Human Behavior is Best Predicted by a Novel Successor Representation Learning Rule

Ari E. Kahn (arik@princeton.edu)
Princeton Neuroscience Institute, Princeton University, Princeton, NJ, USA

Dani S. Bassett (dsb@seas.upenn.edu)
Department of Bioengineering, University of Pennsylvania, Philadelphia, PA, USA

Nathaniel D. Daw (ndaw@princeton.edu)
Princeton Neuroscience Institute and Department of Psychology, Princeton University, Princeton, NJ, USA

Abstract

Human decision making depends on learning and using models that capture the statistical structure of the world, including the long-run expected outcomes of our actions. One prominent approach to abstracting such long-run outcomes is the successor representation (SR), which represents a mapping between current and future states, and has been implicated in both behavioral and neural data. Although much behavioral and neural evidence suggest that people and animals use such a representation, it remains unknown how they learn it. Bootstrapping methods (SR-TD(0)) have been ubiquitously proposed, but bootstrapping a vector-valued function in large state spaces appears biologically implausible. Here we propose an alternative learning rule, termed SR-Trace, which approximates SR-TD(1) using a simpler scalar update process. We examined the behavior of both on a probabilistic graph learning task, and found that trial-by-trial response times were better predicted by the more plausibly realizable SR-Trace model, suggesting that humans may rely on this biologically plausible SR learning rule in graph learning tasks.

Introduction

A range of neural and behavioral results (Schapiro, Turk-Browne, Norman, & Botvinick, 2016; Garvert, Dolan, & Behrens, 2017; Stachenfeld, Botvinick, & Gershman, 2017; Momennejad et al., 2017; Russek, Momennejad, Botvinick, Gershman, & Daw, 2021; Ekman, Kusch, & de Lange, 2023) suggest that the brain uses temporally abstract representations like the successor representation (SR) (Dayan, 1993), which predict future events over multiple steps. However, it remains unknown how these long-range predictions are learned. One setting in which it may be promising to investigate such learning rules is stimulus-by-stimulus reaction times (RTs) in graph learning tasks, which reflect predictions relatively directly. In particular, previous work has shown that in such tasks, people exhibit systematic RT biases, among which is a sensitivity to modular structure, exhibited via a 'cross-cluster surprisal effect', where RTs are heightened when transitioning between densely-interconnected clusters of nodes (Karuza, Kahn, Thompson-Schill, & Bassett, 2017; Kahn, Karuza, Vettel, & Bassett, 2018). These behavioral biases are well captured by a maximum-entropy prediction model that balances representational complexity with accuracy (Lynn, Kahn, Nyema, & Bassett, 2020) and which is mathematically equivalent to the SR. The SR's predictions are higher within than between clusters, because multi-step transitions tend to respect cluster boundaries. These results suggest that peoples' learned representations of graph structure reflect an SR.

While we can observe correlates of the expected SR for block-averaged behavior, there are a number of distinct hypothesized mechanisms for how the SR might be learned in a trial-by-trial fashion (Russek, Momennejad, Botvinick, Gershman, & Daw, 2017). The SR can be directly computed from the one-step transition matrix T (which can itself be learned straightforwardly by a Hebbian update) via its definition $M = (I - \gamma T)^{-1}$, but this requires a matrix inversion upon each update to T. It has instead been widely assumed that the SR is directly learned via a bootstrapped temporal-difference rule, **SR-TD(0)**, where future occupancy is directly learned via temporal-difference learning (Dayan, 1993; Gershman, Moore, Todd, Norman, & Sederberg, 2012; Gardner, Schoenbaum, & Gershman, 2018). However, such a rule requires a vector valued update signal (a separate prediction error for each target state, i.e., each row of M), which appears anatomically implausible (Akam & Walton, 2021) (Figure 1b, left). Here we propose an alternative, **SR-Trace**, which can be learned using a Hebbian update with eligibility traces, without a vector-valued error signal (Figure 1b, right).

