}$ fixed at 48 sec, and the flat dashed line corresponds to the ratio $\mathcal{I}/\mathcal{T}$ fixed at 5.

Let us consider an experiment where pigeons are trained in an appetitive conditioning paradigm (Gallistel & Gibbon, 2000). In each learning trial, the CS (light) is paired with the US (food) following a reinforcement delay $\mathcal{T}$. Let the average intertrial interval be $\mathcal{I}$. After a certain number of learning trials, the pigeons start pecking a key in response to the CS. This key peck seems to be in anticipation of the US (food) because it does not happen when the CS is not reinforced by the US, so the key peck can be thought of as a CR. The number of learning trials needed for the pigeons to acquire a CR shows a striking property: it depends only on the ratio $\mathcal{I}/\mathcal{T}$, and not on $\mathcal{I}$ and $\mathcal{T}$ separately (see Figure 10). We now evaluate the prediction $\mathbf{p}$ generated when the model is trained on this experimental paradigm and compute its informativeness by calculating the entropy of its distribution.

Let us place the start of a trial, the onset of CS, at time $\tau = 0$. The US then occurs at $\tau = \mathcal{T}$, and the trial lasts until $\tau = \mathcal{I}$, after which the next trial immediately begins. Following the first learning trial, the presentation of CS induces a prediction for the US as a function of τ, which can be deduced from equation 3.17 to be

$$\mathrm{p}_{cs.us}(\tau) \sim \frac{1}{\mathcal{T}} \frac{y^k}{(1+y)^{2k+1}}, \quad \text{where} \qquad y = \tau/\mathcal{T}. \tag{5.3}$$

Since on an average the US is repeated at an interval $\mathcal{I}$, it will be associated with its prior presentation. Since $(\mathcal{I} - \mathcal{T} + \tau)$ is the time since the prior presentation of the US, the prediction of the US due to its prior presentation as a function of time within the trial is

$$\mathrm{p}_{us.us}(\tau) \sim \frac{1}{\mathcal{I}} \frac{z^k}{(1+z)^{2k+1}}, \quad \text{where} \qquad z = (\mathcal{I} - \mathcal{T} + \tau)/\mathcal{I}. \tag{5.4}$$

By defining the ratio $\mathcal{I}/\mathcal{T}$ to be $\mathcal{R}$, we have $z = 1 + \mathcal{R}^{-1}(y-1)$ and

$$\mathrm{p}_{us.us}(\tau) \sim \frac{1}{\mathcal{I}} \frac{(1+\mathcal{R}^{-1}(y-1))^k}{(2+\mathcal{R}^{-1}(y-1))^{2k+1}}. \tag{5.5}$$

Note that in equations 5.3 and 5.5, only the nearest presentations of the CS and the US are considered. Ideally, we would expect all the prior presentations of the CS and the US to be involved in the prediction, but since their contributions will be very tiny, we will ignore them in these equations.

For the sake of completeness, we also consider the possibility that the US could be associated with some experimental context cues other than the CS. As a consequence, the experimental context cues will generate a prediction for US, $\mathrm{p}_{ec.us}(\tau)$. A fair assumption is that this will not depend on the time within a trial and will be inversely proportional to the trial length $\mathcal{I}$. For longer trial lengths, the experimental context cues will be more strongly associated with themselves than the US and will contribute less toward predicting the US:

$$\mathrm{p}_{ec.us}(\tau) \sim 1/\mathcal{I}. \tag{5.6}$$

The overall prediction for US as a function of time within a trial is then

$$\begin{aligned} &\mathrm{p}_{cs.us}(\tau) + \mathrm{p}_{us.us}(\tau) + \mathrm{p}_{ec.us}(\tau) \\ &= \frac{1}{\mathcal{T}} \left[C_1 \frac{y^k}{(1+y)^{2k+1}} + C_2 \mathcal{R}^{-1} \frac{(1+\mathcal{R}^{-1}(y-1))^k}{(2+\mathcal{R}^{-1}(y-1))^{2k+1}} + C_3 \mathcal{R}^{-1} \right] \\ &\equiv \frac{G(y,\mathcal{R})}{\mathcal{T}}. \end{aligned} \tag{5.7}$$

Here C_1, C_2, and C_3 are parameters indicating the relative saliencies of the three contributing terms to the prediction of the US.

To construct a probability distribution for the overall prediction for US within a trial, we normalize expression 5.7, such that its integral over the trial length is unity, that is,

$$\mathrm{P}(\tau) = \frac{1}{N} \frac{G(y,\mathcal{R})}{\mathcal{T}}, \tag{5.8}$$

where

$$N = \int_0^{\mathcal{I}} \frac{G(y, \mathcal{R})}{\mathcal{T}} d\tau = \int_0^{\mathcal{R}} G(y, \mathcal{R}) dy. \tag{5.9}$$

Clearly, N depends on just $\mathcal{R}$ and not explicitly on $\mathcal{I}$ and $\mathcal{T}$. The entropy of the distribution $\mathrm{P}(\tau)$ is

$$\mathrm{H_P} = \int_0^{I} -\mathrm{P}(\tau) \ln \mathrm{P}(\tau)\, d\tau. \tag{5.10}$$

To estimate the amount of information communicated about the timing of the US, we calculate how different the distribution $\mathrm{P}(\tau)$ is from a uniform distribution. This is formally done by computing the Kullback-Leibler divergence between $\mathrm{P}(\tau)$ and the uniform distribution, which simply turns out to be the difference in entropy of the two distributions. That is, $\mathrm{H_{com}} = \mathrm{H_{uniform}} - \mathrm{H_P}$:

$$\mathrm{H_{com}} = \ln \mathcal{I} + \int_0^{\mathcal{I}} \mathrm{P}(\tau) \ln \mathrm{P}(\tau)\, d\tau. \tag{5.11}$$

This can be calculated as

$$\begin{aligned}
\mathrm{H_{com}} &= \ln \mathcal{I} + \int_0^{\mathcal{I}} \frac{G}{N\mathcal{T}} \left[\ln G - \ln N - \ln \mathcal{T}\right] d\tau \\
&= \ln \mathcal{I} + \int_0^{I} \frac{G}{N\mathcal{T}} \left[\ln G - \ln N - \ln \mathcal{I} + \ln \mathcal{R}\right] d\tau \\
&= \int_0^{\mathcal{I}} \frac{G}{N\mathcal{T}} \left[\ln G - \ln N + \ln \mathcal{R}\right] d\tau \\
&= \int_0^{\mathcal{R}} \frac{G}{N} \left[\ln G - \ln N + \ln \mathcal{R}\right] dy \\
&= \ln \mathcal{R} - \ln N + \frac{1}{N} \int_0^{\mathcal{R}} G \ln G\, dy.
\end{aligned} \tag{5.12}$$

Since G is a function of y and $\mathcal{R}$ and N depends only on $\mathcal{R}$, we see that $\mathrm{H_{com}}$ will depend only on $\mathcal{R}$ and not explicitly on $\mathcal{I}$ and $\mathcal{T}$.

Following the arguments of Balsam and Gallistel, let us take the learning rate to be directly proportional to the informativeness of $\mathbf{p}$. We expect the number of trials required for the pigeons to acquire CR to be inversely related to $\mathrm{H_{com}}$. This can be heuristically justified by considering $\mathrm{H_{com}}$ to be the information acquired in each trial. When the total accumulated information over n trials, $n\mathrm{H_{com}}$, crosses a threshold, the pigeons acquire the CR. In the right panel of Figure 10, $\mathrm{H_{com}^{-1}}$ is plotted with a factor of 6 so that it

can be explicitly compared with the trials required for CR acquisition in the experiment plotted on the right side. In plotting Figure 10 from equation 5.12, we make certain simplifications. Comparing the functional forms, we note that $p_{us.us}(\tau)$ is much flatter than $p_{cs.us}(\tau)$. Moreover, the fact that $\mathcal{I}$ is not a constant across trials in the experiment makes $p_{us.us}(\tau)$ even flatter. The contribution of $p_{us.us}(\tau)$ is functionally very similar to $p_{ec.us}(\tau)$. Hence for simplicity, we take C_2 to be 0 in plotting Figure 10 and assume that C_3 contains contributions from both $p_{us.us}(\tau)$ and $p_{ec.us}(\tau)$. We have chosen $C_1 = 1$ and $C_3 = 0.0002$. We have chosen C_1 to be much higher than C_3 to ensure that the saliency of CS is much higher than the saliency of the experimental context cues.

Note that it is not essential to assume a simple proportionality between informativeness (H_{com}) and the learning rate to account for a scale-invariant learning rate. All we need is that the informativeness of the **p** distribution be the only determinant of the learning rate in order to account for its scale invariance. In fact, any model that generates scale-invariant prediction values that change over time within each trial can account for the scale-invariant learning rate.

