

The Successor Representation and Temporal Context

Samuel J. Gershman¹, Christopher D. Moore¹, Michael T. Todd¹, Kenneth A. Norman¹, Per B. Sederberg²

¹ Department of Psychology and Princeton Neuroscience Institute, Princeton University

² Department of Psychology, The Ohio State University

Address for correspondence:

Per B. Sederberg

Department of Psychology

The Ohio State University

Columbus, OH 43214

E-mail: sederberg.1@osu.edu

Abstract

The successor representation was introduced into reinforcement learning by Dayan (1993) as a means of facilitating generalization between states with similar successors. Although reinforcement learning in general has been used extensively as a model of psychological and neural processes, the psychological validity of the successor representation has yet to be explored. An interesting possibility is that the successor representation can be used not only for reinforcement learning, but for episodic learning as well. Our main contribution is to show that a variant of the Temporal Context Model (TCM; Howard and Kahana, 2002), an influential model of episodic memory, can be understood as directly estimating the successor representation using the temporal difference learning algorithm (Sutton and Barto, 1998). This insight leads to a generalization of TCM and new experimental predictions. In addition to casting a new normative light on TCM, this equivalence suggests a previously unexplored point of contact between different learning systems.

KEYWORDS: successor representation, temporal context, temporal difference learning, free recall

1 Introduction

Many learning problems have in common the goal of inferring properties of temporally extended sequences, a much thornier computational problem compared to simply predicting what will happen next. For example, a standard task in reinforcement learning (RL; Sutton and Barto, 1998) is to estimate the expected discounted future return (cumulative reward). Similarly, there is a sense in which understanding natural language involves predicting future words: Evidence suggests that people anticipate long-distance dependencies between words (Altmann and Kamide, 1999; Kamide et al., 2003). Perhaps less obviously, episodic memory also involves a predictive component: Plans and scripts retrieved from memory represent predictions about future events (Schank, 1982; Atance and O’Neill, 2001; Schacter et al., 2007).

An important—and computationally tractable—class of such learning problems arises when the environmental dynamics are specified by a Markov chain (a key assumption in RL theory; see Sutton and Barto, 1998) and the property to be inferred has a particular functional form that we describe below. Markovian dynamics arise when the next state of the environment depends *only* on the current state—that is, the state transition distribution is *memoryless*. In this case, calculations are greatly simplified by using an estimate of the expected discounted number of times each state j is visited following a visit to state i . The matrix of these expected discounted visitations was introduced by Dayan (1993) into RL as the *successor representation* (SR), and subsequently more fully explored by White (1995). The SR is essentially identical to the *fundamental matrix* in the theory of Markov chains (Kemeny and Snell, 1976).¹ One simple and effective way to estimate the SR is using temporal difference (TD) learning (Dayan, 1993; White, 1995), an error-driven learning algorithm that incrementally updates an estimate of the SR based on sample paths.

In this paper, we move beyond RL to show how the SR can function as a common computational substrate for multiple forms of learning. In particular, we draw a formal connection between TD learning of the SR and an influential model of episodic memory, the Temporal Context Model (TCM; Howard and Kahana, 2002; Howard et al., 2005; Sederberg et al., 2008; Polyn et al., 2009; Socher et al., 2009), resulting in a generalized form of TCM. The SR interpretation of TCM makes explicit its connections to normative models: TCM (in its generalized form) is a TD algorithm for estimating the SR. This new interpretation in terms of the well-understood SR allows us to generate several novel experimental predictions.

The rest of this paper is organized as follows. In section 2 we formally specify the problem to be solved and provide a mathematical description of the SR. Section 3 presents a TD algorithm for learning the SR. In section 4, we describe TCM and show that it is equivalent to (and a special case of) TD learning of the SR. Then in section 5 we explore the empirical implications of the model. Finally, in section 6 we discuss our model in the wider context of work on memory-based predictions and error-driven learning in the brain.

2 Markov chains and the successor representation

Consider a discrete, finite state space $\mathcal{S} = \{1, \dots, S\}$ and a state variable $s \in \mathcal{S}$ that evolves in discrete time according to a first-order Markov chain: $P(s_{n+1} = j | s_n = i) = T_{ij}$. We refer to \mathbf{T} as

¹It is also closely related to the inverse Laplacian in graph theory. See Mahadevan and Maggioni (2007) for discussion in the context of RL.

the *transition matrix*. We are interested in calculating functions of the following form:

$$\mathcal{F}(i) = \mathbb{E} \left[\sum_{p=0}^{\infty} \gamma^p \phi(s_{n+p+1}) \middle| s_n = i \right], \quad (1)$$

where $\gamma \in [0, 1]$ is a *discount factor* and $\phi : \mathcal{S} \rightarrow \mathbb{R}$ is an *emission function*. Different choices of ϕ lead to different characterizations of the Markov chain. For example, if $\phi(s)$ specifies the reward received upon entering state s , then $\mathcal{F}(s)$ corresponds to the expected discounted future return, or value, the standard target of RL (Sutton and Barto, 1998).