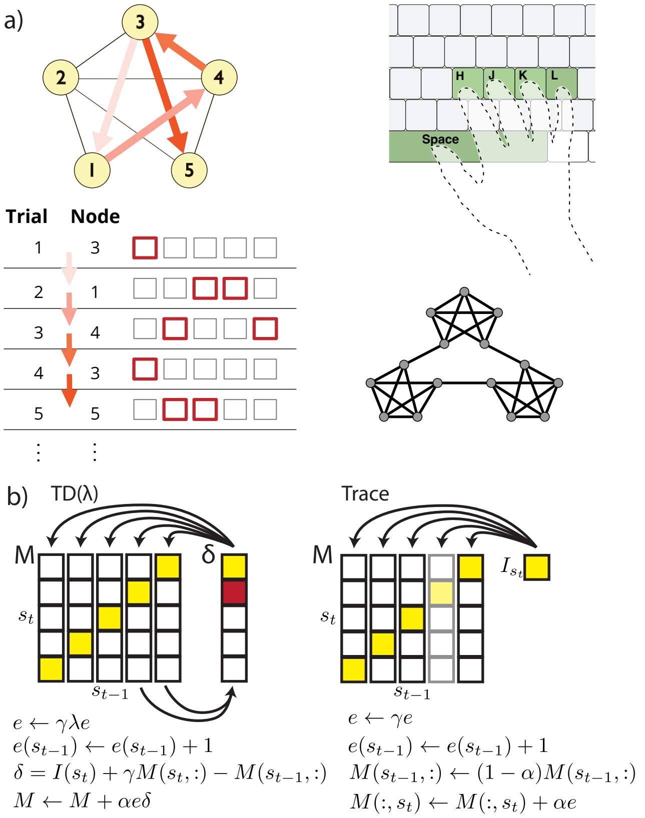
a)
3
2
4
1
5
H
J
K
L
Space
Trial
Node
1
3
2
1
3
4
4
3
5
5
⋮
⋮
b)
TD(λ)
Trace
M
δ
s_t
s_{t-1}
M
I_{s_t}
s_t
s_{t-1}
$e \leftarrow \gamma\lambda e$
$e(s_{t-1}) \leftarrow e(s_{t-1}) + 1$
$\delta = I(s_t) + \gamma M(s_t,:) - M(s_{t-1},:)$
$M \leftarrow M + \alpha e \delta$
$e \leftarrow \gamma e$
$e(s_{t-1}) \leftarrow e(s_{t-1}) + 1$
$M(s_{t-1},:) \leftarrow (1-\alpha)M(s_{t-1},:)$
$M(:,s_t) \leftarrow M(:,s_t) + \alpha e$

Figure 1: **SR Prediction on a Graph Learning Task. a)** Task Design: Participants responded to a sequence of 1500 stimuli derived from a random walk on a modular graph, where each trial required pressing a one- or two-button combination of keys to identify the presented stimulus as quickly as possible. **b)** SR Updating: SR-TD requires transmitting a $1 \times n$ error vector to update future predictions of all states whose trace is nonzero. SR-Trace instead only requires local updates, while having the same fixed point.

SR-Trace

SR-Trace is a Monte Carlo estimator for the SR, derived from SR-TD(λ) for $\lambda = 1$. Recall that TD(λ) interpolates between fully bootstrapped TD ($\lambda = 0$) (where a "bootstrapped" future prediction stands in for longer run outcomes at each update) and a Monte Carlo estimator $\lambda = 1$ (where long-run outcomes are incorporated directly by gating updates with a long-lasting "eligibility trace" e at each state). When $\lambda = 1$, the bootstrap terms have only transient effect as the bootstrap from each successive update replaces the previous one at each state (except when that state is encountered and its eligibility increases). Thus by replacing the bootstrap term in the prediction error with 0 and simplifying, we obtain a variant rule in which the SR is estimated via Hebbian learning with decay and eligibility traces (Figure 1b, right), without requiring a vector-valued error signal to route a $1 \times n$ vector of state-specific bootstraps. Conceptually, this approximation can be thought of as a truncation of the target sample at each step by deferring the rest of the long-run update, which will in turn gradually be incorporated by outcomes encountered at later steps, so that the fixed point (the SR) is the same in the limit.