5.4 Recency Effect in Episodic Recall. In the free recall task, subjects are presented with a list of words one at a time and then asked to recall all the words they can remember in the order the words come to mind. Perhaps because there is no cue for memory other than the subject's memory for the episode itself, free recall is considered a particularly sensitive assay of episodic memory. The recency effect in free recall is dominated by the tendency to initiate recall with items from the end of the list. Recall of subsequent items undoubtedly utilizes previously recalled words as cues (Kahana, 1996; Howard & Kahana, 2002b), but we shall focus only on the first recalled item here. We propose that the state of **T** at the end of the list is the only cue used to generate the first recalled item. While scale invariance has not been unambiguously demonstrated in the recency effect in free recall, recency has been observed over all laboratory timescales, from fractions of a second to hours (Glenberg et al., 1980; Howard et al., 2008; Murdock & Okada, 1970). A memory mechanism based on a scale-invariant representation of time will naturally lead to this finding.

We now show that the scale-invariant stimulus history represented in the **T** layer of the model can be used as a cue to initiate free recall. The recency effect in free recall shows three robust temporal features:

1. A delay at the end of the list, during which subjects perform a demanding distractor task to prevent rehearsal, decreases the sharpness of the recency effect (Glanzer & Cunitz, 1966; Postman & Phillips, 1965; Glenberg et al., 1980).
2. Increasing the time gap between study items, again filled with a distractor task to prevent rehearsal, increases the sharpness of the recency

effect (Bjork & Whitten, 1974; Glenberg et al., 1980; Howard & Kahana, 1999).

3. The recency effect persists across multiple timescales (Glenberg et al., 1980; Howard et al., 2008; Moreton & Ward, 2010).

To demonstrate the model's ability to account for these features, we consider the **T**-layer activity after studying a list of words and derive an analytic expression for the prediction vector **p** at the end of the study list. For analytic simplicity, we take each word in the list to be represented by a separate **f** node. However, in reality, different words would have overlapping representations in the **f** layer, leading to semantic effects in the recall stage (Howard & Kahana, 2002b; Howard, Addis, Jing, & Kahana, 2007), which we shall not address here (but see Howard, Shankar, & Jagadisan, 2011, for a template of how to construct semantic representations). By relating the **p**-component of each word to its probability of recall, we can obtain an analytic expression for the recency effect in the initiation of free recall. We will see that the scale invariance of **p** naturally leads to the features of the recency effect already noted.

We denote the list items by $\mathbf{f}_n$, where n is the position of the item from the end of the list. Let the nth item be presented at time τ_n, and let the time interval between successive items be a constant $\delta\tau$. The **T**-layer activity at the end of the list is $|\mathbf{T}(\tau_o)\rangle$; prior to the presentation of the item $\mathbf{f}_n$, it is $|\mathbf{T}(\tau_n)\rangle$. This can be pictorially represented as

$$\begin{array}{ccccccc} \ldots \downarrow \mathbf{f}_{n+m} & \ldots & \mathbf{f}_{n+1} & \downarrow \mathbf{f}_n & \ldots. & \mathbf{f}_1 & \downarrow \text{End} \\ |\mathbf{T}(\tau_{n+m})\rangle & \xleftarrow[m\delta\tau]{} & |\mathbf{T}(\tau_n)\rangle & \xleftarrow[n\delta\tau]{} & & |\mathbf{T}(\tau_o)\rangle. & \end{array}$$

Note that only the predecessors of $\mathbf{f}_n$ will leave an image in $|\mathbf{T}(\tau_n)\rangle$, which is associated to $\mathbf{f}_n$ in **M** (see equation 2.11).

At the end of the list, during the recall stage, the **T**-layer activity generates a prediction **p** based on the stored associations. To estimate how well the item $\mathbf{f}_n$ is predicted or recalled, we have to compute the inner product $\langle \mathbf{T}(\tau_n)|\mathbf{T}(\tau_o)\rangle$ in accordance with equation 2.15. Again, note that $|\mathbf{T}(\tau_o)\rangle$ will contain activity in all the **T** columns corresponding to all the items in the list, while $|\mathbf{T}(\tau_n)\rangle$ will contain activity in only columns corresponding to the items that were presented prior to $\mathbf{f}_n$. Their inner product will hence involve only those columns corresponding to items presented prior to $\mathbf{f}_n$. To compute the inner product, we denote the **T** column activity associated with the item $\mathbf{f}_{n+m}$ to be $\mathrm{T}^m(\tau_o, \overset{*}{\tau})$ and $\mathrm{T}^m(\tau_n, \overset{*}{\tau})$ at times τ_o and τ_n, respectively. At the end of the list, the component of the prediction **p** corresponding to the item $\mathbf{f}_n$, $\mathrm{p}_n \equiv \langle \mathbf{p}|\mathbf{f}_n\rangle$ is given by

$$\mathrm{p}_n = \langle \mathbf{T}(\tau_n)|\mathbf{T}(\tau_o)\rangle \equiv \sum_{m=1}^{\infty} A_n^m \tag{5.13}$$

where

$$A_n^m = \int \mathrm{T}^m(\tau_o, \overset{*}{\tau})\, \mathrm{T}^m(\tau_n, \overset{*}{\tau})\, g(\overset{*}{\tau})\, d\overset{*}{\tau} \tag{5.14}$$

is the inner product of the **T**-layer activities at times τ_o and τ_n, restricted to the column corresponding to the item $\mathbf{f}_{n+m}$. In effect, A_n^m denotes the contribution of the image of $\mathbf{f}_{n+m}$ in **T** toward predicting $\mathbf{f}_n$. From equation 3.3, we can write

$$\begin{aligned}
\mathrm{T}^m(\tau_n, \overset{*}{\tau}) &= -\frac{1}{\overset{*}{\tau}} \frac{(k)^{k+1}}{k!} \left(\frac{-m\delta\tau}{\overset{*}{\tau}} \right)^k e^{k\left(\frac{m\delta\tau}{\overset{*}{\tau}}\right)}, \\
\mathrm{T}^m(\tau_o, \overset{*}{\tau}) &= -\frac{1}{\overset{*}{\tau}} \frac{(k)^{k+1}}{k!} \left(\frac{-(n+m)\delta\tau}{\overset{*}{\tau}} \right)^k e^{k\left(\frac{(n+m)\delta\tau}{\overset{*}{\tau}}\right)}.
\end{aligned} \tag{5.15}$$

Substituting into equation 5.14, we obtain

$$A_n^m = \frac{k^{2k+2}}{k!^2} (1 + n/m)^k (m\delta\tau)^{2k} \int \left(\frac{1}{\overset{*}{\tau}} \right)^{2k+2} e^{k(2+n/m)\left(\frac{m\delta\tau}{\overset{*}{\tau}}\right)} g(\overset{*}{\tau})\, d\overset{*}{\tau}. \tag{5.16}$$

With $g(\overset{*}{\tau}) = 1$, the above integral can be calculated as

$$\begin{aligned}
A_n^m &= \frac{k^{2k+2}}{k!^2} (1 + n/m)^k (m\delta\tau)^{2k} \left[\frac{(2k)!}{\left(k(2 + n/m)m\delta\tau\right)^{2k+1}} \right] \\
&= \frac{k(2k)!}{k!^2} \left[\frac{(1 + n/m)^k}{(2 + n/m)^{2k+1}} \right] (m\delta\tau)^{-1} \qquad (5.17) \\
&= \frac{k(2k)!}{k!^2} \left[\frac{(n/m)(1 + n/m)^k}{(2 + n/m)^{2k+1}} \right] (n\delta\tau)^{-1}. \qquad (5.18)
\end{aligned}$$

Some important properties of the above functional form are explicated in Figure 11.

To compute p_n for any given n, we have to sum over the contributions from all its predecessors, that is, sum the coefficients A_n^m over all m. Note from equation 5.17 that for $m \gg n$, $A_n^m \sim m^{-1}$, and the summation over m diverges. To obtain a qualitative functional form of p_n, consider the following approximation, where we reduce the summation of A_n^m over m to

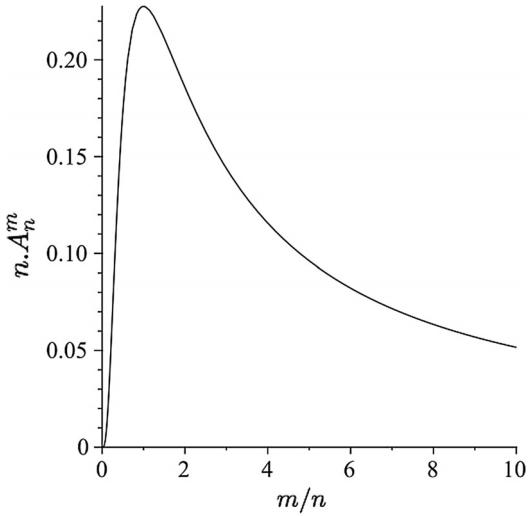

Figure 11: The contribution of the image of item $\mathbf{f}_{n+m}$ in $\mathbf{T}$ in predicting item $\mathbf{f}_n$ when the state of $\mathbf{T}$ at the end of list is used to cue free recall. The function $n \times A_n^m$ is plotted with respect to m/n for $k = 4$. From equation 5.18, note that $A_n^m \to 0$ when either $m \gg n$ or $n \gg m$. The value of A_n^m peaks when m and n are comparable numbers. The exact position of the peak depends on k. For $k = 4$, A_n^m peaks exactly at $m/n = 1$ for any value of n, as can be seen from this plot. For $k = 2$, the peak is at $m/n = 0.62$. One might naively expect that the image of $\mathbf{f}_{n+1}$ in $|\mathbf{T}\rangle$—the representation of the item preceding $\mathbf{f}_n$—would contribute greatly to cueing the item $\mathbf{f}_n$ because of their temporal proximity. If this expectation were true, the function A_n^m would be maximum at $m = 1$. However, this is not the case. When the item $\mathbf{f}_n$ is presented, the image of the immediately preceding item $\mathbf{f}_{n+1}$ in $|\mathbf{T}\rangle$ is relatively sharp. As time passes, however, the image of $\mathbf{f}_{n+1}$ in $|\mathbf{T}\rangle$ becomes more diffuse (see Figure 3). As a consequence, when n is large, the image of $\mathbf{f}_{n+1}$ in the current state of $|\mathbf{T}\rangle$ does not resemble its image in the state of $|\mathbf{T}\rangle$ when the item $\mathbf{f}_n$ was presented. Because the representation of items further in the past changes more slowly, these components in the current state of $|\mathbf{T}\rangle$ contribute more to cueing the item $\mathbf{f}_n$. It turns out that the image of $\mathbf{f}_{n+m}$ in $|\mathbf{T}\rangle$ leads in contributing to the recall of $\mathbf{f}_n$ when the ratio m/n is fixed. This is a consequence of scale invariance intrinsic to the model.