A felicitous property of Markov chains is that such functions can be calculated analytically:

$$\mathcal{F}(i) = \sum_{j=1}^S M_{ij} \phi(j), \quad (2)$$

where M is the successor representation (SR), which encodes the expected discounted future visitations of each state j along trajectories originating in state i :

$$\begin{aligned} M_{ij} &= \mathbb{E} \left[\sum_{p=0}^{\infty} \gamma^p \delta(s_{n+p+1}, j) \middle| s_n = i \right] \\ &= \mathbb{E}[\delta(k, j) + \gamma M_{kj} | s_n = i] \\ &= \sum_k T_{ik} [\delta(k, j) + \gamma M_{kj}] \end{aligned} \quad (3)$$

where $\delta(\cdot, \cdot)$ is the Kronecker delta function whose value is 1 if its arguments are equal and 0 otherwise, and k indexes all of the possible states that can follow i . The recursive expression is analogous to Bellman’s equation for value functions in RL (Bellman, 1957; Sutton and Barto, 1998). An important difference between the SR and the transition matrix is that the latter only looks at the next state, whereas the SR predicts visitations within an extended future window whose effective size is determined by γ .

3 Temporal difference learning of the successor representation

We have seen that the SR renders calculation of a certain class of functions (e.g., expected discounted future return) a linear operation, but how is it learned? Although the SR can be calculated directly from the transition matrix, an agent would still have to learn the transition matrix and then perform an expensive matrix inversion. Here we consider a temporal difference (TD) learning algorithm (Sutton, 1988) that estimates M directly from sample paths (see White, 1995, for more information). The basic idea behind this algorithm, which we refer to as TD-SR, is to minimize the discrepancy (prediction error) between predicted and observed visitation counts.

At each timepoint n , a transition $s_n \rightarrow s_{n+1}$ is observed. For example, in verbal memory experiments (as we describe in the next section), s_n might represent a word item presented on trial n . The “eligibility” $e_n(s_n)$ of state s_n is boosted, while the eligibility of other states is decayed:

$$e_n(i) = \begin{cases} \gamma \lambda e_{n-1}(i) & \text{if } i \neq s_n \\ \gamma \lambda e_{n-1}(i) + 1 & \text{if } i = s_n, \end{cases} \quad (4)$$

where $\lambda \in [0, 1]$ is a trace decay parameter and $e_0(i) = 0$. The eligibility trace vector stores a short-term memory of which items were recently experienced; in dynamical systems terminology, it can be viewed as a leaky integrator. Intuitively, the observed transition will have a greater effect on updating predictions for more eligible states. Large values of λ allow states to be eligible for longer periods of time. When $\lambda = 0$, only the most recently visited state is eligible for updating.

Using $\hat{\mathbf{M}}$ to denote the SR estimate, each component \hat{M}_{ij} is updated according to a TD rule:

$$\hat{M}_{ij} \leftarrow \hat{M}_{ij} + \alpha \left[\delta(s_{n+1}, j) + \gamma \hat{M}_{s_{n+1}, j} - \hat{M}_{s_n, j} \right] e_n(i), \quad (5)$$

where $\alpha \in [0, 1]$ is a learning rate, and the term in brackets is the *prediction error*, the discrepancy between the predicted and observed transition. To understand intuitively why this update rule converges to \mathbf{M} , notice that the recursive definition of the SR (Eq. 3) implies a consistency condition between the SR for successive states. Specifically, it implies that when $\hat{\mathbf{M}} = \mathbf{M}$ we can decompose the expectation of the prediction error as follows:

$$\mathbb{E}[\delta(k, j) + \gamma M_{kj} - M_{s_n, j} | s_n = i] = \mathbb{E}[\delta(k, j) + \gamma M_{kj} | s_n = i] - M_{ij} = 0. \quad (6)$$

The prediction error thus stochastically signals the extent to which this consistency condition has been violated by a transition. If the prediction error is positive, then \hat{M}_{ij} has underestimated future visitations and is therefore increased; in contrast, if the prediction error is negative, then \hat{M}_{ij} has overestimated future visitations and is therefore decreased.

As an example, consider the sequence

A B B B B A C

where A, B and C denote states. Assuming that \hat{M}_{ij} is initialized to 0, there will be a positive prediction error after observing the transition from $A \rightarrow B$, which (according to Eq. 5) will strengthen the \hat{M}_{AB} association. In addition, the eligibility trace $e_n(A) = 1$ after the first transition, and the eligibility of A subsequently decays slowly after each following transition. For each $B \rightarrow B$ transition, M_{AB} continues to be strengthened, despite not observing any more $A \rightarrow B$ transitions:

$$\Delta \hat{M}_{AB} = \alpha \left[1 + \gamma \hat{M}_{BB} - \hat{M}_{AB} \right] e_n(A) > 0. \quad (7)$$

The prediction error will always be greater than zero after $B \rightarrow B$ transitions because the proximal association \hat{M}_{BB} is strengthened more than the distal association \hat{M}_{AB} . When the $A \rightarrow C$ transition is observed, the prediction error will be negative:

$$\Delta \hat{M}_{AB} = \alpha \left[0 + \gamma \hat{M}_{CB} - \hat{M}_{AB} \right] e_n(A) = -\alpha \hat{M}_{AB} e_n(A) < 0. \quad (8)$$

The prediction error is negative in this case because \hat{M}_{CB} is still 0 upon the first observed $A \rightarrow C$ transition, whereas $\hat{M}_{AB} > 0$. This example illustrates an important property of the TD-SR algorithm: changes in the association between two items do not depend solely on direct transitions between them. We shall exploit this property when we present new experimental predictions in Section 5.