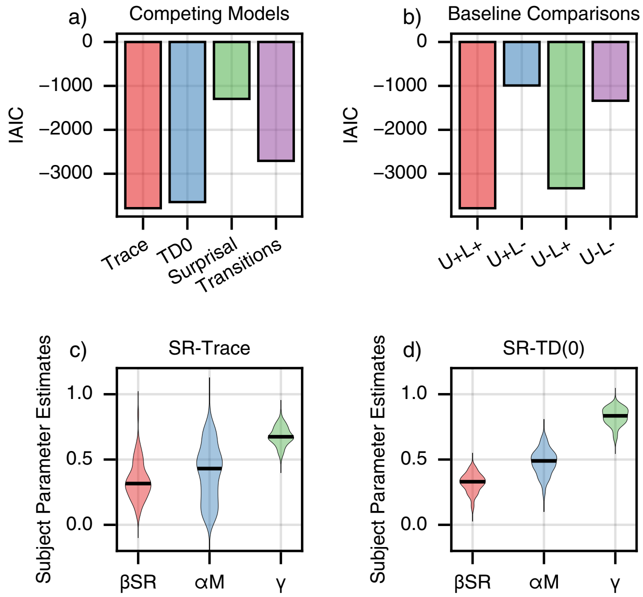

Figure 2: **Model fit and parameter recovery. a)** Model comparison between SR-Trace and alternatives. Bars indicate the integrated AIC, where more negative values indicate a better fit to observed data. **b)** Model comparison between full SR-Trace and simplified versions. U and L, respectively, indicate whether the model was initialized with an uninformed starting point (vs. the expected fix point SR) and whether learning was included. **c)** Distribution of recovered subject-level parameters for SR-trace. $\beta_{anticipation}$ is the coefficient of anticipation, indicating the effect of anticipation on reaction time. αM is the learning rate for the SR M matrix. γ is the discount factor. **d)** Distribution of recovered subject-level parameters for SR-TD(0).

Model

We modeled the RT data from Kahn et al. (2018), where participants completed 1500 trials of a serial reaction time-like task, responding to a cue shown on the screen as rapidly as possible with a one- or two-button key combination. Unbeknownst to participants, the sequence of cues followed a traversal through a graph composed of 15 nodes in 3 clusters. RTs were assumed to be log-normally distributed, including a possible shift in baseline RT. The model took the form: $\log(rt_t - s) \sim N(\mu_t, \sigma^2)$, where μ_t is a function of a per-subject baseline, trial, target (finger combination), and an anticipatory effect, which we hypothesize is best explained by the SR. Per-subject parameters were fit hierarchically over the group via expectation maximization, and models were compared via integrated AIC. For SR models, the anticipatory effect was the entry of the M matrix corresponding to the observed transition, denoted M_{s_{t-1},s_t}.

Results

First, we compared SR-Trace with a standard SR-TD(0) bootstrapped learning rule. In both cases, in order to capture unbiased baseline expectancy, the M matrix was initialized to $(I - \gamma T)^{-1}$ for a uniform transition matrix between all 15 stimuli. After each transition, the SR M matrix was updated using either TD(0) or the simpler trace update, and the resulting M matrix was used to estimate the anticipatory effect for each subject. We find that SR-Trace provides a better fit than SR-TD(0) to trial-by-trial behavioral data. Additionally, as a baseline, we compared SR-Trace and SR-TD(0) to two methods that do not rely on multi-step predictive representations—first, a simple model which modeled expectancy only by a binary indicator for between-cluster transitions, and second, a model that directly learned one-step transition probabilities via a delta rule. Consistent with prior work, the multi-step predictive models provided a better fit than either of the alternative models.