an integral:

$$\mathrm{p}_n \sim \sum_{m=1}^{\infty} \frac{(1+n/m)^k}{m(2+n/m)^{2k+1}} \longrightarrow \int_0^{\infty} dm\, \frac{(1+n/m)^k}{m(2+n/m)^{2k+1}}$$
$$= \int_0^{\infty} \frac{dz}{z} \frac{(1+z^{-1})^k}{(2+z^{-1})^{2k+1}}. \tag{5.19}$$

In the last step, we relabeled m/n as z. As expected, this integral diverges. Note that the integrand is well behaved near $z = 0$, and the divergence is purely due to integration to infinity (or due to summing over infinite m). In reality, any experiment has a starting point, and there is also an upper limit to $|\overset{*}{\tau}|$, so it is reasonable to stop the summation at some finite value m_o. By imposing such a cutoff, equation 5.19 can be rewritten as

$$\mathrm{p}_n \sim \int_0^{m_o/n} \frac{dz}{z} \frac{(1+z^{-1})^k}{(2+z^{-1})^{2k+1}}. \tag{5.20}$$

Let m_o be much greater than the length of the list of items used in the experiment, so that we need to focus only on the regime $m_o/n \gg 1$. Note that when $z \gg 1$, the integrand in equation 5.20 can be well approximated as $2^{-(2k+1)}z^{-1}$. Integrating equation 5.20 yields

$$\mathrm{p}_n \simeq \left[\frac{k(2k)!}{2^{2k+1}(k!)^2\delta\tau}\right] \log(m_o/n). \tag{5.21}$$

The constant in front of $\log(m_o/n)$ simply comes from tracking the coefficients in front of A_n^m (see equation 5.18). Thus, imposing a cutoff at m_o leads to the logarithmic decay of p_n.

Let us now ask the question: What is the probability that the item $\mathbf{f}_n$ will be recalled before the other items? To construct the probability of first recall (PFR), we need to define a retrieval rule that takes in the different components of $\mathbf{p}$ and yields a probability of recall for each item. In general, the different components of $\mathbf{p}$ would compete among each other to generate recall. A successful approach to induce competition is to feed in the different components of $\mathbf{p}$ into leaky accumulators that mutually inhibit each other (Sederberg, Howard, & Kahana, 2008; Polyn, Norman, & Kahana, 2009). For the sake of analytic tractability of equations, we shall work with a simpler retrieval rule used in Howard and Kahana (2002a). Motivated by the finding that the probability of generalization between stimuli decays exponentially with the psychological distance between them (Shepard, 1987), we use the following form for the probability of first recall:

$$\mathrm{PFR}(n) = \frac{\exp(c\mathrm{p}_n)}{\sum_\ell \exp(c\mathrm{p}_\ell)}. \tag{5.22}$$

The summation in the denominator is restricted to the items in the list, so as to restrict the recall process to purely the list items. By fixing $\delta\tau$ to be a constant and plugging in the form of p_n from equation 5.21, we obtain

$$\mathrm{PFR}(n) = \frac{n^{-a}}{\sum_\ell \ell^{-a}}, \tag{5.23}$$

where a is some positive constant that can be inferred from equations 5.21 and 5.22 based on the values of k, $\delta\tau$, and c. Since c is an arbitrary parameter, the value of a does not provide any meaningful constraint on the value of k. So, in effect, a is the only arbitrary parameter. A useful feature that emerges out of this retrieval rule is that the cutoff m_o automatically drops out of the equation for PFR.

The recency effect is a straightforward consequence of the functional form of the PFR: the smaller the value of n, the larger the PFR is. In order to demonstrate the three features of the recency effect mentioned at the beginning of this section, we need to generalize equation 5.23 to incorporate a continuous stream of distracting stimuli in between the list items and a delay at the end of the list. Generally in free recall tasks, the distracting stimuli between the list items are simple arithmetic problems that aid in keeping the participants from rehearsing the list items. For analytical tractability, we assume that the distractor is a sequence of delta function stimuli (similar to the list items) separated by $\delta\tau$. Let the number of distractor stimuli between two list items be $D-1$. Then the total time elapsed between the presentation of two successive list items is $D \times \delta\tau$. Surely there exists a degree of freedom in simultaneously choosing D and $\delta\tau$, because the only experimentally constrained quantity is the total time $D \times \delta\tau$. For the purpose of qualitative analysis, we fix $\delta\tau = 1$ for all the demonstrations here. The delay at the end of the list is also filled with distracting stimuli. We denote the number of distractors in this delay period by d. Now the positions of the list items in the sequence of stimuli are $(d + D \times n)$, where n is item number from the end of list. We can now rewrite equation 5.23 as

$$\mathrm{PFR}(n) = \frac{(d + D \cdot n)^{-a}}{\sum_{\ell}(d + D \cdot \ell)^{-a}}. \tag{5.24}$$

In order to compare this expression to the experimental data, we simply set d and D to appropriately correspond to the experimental values, so that the only free parameter is a. In the following illustrations, we fix $a = 2$, for it gives a good fit to all the experiments.

Property 1: A delay at the end of the list decreases the sharpness of recency. To demonstrate this property, we fix $D = 1$ and plot the PFR for three different values of d: 0, 8, and 16. The top right panel of Figure 12 shows this plot with 16 list items. The top left panel shows an experiment (S. Polyn & Kahana, personal communication, 2011), where subjects recall lists of 16 items following a delay of 0, 8, or 16 seconds. The functional form of equation 5.24 makes it clear why increasing d should decrease the sharpness of recency: as d increases, the effect of n on PFR is decreased.

Property 2: Increasing the gap between the items increases the sharpness of recency. To demonstrate this property, we fix $d = 16$ and plot the PFR for four different values of D: 2, 4, 8, and 16. The bottom right panel of

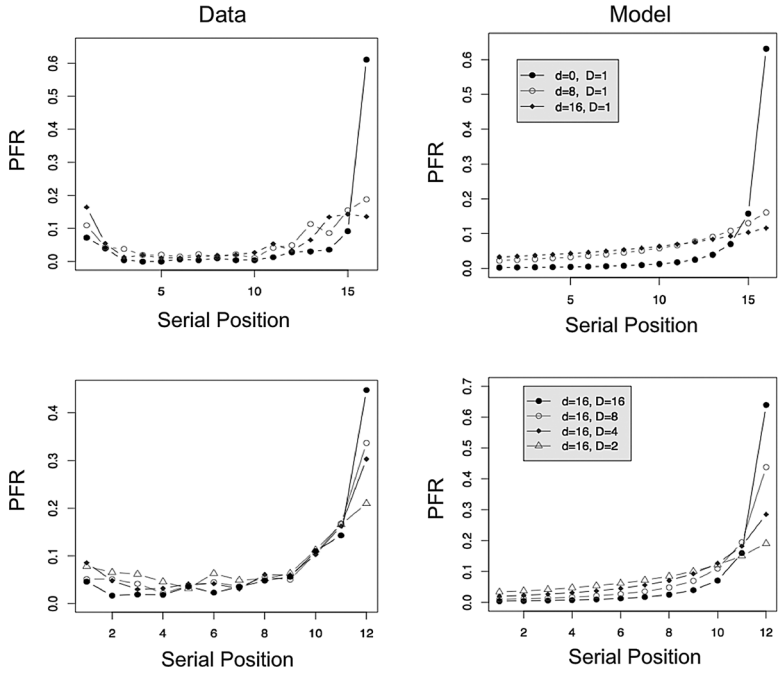

Figure 12: Recency effect depends on the time interval between stimulus presentations and the delay to the recall phase. PFRs from two experiments (Polyn & Kahana, Private communication, at top; Howard & Kahana, 1999, at bottom) with different values of presentation rate and end of list delay are plotted on the left side. On the right side, the corresponding experiments are modeled using equation 5.24 with $a = 2$. See text for further details.