The learning algorithm presented here can be easily generalized to states associated with feature vectors (White, 1995). In this case, the successor representation encodes the expected discounted visitations to future *features* rather than *states*. That is, M_{ij} encodes the expected discounted number of times feature j will be active following a visit to state i .

4 The Temporal Context Model

TCM was originally designed to describe the associative processes that underlie episodic memory and applied to human behavior in free recall experiments (for extensions of the TCM framework to semantic learning, see Howard et al., 2011; Shankar et al., 2009). In a free recall experiment, participants study a list of items presented sequentially and are then asked to recall the items in any order. Thus, “states” in this setting correspond to list items (usually words).

We will first briefly review some of the basic empirical findings from free recall experiments motivating TCM (for a comprehensive review, see Kahana et al., 2008). Then we formally describe the model and its relationship to TD learning of the SR.

4.1 Empirical background: free recall and episodic memory

One of the most important observations about free recall is that although the task itself does not demand temporal ordering of responses (unlike serial recall), human recall behavior appears to recapitulate the temporal structure of the study-list. In particular, Kahana (1996) measured recall contingencies: If the previously-recalled item was studied at serial position j , what is the probability that the next-recalled item will be from serial position $j + lag$? This measure, known as the lag conditional response probability (CRP), revealed two properties of free recall dynamics:

- *The temporal contiguity effect*: transitions tend to occur between items that were presented in nearby serial positions in the study-list.
- *The asymmetry effect*: forward transitions are more frequent than backward transitions.

In addition, Kahana (1996) demonstrated a temporal contiguity effect in conditional response latency: inter-response times tend to increase as a function of lag. A recent meta-analysis by Sederberg et al. (2010) confirmed that these properties of free recall are robust across numerous studies (Figure 1). Moreover, the size of the temporal contiguity effect predicts recall performance across subjects.

The key principle underlying TCM that allows it to explain these findings is the construct of *temporal context*: a slowly changing representation of recently experienced items. When an item is experienced at time step $n + 1$, the association between that item and currently-active contextual features (from time step n) is strengthened; next, the context vector is drifted incrementally towards the just-presented item’s representation. This conceptualization of temporal context as a recency-weighted average of experienced items distinguishes TCM from earlier models of memory that posited randomly fluctuating context representations (Estes, 1955; Anderson and Bower, 1972; Mensink and Raaijmakers, 1988).

During memory retrieval, the current state of the context vector is used as a retrieval cue. Items are sampled from memory according to how well the context cue (at test) matches the context that was associated with the item at study. When an item is retrieved at test, the context vector is updated using the features of the retrieved item and also the *context that was associated with that item at study*; the latter update can be construed as mentally “jumping back in time” to the moment when the just-retrieved item was studied. The updated context is then used to cue for more items. This combination of contextual cuing and temporal context reinstatement can account for the temporal contiguity effect, based on the principle that items studied close in time will have similar temporal contexts. When an item is retrieved, reinstated temporal context associated with that item

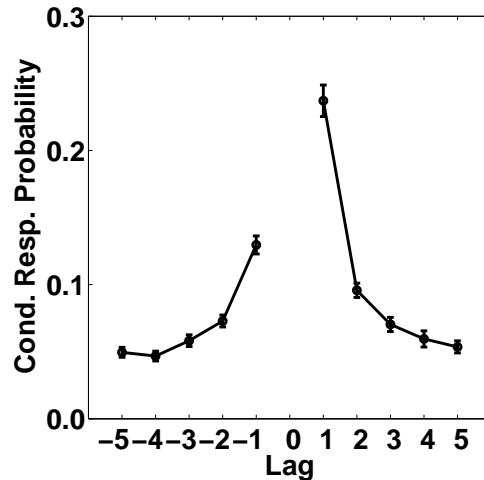


Figure 1: **The lag-CRP.** Graph of the conditional response probability as a function of serial position lag, averaged across nine delayed free recall studies. Reproduced from Sederberg et al. (2010).

will (symmetrically) match the contexts that were associated with items studied *before* and *after* the just-retrieved item, making it more likely that these items will be recalled. The asymmetry effect arises from the fact that (at study) the context vector is updated with item information when that item was presented. When the features of that item are used as a retrieval cue, the temporal context for subsequently-studied items will include the retrieved item’s features, but the temporal context for preceding items will not include these features. Thus, during recall the forward-asymmetric cue of the just-retrieved item and the bidirectional cue of its reinstated context combine to produce the ubiquitous pattern of recall transitions captured by the conditional response probability in Figure 1. To date, TCM is the only published model that has been shown to fit the contiguity and the asymmetry effect of the CRP without the addition of a special mechanism that was devised explicitly to fit the shape of the curve.