Figure 12 shows this plot with 12 list items. The bottom left panel shows an experiment (Howard & Kahana, 1999) where subjects recall lists of 12 items with different interitem durations—roughly 2, 4, 8, or 16 seconds. Again, the functional form of equation 5.24 makes it clear why increasing D while holding d fixed should increase the sharpness of recency: increasing D amplifies the effect of n on PFR. In order to be consistent with the experiment, the values of n and ℓ in equation 5.24 goes from 0 to 11 rather than 1 to 12.

Property 3: The recency effect persists across multiple time scales. Scale invariance enables the model to generate recency effects across multiple timescales. To elaborate this property, we consider an experiment (Howard et al., 2008) where participants were initially given a list of 10 words and 30 seconds to freely recall the words. Following this, the subjects were given a second list of 10 words to recall. Forty-eight such lists were presented back to back interleaved with 30 second recall periods over the course of

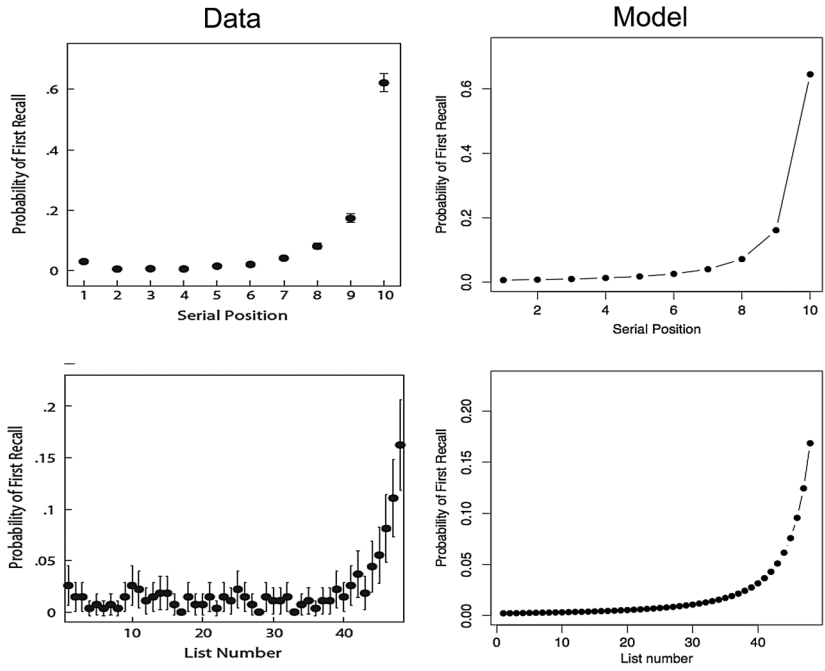

Figure 13: (Left) The recency effect persists across multiple timescales. PFRs of within-list recalls and across-list recalls from Howard et al. (2008). (Right) The experiment is modeled using equation 5.24 with $a = 2$. See text for further details.

the experimental session. The average time between the lists was roughly 49 seconds. The first word recalled within each list was recorded; the PFR is plotted in the top left panel of Figure 13. The PFR from the recall of the lists demonstrates an immediate recency effect. At the end of the experiment, after completion of all 48 lists, there is a break of about 250 seconds, following which the subjects were asked to recall the words from all lists in the order they came to mind. The list from which the first recall occurs is recorded, and the PFR is plotted as a function of list number in the bottom left panel of Figure 13. The overall duration of the experiment is about 1 hour, and we see that the final recall phase clearly shows a recency effect. We refer to this finding as a long-term recency effect.

It turns out that equation 5.24 can account for both the immediate and long-term recency effects. To account for the immediate recency effect, we simply restrict the summation over ℓ in the denominator from 1 to 10 (size of each list). This is plotted in the top right panel of Figure 13. To account for the long-term recency effect, ideally we will have to sum over the contributions from all items within each list and compare the relative

activations of different lists. However, for a qualitative demonstration, we consider a representative element from each list and assume that every item within the list is equally recallable. Moreover, we treat the average distance between the representative items from two successive lists in a manner similar to the distractors D in equation 5.24. To account for the long-term recency effect, we choose $d = 250$, and $D = 49$, in accordance with the experiment. The summation over ℓ in the denominator of equation 5.24 goes from 1 to 48 (number of lists). This function is plotted in the bottom-right panel of Figure 13. It is clear that the model shows a recency effect over this longer timescale.

The scale-invariant behavior of the prediction $\mathbf{p}$ generated by the model can thus be adopted to generate a recency effect in memory that satisfies the three basic properties described above.

6 Summary and Discussion

Understanding how the brain represents and utilizes the stimulus history leading up to a given moment may be a key to understanding problems in a variety of cognitive domains. We have described a three-layer neural network model ($\mathbf{f}$, $\mathbf{t}$, and $\mathbf{T}$; see Figure 1) that integrates the stimulus history into the leaky integrator layer $\mathbf{t}$, in the form of a Laplace transform, and then approximately inverts the transformation (through $\mathbf{L}_k^{-1}$) at each moment to reconstruct the stimulus history in the timing layer $\mathbf{T}$. The inversion performed by $\mathbf{L}_k^{-1}$ can be realized through alternating bands of balanced excitatory and inhibitory connections from k local neighbors in the $\mathbf{t}$ layer onto the $\mathbf{T}$ layer (see Figure 5). The resulting representation of stimulus history in the $\mathbf{T}$ layer is scale invariant and can be naturally interpreted as a stimulus-specific representation of time.

Once we theoretically accept such an internal representation of time and stimulus history, it is rather simple to induce scale invariance at the behavioral level. We have demonstrated this using simple Hebbian association between the $\mathbf{f}$ and $\mathbf{T}$ layers. We have shown that the behavioral-level predictions generated are resistant to noise in the $\mathbf{f}$ and $\mathbf{t}$ layers. The model can generate the scalar variability in the response distributions of interval timing experiments. In particular, we showed model descriptions of the response distributions from a fear conditioning experiment on goldfish (Drew et al., 2005) and the response distributions of a human interval timing experiment (Rakitin et al., 1998). The model can also explain the fact that there is no characteristic scale or a time window for associative learning to take place (Balsam & Gallistel, 2009). We have demonstrated that the timing information encoded in the model can account for the scale-free learning rate observed in autoshaping experiments on pigeons (Gallistel & Gibbon, 2000). Finally, the model's ability to form associations between stimuli from multiple timescales renders it suitable to address the recency effect in episodic memory.

6.1 Time Cells. Because their responses peak at specific latencies following the presentation of stimuli, the nodes in the **T** layer can be thought of as time cells. Recent findings indicate the possibility of the presence of such cells in the hippocampus of rats. In an experiment (MacDonald, Lepage, Eden, & Eichenbaum, 2011), rats were presented with one of two sample stimuli followed by a delay. The identity of the stimulus predicted which of the two choice stimuli available after the delay was associated with reward. During the delay between the sample stimulus and the choice stimuli, the animal stayed in a small enclosure. During the delay period, different cells were found to respond at different intervals of time. Moreover, the cells that responded early in the delay interval had a smaller temporal spread in their activity than the cells that responded later in the delay interval. This observation is consistent with the property of the time cells in the model. The model makes a more specific prediction that the temporal spread in the activity of the time cells should exhibit scale invariance, which has not been explicitly verified with the available data. Moreover, the activity of cells in the delay period was found to depend on the identity of the sample stimulus, a finding that is also consistent with the model. If we knew that these particular cells were encoding **T** rather than receiving a transformed version of **T**, we could use neurophysiological observations to constrain the number of density functions $g(\overset{*}{\tau})$ from the distribution of peak responses across cells and obtain the value of k from the temporal spread in their activity.

Note that the nodes in **T** are sequentially activated. The activity of a **T** cell responding at a shorter latency does not directly influence the activity of the larger latency cells; rather, both are caused by activity in **t**. There are many models that assume that cells that are activated in series cause one another to fire (Hasselmo, 2009; Itskov, Curto, Pastalkova, & Buzsaki, 2011; Jensen & Lisman, 1996; Tieu, Keidel, McGann, Faulkner, & Brown, 1999). A subset of these has been applied to problems in timing behavior. For example, in the model of Tieu et al. (1999), late-spiking neurons that can fire spikes after a lengthy depolarization are connected in chains. One could also adaptively modulate the spiking threshold in neurons with Mexican hat–like connectivity with neighbors to generate sequential activity (Itskov et al., 2011) in densely connected cells, similar to that observed in hippocampal neurons (Pastalkova, Itskov, Amarasingham, & Buzsaki, 2008; MacDonald et al., 2011). TILT, on the other hand, illustrates that the observation of sequentially activated cells does not necessarily imply that they are connected in series. In principle, it should be possible to distinguish chained firing models from TILT by observing correlations between the activity of different time cells across trials.

6.2 Leaky Integrators. Though the brain contains cells with properties like those hypothesized for the **T** layer, this does not imply that they result from the mechanism we have hypothesized here. Our mechanism requires a set of leaky integrators with a variety of time constants. In order to represent

timescales up to $\tau_{\max}$, the leaky integrators in the **t** layer should have time constants up to $\tau_{\max}/k$. With $k = 4$, to represent timescales up to hundreds of minutes, we would need to have time constants that go up to at least dozens of minutes. Slice electrophysiology work has shown populations of integrator cells in various parts of the medial temporal lobe that can maintain a stable firing rate for very long periods of time, up to at least tens of minutes (Egorov, Hamam, Fransén, Hasselmo, & Alonso, 2002; Fransén, Tahvildari, Egorov, Hasselmo, & Alonso, 2006). When provided an appropriate input, these integrator cells take on a new firing rate and again sustain it for tens of minutes. These properties reflect processes intrinsic to the cell, as they are observed under complete synaptic blockade. Integrator cells have been reported thus far in the entorhinal cortex (Egorov et al., 2002), the amygdala (Egorov, Unsicker, & von Bohlen und Halbach, 2006), the perirhinal cortex (Brown, 2008), and in interneurons in the hippocampus (Sheffield, Best, Mensh, Kath, & Spurston, 2011). Homologous cells that integrate inhibitory signals have been observed in the prefrontal cortex (Winograd, Destexhe, & Sanchez-Vives, 2008). It has been suggested that a range of time constants could be constructed from populations of integrator cells coupled with appropriate network properties (Howard, Fotedar, Datey, & Hasselmo, 2005), but this is not the only way leaky integrators could be constructed (see Guy & Tank, 2004, for a review comparing the intrinsic cellular mechanisms and network-level dynamics as underlying mechanisms for persistent activity).

TILT not only relies on the persistent activity of the integrator cells, but also on the exponential decay of their activity. Some interneurons in the hippocampus show a gradual decay in their firing rate for over 10 minutes (Sheffield et al., 2011). Neurons in prefrontal, parietal, and cingulate cortices have been found to satisfy exponential decay with time constants ranging from hundreds of milliseconds to tens of seconds (Bernacchia, Seo, Lee, & Wang, 2011). Interestingly, in that study, the number of neurons with longer time constants falls off as a power law function. If we knew that this population encoded **t**, this would constrain the number density along the s-axis. This would in turn constrain $g(\overset{*}{\tau})$ if s and $\overset{*}{\tau}$ are in one-to-one correspondence. Bernacchia et al. (2011) reported an exponent of -2 for the distribution of time constants; this would correspond to $g(\overset{*}{\tau}) = |\overset{*}{\tau}|^{-2}$.

6.3 Timing Models. Over the past few decades, researchers have developed many models of timing (Gibbon, Malapani, Dale, & Gallistel, 1997; Miall, 1996; Mauk & Buonomano, 2004; Ivry & Schlerf, 2008). A popular approach has been to propose an internal clock whose ticks are accumulated by a counter to represent perceived time. Different models use different mechanisms for the clock. Some use a pacemaker that emits pulses at regular intervals (Gibbon, 1977; Church, 1984; Killeen & Fetterman, 1988; Gallistel & Gibbon, 2000). Some use a population of neural oscillators of different frequencies (Church & Broadbent, 1990; Treisman, Faulkner, Naish, & Brogan, 1990; Miall, 1990). Some use a distributed idea of detecting the coincidental

activity of different neural populations to represent the ticks of the internal clock (Matell & Meck, 2004; Buhusi & Meck, 2005). In these models, the scale-invariance property is essentially tailored in by assuming the clocks to be intrinsically noisy. For example, Gibbon (1977) assumes a multiplicative noise variable underlying the pacemaker. Similarly Church and Broadbent (1990) take the noise in the different oscillators to be proportional to the respective frequencies.

Nonclock models of timing generally require a distributed population of specialized neural units that respond differently to different external stimuli. Some models in this category use tapped delay lines (Moore & Choi, 1997) or chained connectivity between late spiking neurons (Tieu et al., 1999), where the delays accumulated while traversing through each link of the chain add up, thereby making the different links of the chain respond to the external stimulus at different latencies. The existence of time cells is consistent with such models. The scale invariance in such a representation of time can be introduced by injecting low-frequency noise with appropriate properties into the chain (Tieu et al., 1999).

The spectral timing model (Grossberg & Schmajuk, 1989; Grossberg & Merrill, 1992) and the multi-timescale model (Staddon, Chelaru, & Higa, 2002) are more sophisticated examples of nonclock models. TILT can be most aptly placed in a category with these models. The spectral timing model proposes a population of neural units that span a spectrum of reaction rates in the relevant temporal range. A reinforcement learning procedure adaptively selects the appropriate units leading to well-timed behavior. The different decay rates of the **t** nodes in TILT are analogous to the spectrum of reaction rates in the spectral timing model. In this way, TILT shares the same fundamental spirit as the spectral timing model.

TILT shares an even closer resemblance with the multi-timescale (MTS) model. Following the activation by an external stimulus, the cascade of leaky integrators used in the MTS model decays exponentially, which is functionally identical to the **t** nodes in TILT. A distinguishing feature in MTS is that the external stimulus differentially activates different units. The activation of a unit following a stimulus is suppressed to an extent determined by the preexisting activities of other units. In the framework of TILT, this feature would be analogous to having a more complicated **C** operator in Figure 1 that activates the **t** nodes nonuniformly. By suitably choosing the rate constants of the different units in the population, both the spectral timing model and the MTS model yield an approximate Weber law behavior. Nevertheless, it has been hard to analytically pinpoint the features of these models that lead to explicit scale invariance in the internal representation of time.

6.4 Episodic Memory Models. Serial position effects in free recall played a major role in the development of models based on the distinction between short-term memory and long-term memory (Atkinson & Shiffrin, 1968; Raaijmakers & Shiffrin, 1980; Davelaar et al., 2005). It is quite evident

that the short-term store cannot account for recency effects across multiple timescales. Some models (Davelaar et al., 2005; Atkinson & Shiffrin, 1968) have adopted multiple memory stores at different scales to account for the recency effects observed at those different scales. Two recent models account for the recency effect with a single memory mechanism across scales, the scale-invariant memory and perceptual learning model, SIMPLE (Brown et al., 2007; Neath & Brown, 2006), and the temporal context model, TCM (Howard & Kahana, 2002a; Sederberg et al., 2008).

The SIMPLE model assumes that a memory trace for any item is represented along an internal time axis and that the time line is logarithmically compressed such that recently studied items are more discriminable. In some sense, SIMPLE assumes the existence of something like the **T** layer of TILT (see equation 5.21). It might be possible to adopt the internal representation of time and stimulus history proposed by TILT to give a more mechanistic interpretation of SIMPLE.

The connection between TCM and TILT is particularly strong, primarily because TILT is constructed based on TCM (Shankar & Howard, 2010). The gradually decaying activity of the **t** layer, the Hebbian association operator **M**, and the prediction vector **p** are all directly adopted from TCM (Shankar et al., 2009). The key difference between TCM and TILT is that the **t** layer of TCM consisted of nodes with just one decay rate, as opposed to a whole range of decay rates in TILT. As a consequence, the recency effect described by TCM shows an exponential decay as opposed to a power law decay in TILT and SIMPLE. This makes it impossible for TCM to predict scale-invariant recency effects (Howard, 2004).

Another robust effect observed in free recall experiments is that successively recalled items generally tend to have been studied in close temporal proximity, the contiguity effect (Kahana, 1996). Regardless of the scale we choose to define this "temporal proximity," the contiguity effect always seems to exist (Howard & Kahana, 1999; Howard et al., 2008; Moreton & Ward, 2010; Unsworth, 2008). TCM accounts for the contiguity effect using the idea of reinstatement of context, whereby the information recovered by recalled items cues for the subsequent recall attempts. Analogously, if the entire state of **t**-layer activity at the time of study of an item is reconstructed following the recall of that item, TILT should be able to account for the contiguity effect and its persistence across timescales.

Appendix A: Deriving the Differential Equation for T

We start with the equation

$$\mathbf{t}(\tau, s) = \int_{-\infty}^{\tau} \mathbf{f}(\tau') e^{s(\tau'-\tau)} d\tau' = e^{-s\tau} \int_{-\infty}^{\tau} \mathbf{f}(\tau') e^{s\tau'} d\tau' \tag{A.1}$$

$$\frac{\partial}{\partial \tau} \mathbf{t}(\tau, s) = -s.\mathbf{t}(\tau, s) + \mathbf{f}(\tau). \tag{A.2}$$

Differentiating k times with respect to s yields

$$\frac{\partial^k}{\partial s^k}\frac{\partial}{\partial \tau}\mathbf{t}(\tau, s) = -k\mathbf{t}^{(k-1)}(\tau, s) - s\mathbf{t}^{(k)}(\tau, s). \tag{A.3}$$

We now express $\mathbf{T}(\tau, \overset{*}{\tau})$ in terms of s instead of $\overset{*}{\tau}$. Equation 2.3 becomes

$$\mathbf{T}(\tau, s) = \frac{(-1)^k}{k!}s^{k+1}\mathbf{t}^{(k)}(\tau, s), \tag{A.4}$$

$$\frac{\partial \mathbf{T}(\tau, s)}{\partial \tau} = \frac{(-1)^k}{k!}s^{k+1}\frac{\partial}{\partial \tau}\mathbf{t}^{(k)}(\tau, s) = \frac{(-1)^k}{k!}s^{k+1}\frac{\partial^k}{\partial s^k}\frac{\partial}{\partial \tau}\mathbf{t}(\tau, s), \tag{A.5}$$

$$\frac{\partial \mathbf{T}(\tau, s)}{\partial \tau} = -s\mathbf{T}(\tau, s) - k\frac{(-1)^k}{k!}s^{k+1}\mathbf{t}^{(k-1)}(\tau, s), \tag{A.6}$$

$$\frac{1}{s^{k+1}}\left[\frac{\partial \mathbf{T}(\tau, s)}{\partial \tau} + s\mathbf{T}(\tau, s)\right] = -k\frac{(-1)^k}{k!}\mathbf{t}^{(k-1)}(\tau, s). \tag{A.7}$$