Temporal context effects manifest themselves in other ways, as well. For example, the *recency effect* refers to the observation that subjects tend to initiate recall with items that were presented at the end of the study-list (Deese and Kaufman, 1957); interposing a distractor task between study and test dramatically reduces this effect (Howard and Kahana, 1999). As with the contiguity effect, TCM explains the recency effect in terms of associations between the items and the drifting temporal context: the temporal context at test is more similar to items in late serial positions than to items in early serial positions. The distractor task attenuates this similarity by perturbing the context vector away from the end-of-list context, decreasing the overlap of the test context with the context bound to the items on the list. This explanation of recency effects is consistent with the principle of temporal distinctiveness posited by many previous models (e.g., Glenberg and Swanson, 1986; Brown et al., 2007): recall of an item depends on its relative recency to other list items, not its absolute recency. The temporal context vector provides a mechanistic implementation of this abstract principle.

4.2 Formal description of TCM

We shall focus on the encoding operations of TCM. The two key components of TCM are the context vector \mathbf{t}_n and the matrix of item-context associations, \mathbf{M} . Our notation is deliberately suggestive: as we will show, TCM’s learning rule for the item-context association matrix is equivalent to TD(λ) learning of the SR.

The context vector \mathbf{t}_n encodes a slowly-changing representation of temporal context, operationally defined as a superposition of recently experienced items. Formally, \mathbf{t}_n is updated as follows:

$$\mathbf{t}_n = \rho \mathbf{t}_{n-1} + \mathbf{f}_n, \quad (9)$$

where \mathbf{f}_n is an “item vector” typically assumed to be a binary unit vector with a $f_{ni} = 1$ if item i is experienced at time n , and 0 otherwise.² The parameter $\rho \in [0, 1]$ determines the drift rate of the context vector. In TCM, the context vector is used to cue memory retrieval, such that items experienced in a similar context are more likely to be retrieved; we will not discuss the recall operations further in this paper (see Howard and Kahana, 2002; Sederberg et al., 2008).

The item-context associations are updated using an outer-product Hebbian learning rule:

$$\hat{M}_{ij} \leftarrow \hat{M}_{ij} + \alpha f_{n+1,j} t_{ni} \quad (10)$$

where α is a learning rate parameter. This learning rule serves to bind the new item \mathbf{f}_{n+1} to the context that was present when it was experienced \mathbf{t}_n . This learning rule is equivalent to TD(λ) learning of the SR in the special case that items are only presented once. To see this, notice that if \mathbf{M} is initialized to all zeroes, then upon the first presentation of item i the second and third terms in the prediction error part of Equation 5 ($\gamma \hat{M}_{s_{n+1},j} - \hat{M}_{s_n,j}$) will be equal to 0 (if items are only presented once, then neither the current item s_n nor the following item s_{n+1} will have been presented previously, so \hat{M} will not have been updated for these items). Thus, in this case the TD update reduces to:

$$\hat{M}_{ij} \leftarrow \hat{M}_{ij} + \alpha \delta(s_{n+1}, j) e_n(i), \quad (11)$$

which corresponds to Equation 10 with the eligibility trace mapping onto the context vector ($e_n(i) = t_{ni}$) and the delta function mapping onto the item vector ($\delta(s_{n+1}, j) = f_{n+1,j}$). The dependence on ρ , λ and γ is left implicit in these equations; equivalence between the two updates is obtained when $\rho = \lambda\gamma$. In other words, the context drift rate ρ correspondingly governs the decay rate of the eligibility trace. In the more general case (in which items are presented more than once), the TD and Hebbian updates diverge, since the second and third terms in the prediction error are no longer 0. Specifically, TD replaces $\delta(s_{n+1}, j)$ with $\delta(s_{n+1}, j) + \gamma \hat{M}_{s_{n+1},j} - \hat{M}_{s_n,j}$. In essence, the Hebbian update rule is purely correlational, whereas the TD update rule is error-driven.

While the TD-SR algorithm provides insight into the normative rationale of TCM, it does not offer guidance in the setting of parameters, since λ and γ are considered idiosyncratic properties of the agent (but see Doya, 2002). Experimentally, a number of individual differences have been correlated with these parameters, such as age (Wingfield et al., 1998) and working memory capacity (Unsworth, 2007).

²In addition, a scaling operation is typically performed to constrain the context vector to be of unit length. For simplicity, we omit this step. See Howard and Kahana (2002) for more details.