Differentiating with respect to s,

$$\frac{\partial}{\partial s}\left[\frac{1}{s^{k+1}}\left[\frac{\partial \mathbf{T}(\tau, s)}{\partial \tau} + s\mathbf{T}(\tau, s)\right]\right] = -k\frac{(-1)^k}{k!}\mathbf{t}^{(k)}(\tau, s), \tag{A.8}$$

$$\frac{1}{s^{k+1}}\frac{\partial}{\partial s}\left[\frac{\partial}{\partial \tau}\mathbf{T}(\tau, s) + s\mathbf{T}(\tau, s)\right] - \frac{k+1}{s^{k+2}}\left[\frac{\partial}{\partial \tau}\mathbf{T}(\tau, s) + s\mathbf{T}(\tau, s)\right]$$

$$= -\frac{k}{s^{k+1}}\mathbf{T}(\tau, s), \tag{A.9}$$

$$\frac{\partial^2}{\partial s \partial \tau}\mathbf{T}(\tau, s) + s\frac{\partial}{\partial s}\mathbf{T}(\tau, s) - \frac{k+1}{s}\frac{\partial}{\partial \tau}\mathbf{T}(\tau, s) = 0. \tag{A.10}$$

We now rewrite the equation in terms of $\overset{*}{\tau}$. Since $s = -k/\overset{*}{\tau}$, note that $\partial/\partial s = (\overset{*}{\tau}^2/k)\partial/\partial\overset{*}{\tau}$. Equation A.10 becomes

$$\frac{\overset{*}{\tau}^2}{k}\frac{\partial^2}{\partial \overset{*}{\tau}\partial \tau}\mathbf{T}(\tau, \overset{*}{\tau}) - \overset{*}{\tau}\frac{\partial}{\partial \overset{*}{\tau}}\mathbf{T}(\tau, \overset{*}{\tau}) + \left(\frac{k+1}{k}\right)\overset{*}{\tau}\frac{\partial}{\partial \tau}\mathbf{T}(\tau, \overset{*}{\tau}) = 0. \tag{A.11}$$

This is a second-order differential equation involving two variables $(\tau, \overset{*}{\tau})$. By prescribing the function on two boundaries, namely, $\overset{*}{\tau} = 0$ and $\tau = 0$, equation A.11 can be numerically solved. First note that by

differentiating $\mathbf{t}(\tau, s)$ explicitly k times with respect to s, we obtain

$$\mathbf{T}(\tau, s) = \frac{(-1)^k}{k!} s^{k+1} \int_{-\infty}^{\tau} (\tau' - \tau)^k \mathbf{f}(\tau') e^{s(\tau'-\tau)} d\tau'$$

$$= \frac{1}{k!} \int_0^{\infty} y^k \mathbf{f}(\tau + y/s) e^{-y} dy, \qquad \left\{ y = s(\tau - \tau') \right\}. \tag{A.12}$$

In the limit $s \to \infty$ ($\overset{*}{\tau} \to 0$), the above integral A.12 is simply $\mathbf{f}(\tau)$. Thus, we obtain one of our boundary conditions to be

$$\mathbf{T}(\tau, \overset{*}{\tau} = 0) = \mathbf{f}(\tau). \tag{A.13}$$

If the stimulus was never presented prior to $\tau = 0$, then we can assume that $\mathbf{T}(\tau = 0, \overset{*}{\tau}) = 0$, which will be the other boundary condition for solving the differential equation A.11.

Appendix B: Evaluating the Discretized kth Derivative

Let us calculate the derivatives of a function $F(s)$ at s_o on a discretized s-axis. Consider discrete steps of width $\Delta/2$ around the point of interest s_o. The first derivative of the function about s_o is given by

$$F^{(1)}(s_o) = \frac{F(s_o + \Delta/2) - F(s_o - \Delta/2)}{\Delta}. \tag{B.1}$$

If we have the kth derivative of the function at the neighboring points, the $(k+1)$th derivative at the point s_o is given by

$$F^{(k+1)}(s_o) = \frac{F^{(k)}(s_o + \Delta/2) - F^{(k)}(s_o - \Delta/2)}{\Delta}. \tag{B.2}$$

Formulas B.1 and B.2 will be accurate in the limit Δ goes to 0. With a nonzero Δ, this formula will yield an error of $\mathcal{O}(k\Delta^2)$ in the calculation of the kth derivative. Let us denote the true derivatives in the continuum by $\mathcal{F}^{(k)}(s_o)$. We now show that

$$F^{(k)}(s_o) - \mathcal{F}^{(k)}(s_o) = \mathcal{O}(k\Delta^2).$$

Taylor expansion gives

$$\begin{aligned}
&\Delta F^{(1)}(s_o)\\
&=\left(F(s_o)+(\Delta/2)\mathcal{F}^{(1)}(s_o)+\frac{1}{2}(\Delta/2)^2\mathcal{F}^{(2)}(s_o)+\frac{1}{3!}(\Delta/2)^3\mathcal{F}^{(3)}(s_o)+\ldots\right)\\
&-\left(F(s_o)-(\Delta/2)\mathcal{F}^{(1)}(s_o)+\frac{1}{2}(\Delta/2)^2\mathcal{F}^{(2)}(s_o)-\frac{1}{3!}(\Delta/2)^3\mathcal{F}^{(3)}(s_o)+\ldots\right)\\
&F^{(1)}(s_o)=\mathcal{F}^{(1)}(s_o)+\frac{\Delta^2}{2^2 3!}\mathcal{F}^{(3)}(s_o)+\frac{\Delta^4}{2^4 5!}\mathcal{F}^{(5)}(s_o)+\ldots \qquad \text{(B.3)}
\end{aligned}$$

More generally, it turns out that

$$F^{(k)}(s_o)=\mathcal{F}^{(k)}(s_o)+k\frac{\Delta^2}{2^2 3!}\mathcal{F}^{(k+2)}(s_o)+\Delta^4 f(s_o,k)+\ldots \qquad \text{(B.4)}$$

Here we are not interested in the functional form of $f(s_o,k)$. Equation B.4 can be shown using mathematical induction. We know that the above hypothesis is true for $k=1$. Assuming that it is true for an arbitrary k, we show that the hypothesis holds true for $k+1$. Since s_o is not a special point, the induction hypothesis, equation B.4, immediately leads to

$$\begin{aligned}
&F^{(k)}(s_o+\Delta/2)\\
&\quad=\mathcal{F}^{(k)}(s_o+\Delta/2)+k\frac{\Delta^2}{2^2 3!}\mathcal{F}^{(k+2)}(s_o+\Delta/2)+\Delta^4 f(s_o+\Delta/2,k)+\ldots\\
&F^{(k)}(s_o-\Delta/2)\\
&\quad=\mathcal{F}^{(k)}(s_o-\Delta/2)+k\frac{\Delta^2}{2^2 3!}\mathcal{F}^{(k+2)}(s_o-\Delta/2)+\Delta^4 f(s_o-\Delta/2,k)+\ldots
\end{aligned}$$

We now Taylor-expand the above equations and retain terms up to order Δ^4:

$$\begin{aligned}
F^{(k)}(s_o+\Delta/2)=&\left(\mathcal{F}^{(k)}(s_o)+(\Delta/2)\mathcal{F}^{(k+1)}(s_o)+\frac{1}{2}(\Delta/2)^2\mathcal{F}^{(k+2)}(s_o)\right.\\
&\left.+\frac{1}{3!}(\Delta/2)^3\mathcal{F}^{(k+3)}(s_o)+\frac{1}{4!}(\Delta/2)^4\mathcal{F}^{(k+4)}(s_o)+\ldots\right)\\
&+k\frac{\Delta^2}{2^2 3!}\left(\mathcal{F}^{(k+2)}(s_o)+(\Delta/2)\mathcal{F}^{(k+3)}(s_o)\right.\\
&\left.+\frac{1}{2}(\Delta/2)^2\mathcal{F}^{(k+4)}(s_o)+\ldots\right)\\
&+\Delta^4\left(f(s_o,k)+\Delta/2\,f'(s_o,k)+\ldots\right)+\ldots \qquad \text{(B.5)}
\end{aligned}$$

$$
\begin{aligned}
F^{(k)}(s_o + \Delta/2) = &\left(\mathcal{F}^{(k)}(s_o) - (\Delta/2)\mathcal{F}^{(k+1)}(s_o) + \frac{1}{2}(\Delta/2)^2 \mathcal{F}^{(k+2)}(s_o) \right. \\
&\left. - \frac{1}{3!}(\Delta/2)^3 \mathcal{F}^{(k+3)}(s_o) + \frac{1}{4!}(\Delta/2)^4 \mathcal{F}^{(k+4)}(s_o) + \ldots \right) \\
&+ k\frac{\Delta^2}{2^2 3!}\left(\mathcal{F}^{(k+2)}(s_o) - (\Delta/2)\mathcal{F}^{(k+3)}(s_o) \right. \\
&\left. + \frac{1}{2}(\Delta/2)^2 \mathcal{F}^{(k+4)}(s_o) + \ldots \right) \\
&+ \Delta^4 \left(f(s_o, k) - \Delta/2 f'(s_o, k) + \ldots \right) + \ldots
\end{aligned} \tag{B.6}
$$