5 New experimental predictions

Previous applications of TCM to free recall data have focused on paradigms where each item only appeared once at study (but see Howard et al., 2009); as noted above, the TD and Hebbian models generate identical predictions in this situation. Other free recall studies have presented items repeatedly at study (e.g., to investigate spacing effects; Kahana and Howard, 2005), but none of these studies manipulated successor relationships in a manner that would allow us to tease apart the models. In this section, we discuss how to construct an experiment that predicts qualitatively different outcomes given TD vs. Hebbian learning.

There are two key differences between the TD and Hebbian learning rules that can be exploited to generate differential predictions. The first is that TD is *error-driven*, whereas Hebbian learning is purely associative. What this means is that TD will only update the SR if there is a discrepancy between what it predicts and what it observes. In contrast, the Hebbian rule will continue to strengthen the association between item and contextual features even if there is no prediction error. It is worth noting that, from a neurobiological point of view, Hebbian learning of this sort can lead to instabilities that have disastrous consequences for a memory system (Brown et al., 1990). Furthermore, error-driven learning has played a key role in theoretical explanations of a wide variety of animal learning phenomena, such as blocking and conditioned inhibition (e.g., Rescorla and Wagner, 1972), and signatures of state-based prediction errors have been recently observed neurally and behaviorally (Gläscher et al., 2010). The second key difference between the TD and Hebbian learning rules is that the TD rule also learns to predict discounted future states given an observed transition (via the γ term in Eq. 5). For example, in cases where a previously experienced item (call it item X) is repeated after a novel item (call it item Y), the TD learning rule will learn associations between item Y and the successors of X (i.e., items that have followed X in the past), whereas the Hebbian rule will not form these associations.

To illustrate the two predicted differences between TD and Hebbian learning, we simulated a free recall paradigm with lists containing a repeated sub-sequence (AB):

A B C D A B E F A B G H I A,

where the letters indicate the study-words (the bold is added only for explanatory emphasis). As shown in Figure 2, both the TD and Hebbian learning rules predict almost the same associative strength stored in \hat{M} from $I \rightarrow A$, which we can use as a baseline for the subsequent comparisons. As described above, repeating the AB sub-sequence forms a strong association from A to B, but the Hebbian learning rule grows without bounds with each repetition, which gives rise to an associative strength from $A \rightarrow B$ that is much larger for the Hebbian than the TD learning rule. We simulated recall³ for the list described above with both the TD and Hebbian learning rules and counted the number of $A \rightarrow B$ transitions (i.e., the number of times the model recalled A and then retrieved B on the following recall). Because recall in the model is a stochastic process, we ran the simulation 2000

³See Sederberg et al. (2008) for details of the recall mechanism in TCM. In short, the current state of context is multiplied by the learned context-to-item association matrix, giving rise to a distribution of item strengths that are used to drive item-specific accumulators. The one difference between the present simulations and the Sederberg et al. (2008) simulations is that we replaced the leaky competitive accumulator (Usher and McClelland, 2001) with the linear ballistic accumulator (LBA; Brown and Heathcote, 2008) for computational speed. The LBA samples two random numbers for each candidate item per simulated recall (one for the noise in the accumulator slope, one for noise in the accumulator offset) and uses these numbers to compute how long it will take each item to cross the retrieval threshold; the first item to cross threshold is recalled.

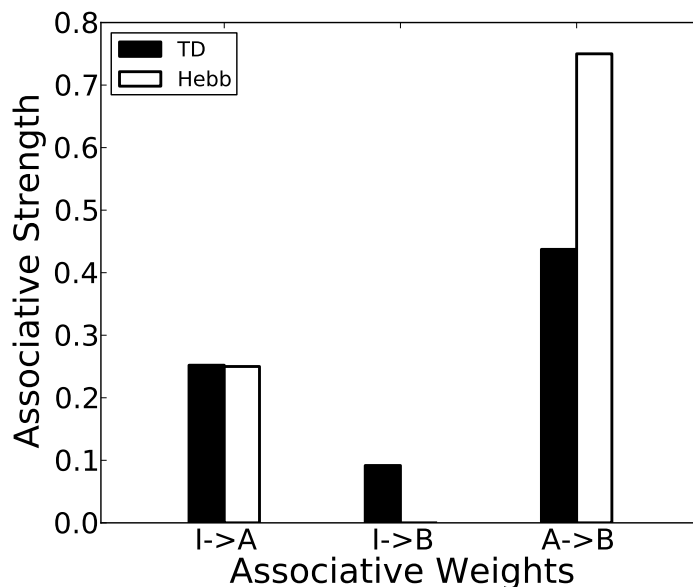


Figure 2: **New experimental predictions.** Graph of TCM simulations that dissociate the Hebbian and TD learning rules. See text for details.

times and averaged the results. The ratio of total $A \rightarrow B$ recall transitions for the Hebbian learning rule vs. the TD learning rule was 1.258, confirming that the larger association in \hat{M} is proportional to recall and translates into a higher $A \rightarrow B$ transition probability at test.