Subtracting the above two equations and applying equation B.2 yields

$$
\Delta F^{(k+1)}(s_o) = \Delta \mathcal{F}^{(k+1)}(s_o) + (k+1)\frac{\Delta^3}{2^2 3!}\mathcal{F}^{(k+3)}(s_o) + \mathcal{O}(\Delta^5). \tag{B.7}
$$

This shows that equation B.4 is true for all k. The error in estimation of the kth derivative of a function is $\mathcal{O}(k\Delta^2)$:

$$
F^{(k)}(s_o) - \mathcal{F}^{(k)}(s_o) = k\frac{\Delta^2}{2^2 3!}\mathcal{F}^{(k+2)}(s_o) + \mathcal{O}(\Delta^4) = \mathcal{O}(k\Delta^2). \tag{B.8}
$$

By successively reexpressing the kth derivative in terms of lower derivatives of the function, we can ultimately express the kth derivative of the function purely in terms of the functional values in a region around s_o. That is, $F^{(k)}(s_o)$ can be expressed as a linear combination of the functional values around s_o. For convenience, we define the functions

$$
\begin{aligned}
E_0 &= F(s_o), \\
E_n &= F(s_o + n\Delta/2) + F(s_o - n\Delta/2), \\
O_n &= F(s_o + n\Delta/2) - F(s_o - n\Delta/2).
\end{aligned} \tag{B.9}
$$

The derivatives of these functions have the property that

$$
\begin{aligned}
\Delta E_0' &= O_1, \\
\Delta E_n' &= O_{n+1} - O_{n-1}, \\
\Delta O_n' &= E_{n+1} - E_{n-1}, \\
\Delta O_1' &= E_2 - 2E_0.
\end{aligned} \tag{B.10}
$$

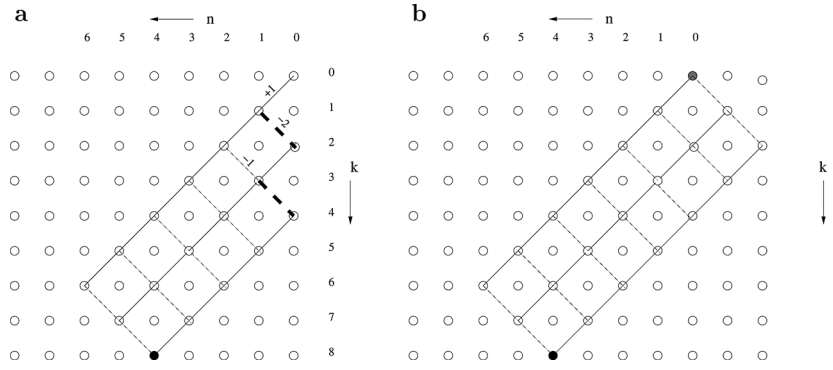

Figure 14: (a) To compute the coefficient of E_n or O_n in $F^{(k)}(s_o)$, choose the (n, k) node in the graph. In this example, $k = 8$ and $n = 4$. Start from $(0, 0)$ and traverse the nodes diagonally in the downward direction. Traversing the leftward diagonal accumulates a factor of +1 (solid lines), while traversing the rightward diagonal generally accumulates a factor of -1 (dashed lines). But when the rightward diagonal link hits the boundary at $n = 0$, it accumulates a factor of -2 (thick dashed lines). These rules basically summarize equation B.10. There can be many distinct routes to reach (n, k) from $(0, 0)$, and each path will carry a weight in accordance with the rules. The sum of the weights of the distinct paths is essentially the coefficient of E_n or O_n in equations B.11 and B.12. (b) Add columns of nodes on the other side of $n = 0$ so as to complete a rectangle. Change the rules such that every leftward diagonal link holds a factor of $+1$ and every rightward diagonal link holds a factor of -1. Acknowledging the equivalence between panels a and b is very useful. In panel b, every path always has a weight of either $+1$ or -1, depending on whether $(k - n)$ is even or odd, respectively. For a given (n, k), the number of distinct paths connecting it to $(0, 0)$ is a simple combinatorics exercise that leads to equations B.11 and B.12.

Starting with E_0 and successively taking k derivatives yields the kth derivative as a linear combination of various E_n's and O_n's. It turns out that when k is even,

$$F^{(k)}(s_o) = \sum_{r=0}^{r=k/2} (-1)^r \Delta^{-k} \frac{k!}{r!(k-r)!} E_{k-2r}, \tag{B.11}$$

and when k is odd,

$$F^{(k)}(s_o) = \sum_{r=0}^{r=(k-1)/2} (-1)^r \Delta^{-k} \frac{k!}{r!(k-r)!} O_{k-2r}. \tag{B.12}$$

A graphical method of computing the coefficients in front of the E_n's and the O_n's in equations B.11 and B.12 are described in Figure 14.

References

Atkinson, R. C., & Shiffrin, R. M. (1968). Human memory: A proposed system and its control processes. In K. W. Spence & J. T. Spence (Eds.), *The psychology of learning and motivation* (Vol. 2, pp. 89–105). New York: Academic Press.

Balsam, P. D., & Gallistel, C. R. (2009). Temporal maps and informativeness in associative learning. *Trends in Neuroscience*, *32*(2), 73–78.

Bernacchia, A., Seo, H., Lee, D., & Wang, X. (2011). A reservoir of time constants for memory traces in cortical neurons. *Nature Neuroscience*, *14*, 366–372.

Bjork, R. A., & Whitten, W. B. (1974). Recency-sensitive retrieval processes in long-term free recall. *Cognitive Psychology*, *6*, 173–189.

Brown, G. D., Neath, I., & Chater, N. (2007). A temporal ratio model of memory. *Psychological Review*, *114*(3), 539–576.

Brown, T. H. (2008, February). *Persistent multistable firing in perirhinal cortex*. Paper presented at the 20th Annual Meeting of the Winter Conference on Neural Plasticity, St. Lucia.

Buhusi, C. V., & Meck, W. H. (2005). What makes us tick? Functional and neural mechanisms of interval timing. *Nature Reviews Neuroscience*, *6*, 755–765.

Chater, N., & Brown, G. (2008). From universal laws of cognition to specific cognitive models. *Cognitive Science*, *32*(1), 36–67.

Church, R. M. (1984). Properties of the internal clock. In J. Gibbon & L. Allan (Eds.), *Timing and time perception* (pp. 566–582). New York: New York Academy of Sciences.

Church, R. M., & Broadbent, H. (1990). Alternative representation of time, number and rate. *Cognition*, *37*, 55–81.

Davelaar, E. J., Goshen-Gottstein, Y., Ashkenazi, A., Haarmann, H. J., & Usher, M. (2005). The demise of short-term memory revisited: Empirical and computational investigations of recency effects. *Psychological Review*, *112*(1), 3–42.

Desmond, J. E., & Moore, J. W. (1988). Adaptive timing in neural network: The conditioned response. *Biological Cybernetics*, *58*, 405–415.

Drew, M. R., Couvillon, P. A., Zupan, B., Cooke, A., & Balsam, P. (2005). Temporal control of conditioned responding in goldfish. *Journal of Experimental Psychology: Animal Behavior Processes*, *31*, 31–39.

Egorov, A. V., Hamam, B. N., Fransén, E., Hasselmo, M. E., & Alonso, A. A. (2002). Graded persistent activity in entorhinal cortex neurons. *Nature*, *420*(6912), 173–178.

Egorov, A. V., Unsicker, K., & von Bohlen und Halbach, O. (2006). Muscarinic control of graded persistent activity in lateral amygdala neurons. *European Journal of Neuroscience*, *24*(11), 3183–3194.

Fransén, E., Tahvildari, B., Egorov, A. V., Hasselmo, M. E., & Alonso, A. A. (2006). Mechanism of graded persistent cellular activity of entorhinal cortex layer V neurons. *Neuron*, *49*(5), 735–746.

Gallistel, C. R., & Gibbon, J. (2000). Time, rate, and conditioning. *Psychological Review*, 1072, 289–344.

Gibbon, J. (1977). Scalar expectancy theory and Weber's law in animal timing. *Psychological Review*, *84*(3), 279–325.

Gibbon, J., Malapani, C., Dale, C. L., & Gallistel, C. R. (1997). Toward a neurobiology of temporal cognition: Advances and challanges. *Current Opinion in Neurobiology*, *7*, 170–184.

Glanzer, M., & Cunitz, A. R. (1966). Two storage mechanisms in free recall. *Journal of Verbal Learning and Verbal Behavior*, *5*, 351–360.