Although the $A \rightarrow B$ association is smaller for the TD learning rule, it still has an effect on other associations. When A is repeated after I , the fact that A predicts B factors into the TD learning rule such that an $I \rightarrow B$ association is learned in addition to the $I \rightarrow A$ association (see the middle column of Figure 2). In our simulations, this association manifested as a relative increase in $I \rightarrow B$ transitions during recall, with the TD learning rule producing 6.31 times more $I \rightarrow B$ transitions than the Hebbian learning rule.

6 Discussion

The main contribution of this work is a demonstration of how seemingly disparate forms of learning may rest on a common representational foundation. Starting from a generic prediction problem in Markov chains, we presented a learning algorithm (TD-SR) for solving this problem (see also Dayan, 1993; White, 1995) and then established its equivalence to TCM, a well-known model of episodic memory. In addition to providing new insight into the normative basis of TCM (which was developed as a purely descriptive mathematical model), this equivalence led to a new error-driven learning rule and several new experimental predictions that have yet to be tested.

The idea that memories are based (at least in part) on predictions about the future resonates with recent neuroscientific work suggesting a close connection between remembering the past and envisioning the future (Schacter et al., 2007; Hassabis and Maguire, 2009). For example, when subjects are asked to imagine a future situation related to a cue, many of the same brain regions are

active as when subjects recall autobiographical memories (Addis et al., 2007; Szpunar et al., 2007). Damage to this network, particularly the hippocampus, impairs both autobiographical memory and episodic future thinking (Hassabis et al., 2007). These observations are paralleled by findings from rodent physiology showing that hippocampal cells encode predictions about future spatial locations (Lisman and Redish, 2009). The SR formalizes the relationship between memory and prediction in terms of item-context associations that encode predictions about the future. Retrieving an association from memory thereby corresponds to activating a prediction. From this perspective it makes sense why the same network would be involved in both episodic memory retrieval and episodic future thinking.

A number of computational theories have attempted to explain how episodic memory and prediction arise from the same hippocampal circuitry (e.g., Levy et al., 2005; Byrne et al., 2007); of particular relevance is the work of Zilli and Hasselmo (2008), who showed how episodic memories stored in the hippocampus could facilitate long-term reward prediction for Markov chains. An interesting question for future work is how the computations posited by the TD-SR algorithm might be implemented neurally. Howard et al. (2005) have proposed a detailed mapping of TCM onto medial temporal lobe structures (including the hippocampus); given the equivalence between (generalized) TCM and TD-SR, it may be sufficient simply to replace the Hebbian learning rule used by Howard et al. (2005) to model plasticity at hippocampal synapses with the TD rule described above.

At the circuit level, the architecture of the hippocampus seems well-suited to the task of computing the prediction error required by TD-SR. In particular, the CA1 sub-field receives input from entorhinal cortex (EC) both directly (via the perforant path) and indirectly (via the dentate gyrus-CA3-CA1 pathway). If one assumes that the successor prediction is formed through recurrent dynamics in CA3, this prediction could be compared to sensory input arriving at CA1 directly from EC. This view of CA1 as the locus of prediction error computation is consistent with earlier proposals characterizing CA1 as a “comparator” whose output represents a novelty (mismatch) signal (e.g., Lisman and Otmakhova, 2001; Vinogradova, 2001; Kumaran and Maguire, 2007). One implication of our theory is that the direct EC inputs should not be purely sensory—they should partially reflect future successor predictions (the $M_{s_{n+1}j}$ term in the prediction error), which might be realized through feedback from CA3 to EC (Jones, 1993).

Beyond the hippocampus, there is evidence that other brain areas might encode item representations in terms of their successors. For example, Sakai and Miyashita (1991) recorded cells in the anterior temporal cortex that preferentially responded to one stimulus and gradually increased their firing rate during a delay interval when cued with a paired associate of the preferred stimulus. Functionally similar cells have been recorded in the lateral prefrontal cortex by Rainer et al. (1999). These findings are consistent with the use of a prospective or predictive code in a number of different brain regions, but it is still an open question whether they support the particular SR formalism we have proposed.

Once the affinities between RL and episodic memory models have been recognized, a number of variations become salient. For example, Sutton (1995) observed that the SR learns a representation of states at a single time-scale; he proposed a generalization in which learning occurs at multiple time-scales by mixing together models with different temporal horizons. Sutton’s work served as the precursor to later theories of *hierarchical* RL, in which values can be defined over temporally abstract sequences of states and actions (Sutton et al., 1999). Recent studies have suggested that such temporal abstractions operate in the brain during decision-making tasks (Botvinick et al., 2009;

Ribas-Fernandes et al., 2011). There are hints that episodic memory may also operate at a mixture of time-scales (Mozer et al., 2009). These possibilities leave fertile ground for future work at the intersection of RL and episodic memory.

Acknowledgements

Respective contributions: The link between TCM and the successor representation was first worked out by CDM and PBS, in consultation with SJG, MTT, and KAN; the paper was primarily written by SJG and PBS, with additional contributions from KAN and MTT. SJG was supported by a graduate research fellowship from the National Science Foundation. We thank Dylan Simon and Peter Dayan for helpful discussions.

References

- Addis, D., Wong, A., and Schacter, D. (2007). Remembering the past and imagining the future: Common and distinct neural substrates during event construction and elaboration. *Neuropsychologia*, 45(7):1363–1377.
- Altmann, G. and Kamide, Y. (1999). Incremental interpretation at verbs: Restricting the domain of subsequent reference. *Cognition*, 73(3):247–264.
- Anderson, J. and Bower, G. (1972). Recognition and retrieval processes in free recall. *Psychological Review*, 79(2):97–123.
- Atance, C. and O’Neill, D. (2001). Episodic future thinking. *Trends in Cognitive Sciences*, 5(12):533–539.
- Bellman, R. (1957). *Dynamic Programming*. Princeton University Press.
- Botvinick, M., Niv, Y., and Barto, A. (2009). Hierarchically organized behavior and its neural foundations: A reinforcement learning perspective. *Cognition*, 113(3):262–280.
- Brown, G., Neath, I., and Chater, N. (2007). A temporal ratio model of memory. *Psychological Review*, 114(3):539–576.
- Brown, S. D. and Heathcote, A. (2008). The simplest complete model of choice response time: Linear ballistic accumulation. *Cognitive Psychology*, 57:153–178.
- Brown, T., Kairiss, E., and Keenan, C. (1990). Hebbian synapses: biophysical mechanisms and algorithms. *Annual Review of Neuroscience*, 13(1):475–511.
- Byrne, P., Becker, S., and Burgess, N. (2007). Remembering the past and imagining the future: a neural model of spatial memory and imagery. *Psychological Review*, 114(2):340–375.
- Dayan, P. (1993). Improving generalization for temporal difference learning: The successor representation. *Neural Computation*, 5(4):613–624.
- Deese, J. and Kaufman, R. (1957). Serial effects in recall of unorganized and sequentially organized verbal material. *Journal of Experimental Psychology*, 54(3):180–187.
- Doya, K. (2002). Metalearning and neuromodulation. *Neural Networks*, 15(4-6):495–506.
- Estes, W. (1955). Statistical theory of spontaneous recovery and regression. *Psychological Review*, 62(3):145–154.
- Gläscher, J., Daw, N., Dayan, P., and O’Doherty, J. (2010). States versus Rewards: Dissociable Neural Prediction Error Signals Underlying Model-Based and Model-Free Reinforcement Learning. *Neuron*, 66(4):585–595.
- Glenberg, A. and Swanson, N. (1986). A temporal distinctiveness theory of recency and modality effects. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 12(1):3–15.

- Hassabis, D., Kumaran, D., Vann, S., and Maguire, E. (2007). Patients with hippocampal amnesia cannot imagine new experiences. *Proceedings of the National Academy of Sciences*, 104(5):1726–1731.
- Hassabis, D. and Maguire, E. (2009). The construction system of the brain. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 364(1521):1263–1271.
- Howard, M., Fotedar, M., Datey, A., and Hasselmo, M. (2005). The temporal context model in spatial navigation and relational learning: Explaining medial temporal lobe function across domains. *Psychological Review*, 112:75–116.
- Howard, M., Jing, B., Rao, V., Probyn, J., and Datey, A. (2009). Bridging the gap: Transitive associations between items presented in similar temporal contexts. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 35(2):391–407.
- Howard, M. and Kahana, M. (1999). Contextual Variability and Serial Position Effects in Free Recall. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 25(4):923–941.
- Howard, M. and Kahana, M. (2002). A distributed representation of temporal context. *Journal of Mathematical Psychology*, 46(3):269–299.
- Howard, M., Shankar, K., and Jagadisan, U. (2011). Constructing semantic representations from a gradually-changing representation of temporal context. *Topics in Cognitive Science*, 3:48–73.
- Jones, R. (1993). Entorhinal-hippocampal connections: a speculative view of their function. *Trends in Neurosciences*, 16(2):58–64.
- Kahana, M. (1996). Associate retrieval processes in free recall. *Memory & Cognition*, 24:103–109.
- Kahana, M. and Howard, M. (2005). Spacing and lag effects in free recall of pure lists. *Psychonomic Bulletin and Review*, 12:159–164.
- Kahana, M., Howard, M., and Polyn, S. (2008). Associative retrieval processes in episodic memory. In Roediger, H., editor, *Learning and Memory: A Comprehensive Reference*. Academic Press.
- Kamide, Y., Altmann, G., and Haywood, S. (2003). The time-course of prediction in incremental sentence processing: Evidence from anticipatory eye movements. *Journal of Memory and Language*, 49(1):133–156.
- Kemeny, J. and Snell, J. (1976). *Finite Markov Chains*. Springer.
- Kumaran, D. and Maguire, E. (2007). Which computational mechanisms operate in the hippocampus during novelty detection? *Hippocampus*, 17(9):735–748.
- Levy, W., Hocking, A., and Wu, X. (2005). Interpreting hippocampal function as recoding and forecasting. *Neural Networks*, 18(9):1242–1264.
- Lisman, J. and Otmakhova, N. (2001). Storage, recall, and novelty detection of sequences by the hippocampus: elaborating on the SOCRATIC model to account for normal and aberrant effects of dopamine. *Hippocampus*, 11(5):551–568.