Glenberg, A. M., Bradley, M. M., Stevenson, J. A., Kraus, T. A., Tkachuk, M. J., & Gretz, A. L. (1980). A two-process account of long-term serial position effects. *Journal of Experimental Psychology: Human Learning and Memory*, *6*, 355–369.

Grossberg, S., & Merrill, J. (1992). A neural network model of adaptively timed reinforcement learning and hippocampal dynamics. *Cognitive Brain Research*, *1*, 3–38.

Grossberg, S., & Schmajuk, N. A. (1989). Neural dynamics of adaptive timing and temporal discrimination during associative learning. *Neural Networks*, *2*, 79–102.

Guy, M., & Tank, D. (2004). Persistent neural activity: Prevalence and mechanisms. *Current Opinion in Neurobiology*, *14*, 675–684.

Hasselmo, M. (2009). A model of episodic memory: Mental time travel along encoded trajectories using grid cells. *Neurobiology of Learning and Memory*, *92*, 559–573.

Howard, M. W. (2004). Scaling behavior in the temporal context model. *Journal of Mathematical Psychology*, *48*, 230–238.

Howard, M. W., Addis, K. A., Jing, B., & Kahana, M. (2007). Semantic structure and episodic memory. In T. K. Landauer, D. S. McNamara, S. Dennis, & W. Kintsch (Eds.), *LSA: A road towards meaning* (pp. 121–141). Mahwah, NJ: Erlbaum.

Howard, M. W., Fotedar, M. S., Datey, A. V., & Hasselmo, M. E. (2005). The temporal context model in spatial navigation and relational learning: Toward a common explanation of medial temporal lobe function across domains. *Psychological Review*, *112*(1), 75–116.

Howard, M. W., & Kahana, M. J. (1999). Contextual variability and serial position effects in free recall. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *25*, 923–941.

Howard, M. W., & Kahana, M. J. (2002a). A distributed representation of temporal context. *Journal of Mathematical Psychology*, *46*(3), 269–299.

Howard, M. W., & Kahana, M. J. (2002b). When does semantic similarity help episodic retrieval. *Journal of Memory and Language*, *46*(1), 85–98.

Howard, M. W., Shankar, K. H., & Jagadisan, U.K.K. (2011). Constructing semantic representations from a gradually-changing representation of temporal context. *Topics in Cognitive Science*, *3*, 48–73.

Howard, M. W., Youker, T. E., & Venkatadass, V. (2008). The persistence of memory: Contiguity effects across several minutes. *Psychonomic Bulletin and Review*, *15*, 58–63.

Itskov, V., Curto, C., Pastalkova, E., & Buzsaki, G. (2011). Cell assembly sequences arising from spike threshold adaptation keep track of time in the hippocampus. *Journal of Neuroscience*, *31*(8), 2828–2834.

Ivry, R. B., & Hazeltine, R. E. (1995). Perception and production of temporal intervals across a range of durations: Evidence for a common timing mechanism. *Journal of Experimental Psychology: Human Perception and Performance*, *21*, 1–12.

Ivry, R. B., & Schlerf, J. E. (2008). Dedicated and intrinsic models of time perception. *Trends in Cognitive Science*, *12*, 273–280.

Jensen, O., & Lisman, J. E. (1996). Hippocampal CA3 region predicts memory sequences: Accounting for the phase precession of place cells. *Learning and Memory*, *3*, 279–287.

Kahana, M. J. (1996). Associative retrieval processes in free recall. *Memory and Cognition*, *24*, 103–109.

Killeen, P. R., & Fetterman, G. J. (1988). A behavioral theory of timing. *Psychological Review*, *95*(2), 274–295.

MacDonald, C. J., Lepage, K. O., Eden, U. T., & Eichenbaum, H. (2011). *Hippocampal time cells bridge the gap in memory for discontiguous events*. *Neuron*, *71*(4), 737–749.

Matell, M. S., & Meck, W. H. (2004). Cortico-striatal circuits and interval timing: Coincidence detection of oscillatory processes. *Cognitive Brain Research*, *21*, 139–170.

Mauk, M. D., & Buonomano, D. V. (2004). The neural basis of temporal processing. *Annual Review of Neuroscience*, *27*, 307–340.

Miall, R. C. (1990). The storage of time intervals using oscillating neurons. *Neural Computation*, *37*, 55–81.

Miall, R. C. (1996). Models of neural timing. In M. A. Pastor & J. Artieda (Eds.), *Time, internal clocks and movements* (pp. 69–94). Amsterdam: Elsevier Science.

Moore, J. W., & Choi, J. S. (1997). Conditioned response timing and integration in cerebellum. *Learning and Memory*, *4*, 116–129.

Moreton, B. J., & Ward, G. (2010). Time scale similarity and long-term memory for autobiographical events. *Psychonomic Bulletin and Review*, *17*, 510–515.

Murdock, B. B., & Okada, R. (1970). Interresponse times in single-trial free recall. *Journal of Verbal Learning and Verbal Behavior*, *86*, 263–267.

Nairne, J. S., Neath, I., Serra, M., & Byun, E. (1997). Positional distinctiveness and the ratio rule in free recall. *Journal of Memory and Language*, *37*, 155–166.

Neath, I., & Brown, G.D.A. (2006). SIMPLE: Further applications of a local distinctiveness model of memory. In B. H. Ross (Ed.), *The psychology of learning and motivation.* San Diego, CA: Academic Press.

Pastalkova, E., Itskov, V., Amarasingham, A., & Buzsaki, G. (2008). Internally generated cell assembly sequences in the rat hippocampus. *Science*, *321*(5894), 1322–1327.

Polyn, S. M., Norman, K. A., & Kahana, M. J. (2009). A context maintenance and retrieval model of organizational processes in free recall. *Psychological Review*, *116*, 129–156.

Post, E. (1930). Generalized differentiation. *Transactions of the American Mathematical Society*, *32*, 723–781.

Postman, L., & Phillips, L. W. (1965). Short-term temporal changes in free recall. *Quarterly Journal of Experimental Psychology*, *17*, 132–138.

Raaijmakers, J. G. W., & Shiffrin, R. M. (1980). SAM: A theory of probabilistic search of associative memory. In G. H. Bower (Ed.), *The psychology of learning and motivation: Advances in research and theory* (Vol. 14, pp. 207–262). New York: Academic Press.

Rakitin, B. C., Gibbon, J., Penny, T. B., Malapani, C., Hinton, S. C., & Meck, W. H. (1998). Scalar expectancy theory and peak-interval timing in humans. *Journal of Experimental Psychology: Animal Behavior Processes*, *24*, 15–33.

Roberts, S. (1981). Isolation of an internal clock. *Journal of Experimental Psychology: Animal Behavior Processes*, *7*, 242–268.

Sederberg, P. B., Howard, M. W., & Kahana, M. J. (2008). A context-based theory of recency and contiguity in free recall. *Psychological Review*, *115*, 893–912.

Shankar, K. H., & Howard, M. W. (2010). Timing using temporal context. *Brain Research*, *1365*, 3–17.

Shankar, K. H., Jagadisan, U.K.K., & Howard, M. W. (2009). Sequential learning using temporal context. *Journal of Mathematical Psychology*, *53*, 474–485.

Sheffield, M.E.J., Best, T. K., Mensh, B. D., Kath, W. L., & Spurston, N. (2011). Slow integration leads to persistent action potential firing in distal axons of coupled interneurons. *Nature Neuroscience*, *14*, 200–209.

Shepard, R. N. (1987). Toward a universal law of generalization for psychological science. *Science*, *237*, 1317–1323.

Smith, M. C. (1968). CS-US interval and US intensity in classical conditioning of rabbit's nictitating membrane response. *Journal of Comparative and Physiological Psychology*, *3*, 679–687.

Staddon, J. E., Chelaru, I. M., & Higa, J. J. (2002). Habituation, memory and the brain: The dynamics of interval timing. *Behavioural Processes*, *57*, 71–88.

Tieu, K. H., Keidel, A. L., McGann, J. P., Faulkner, B., & Brown, T. H. (1999). Perirhinal-amygdala circuit level computational model of temporal encoding in fear conditioning. *Psychobiology*, *27*, 1–25.

Treisman, M. (1963). Temporal discrimination and the indifference interval: Implications for a model of the internal clock. *Psychological Monographs: General and Applied*, *77*(576), 1–31.

Treisman, M., Faulkner, A., Naish, P. L., & Brogan, D. (1990). The internal clock: Evidence for a temporal oscillator underlying time perception with some estimates of its characteristic frequency. *Perception*, *19*, 705–743.

Unsworth, N. (2008). Exploring the retrieval dynamics of delayed and final free recall: Further evidence for temporal-contextual search. *Journal of Memory and Language*, *59*, 223–236.

Wearden, J. H., & Lejeune, H. (2008). Scalar properties in human timing: Conformity and violations. *Quarterly Journal of Experimental Psychology*, *61*, 569–587.

Winograd, M., Destexhe, A., & Sanchez-Vives, M. V. (2008). Hyperpolarization-activated graded persistent activity in the prefrontal cortex. *Proceedings of the National Academy of Science, USA*, *105*(20), 7298–7303.

Wixted, J. T. (2004). The psychology and neuroscience of forgetting. *Annual Review of Psychology*, *55*, 235–269.

Received September 16, 2010; accepted June 13, 2011.