- Lisman, J. and Redish, A. (2009). Prediction, sequences and the hippocampus. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 364(1521):1193–1201.
- Mahadevan, S. and Maggioni, M. (2007). Proto-value functions: A Laplacian framework for learning representation and control in Markov decision processes. *Journal of Machine Learning Research*, 8:2169–2231.
- Mensink, G. and Raaijmakers, J. (1988). A model for interference and forgetting. *Psychological Review*, 95(4):434–455.
- Mozer, M., Pashler, H., Cepeda, N., Lindsey, R., and Vul, E. (2009). Predicting the optimal spacing of study: A multiscale context model of memory. *Advances in Neural Information Processing Systems*, 22:1321–1329.
- Polyn, S., Norman, K., and Kahana, M. (2009). A context maintenance and retrieval model of organizational processes in free recall. *Psychological Review*, 116(1):129–156.
- Rainer, G., Rao, S., and Miller, E. (1999). Prospective coding for objects in primate prefrontal cortex. *The Journal of Neuroscience*, 19(13):5493–5505.
- Rescorla, R. A. and Wagner, A. R. (1972). A theory of of Pavlovian conditioning: variations in the effectiveness of reinforcement and nonreinforcement. In Black, A. and Prokasy, W., editors, *Classical Conditioning II: Current Research and theory*, pages 64–99. Appleton-Century-Crofts, New York, NY.
- Ribas-Fernandes, J., Solway, A., Diuk, C., McGuire, J., Barto, A., Niv, Y., Botvinick, M., et al. (2011). A neural signature of hierarchical reinforcement learning. *Neuron*, 71(2):370–379.
- Sakai, K. and Miyashita, Y. (1991). Neural organization for the long-term memory of paired associates. *Nature*, 354(6349):152–155.
- Schacter, D., Addis, D., and Buckner, R. (2007). Remembering the past to imagine the future: the prospective brain. *Nature Reviews Neuroscience*, 8(9):657–661.
- Schank, R. (1982). *Dynamic Memory*. Cambridge University Press.
- Sederberg, P., Howard, M., and Kahana, M. (2008). A context-based theory of recency and contiguity in free recall. *Psychological Review*, 115(4):893–912.
- Sederberg, P., Miller, J., Howard, M., and Kahana, M. (2010). Temporal contiguity between recalls predicts episodic memory performance. *Psychonomic Bulletin & Review*, 38:689–699.
- Shankar, K., Jagadisan, U., and Howard, M. (2009). Sequential learning using temporal context. *Journal of Mathematical Psychology*, 53(6):474–485.
- Socher, R., Gershman, S., Perotte, A., Sederberg, P., Blei, D., and Norman, K. (2009). A bayesian analysis of dynamics in free recall. In Bengio, Y., Schuurmans, D., Lafferty, J., Williams, C. K. I., and Culotta, A., editors, *Advances in Neural Information Processing Systems 22*, pages 1714–1722.
- Sutton, R. (1988). Learning to predict by the methods of temporal differences. *Machine Learning*, 3(1):9–44.

- Sutton, R. (1995). TD models: Modeling the world at a mixture of time scales. *International Conference of Machine Learning*, pages 531–539.
- Sutton, R. and Barto, A. (1998). *Reinforcement Learning: An Introduction*. MIT press.
- Sutton, R., Precup, D., and Singh, S. (1999). Between mdps and semi-mdps: A framework for temporal abstraction in reinforcement learning. *Artificial Intelligence*, 112(1):181–211.
- Szpunar, K., Watson, J., and McDermott, K. (2007). Neural substrates of envisioning the future. *Proceedings of the National Academy of Sciences*, 104(2):642–647.
- Unsworth, N. (2007). Individual differences in working memory capacity and episodic retrieval: Examining the dynamics of delayed and continuous distractor free recall. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 33(6):1020–1034.
- Usher, M. and McClelland, J. L. (2001). The time course of perceptual choice: The leaky, competing accumulator model. *Psychological Review*, 108(3):550–592.
- Vinogradova, O. (2001). Hippocampus as comparator: role of the two input and two output systems of the hippocampus in selection and registration of information. *Hippocampus*, 11(5):578–598.
- White, L. (1995). *Temporal Difference Learning: Eligibility Traces and the Successor Representation for Actions*. Unpublished master’s thesis, Department of Computer Science, University of Toronto.
- Wingfield, A., Lindfield, K., and Kahana, M. (1998). Adult age differences in the temporal characteristics of category free recall. *Psychology and Aging*, 13(2):256–266.
- Zilli, E. and Hasselmo, M. (2008). Modeling the role of working memory and episodic memory in behavioral tasks. *Hippocampus*, 18(2):193–209.