Next, we wanted to verify or disprove a number of alternative explanations for the success of the TD-Trace method. One possibility is that the effect is an artifact of initialization, e.g. that the models differ only in initial acquisition of M but not in steady-state adjustments around the fixed point. To address this possibility, we initialized M to the steady-state expected from the full sequence of 1500 stimuli (i.e., $(I - \gamma T)^{-1}$), and compared performance with and without additional learning. We observe that even when initialized to a 'converged' M matrix, trial-by-trial learning is highly predictive of RTs. In addition to the model fit capturing trial-by-trial fluctuations in anticipation, we indeed find that initialization to an uninformed baseline provides a better fit than starting with the converged matrix, showing that our model provides an optimal fit both when initialized from a subject's true belief state (uniform transitions) and allows for trial-by-trial learning. Together, these results establish that a biologically plausible implementation of SR learning may underlie predictive representations and give rise to RTs observed in graph learning experiments.

References

Akam, T., & Walton, M. E. (2021). What is dopamine doing in model-based reinforcement learning? *Current Opinion in Behavioral Sciences*, *38*, 74–82.

Dayan, P. (1993, July). Improving Generalization for Temporal Difference Learning: The Successor Representation. *Neural Computation*, *5*(4), 613–624. doi: 10.1162/neco.1993.5.4.613

Ekman, M., Kusch, S., & de Lange, F. P. (2023, February). Successor-like representation guides the prediction of future events in human visual cortex and hippocampus. *eLife*, *12*, e78904. doi: 10.7554/eLife.78904

Gardner, M. P., Schoenbaum, G., & Gershman, S. J. (2018). Rethinking dopamine as generalized prediction error. *Proceedings of the Royal Society B*, *285*(1891), 20181645.

Garvert, M. M., Dolan, R. J., & Behrens, T. E. (2017, April). A map of abstract relational knowledge in the human hippocampal–entorhinal cortex. *eLife*, *6*, e17086. doi: 10.7554/eLife.17086

Gershman, S. J., Moore, C. D., Todd, M. T., Norman, K. A., & Sederberg, P. B. (2012). The successor representation and temporal context. *Neural Computation*, *24*(6), 1553–1568.

Kahn, A. E., Karuza, E. A., Vettel, J. M., & Bassett, D. S. (2018, December). Network constraints on learnability of probabilistic motor sequences. *Nature Human Behaviour*, *2*(12), 936–947. doi: 10.1038/s41562-018-0463-8

Karuza, E. A., Kahn, A. E., Thompson-Schill, S. L., & Bassett, D. S. (2017, October). Process reveals structure: How a network is traversed mediates expectations about its architecture. *Scientific Reports*, *7*(1), 12733. doi: 10.1038/s41598-017-12876-5

Lynn, C. W., Kahn, A. E., Nyema, N., & Bassett, D. S. (2020, May). Abstract representations of events arise from mental errors in learning and memory. *Nature Communications*, *11*(1), 2313. doi: 10.1038/s41467-020-15146-7

Momennejad, I., Russek, E. M., Cheong, J. H., Botvinick, M. M., Daw, N. D., & Gershman, S. J. (2017, September). The successor representation in human reinforcement learning. *Nature Human Behaviour*, *1*(9), 680–692. doi: 10.1038/s41562-017-0180-8

Russek, E. M., Momennejad, I., Botvinick, M. M., Gershman, S. J., & Daw, N. D. (2017, September). Predictive representations can link model-based reinforcement learning to model-free mechanisms. *PLOS Computational Biology*, *13*(9), e1005768. doi: 10.1371/journal.pcbi.1005768

Russek, E. M., Momennejad, I., Botvinick, M. M., Gershman, S. J., & Daw, N. D. (2021, August). *Neural evidence for the successor representation in choice evaluation.* bioRxiv. doi: 10.1101/2021.08.29.458114

Schapiro, A. C., Turk-Browne, N. B., Norman, K. A., & Botvinick, M. M. (2016). Statistical learning of temporal community structure in the hippocampus. *Hippocampus*, *26*(1), 3–8. doi: 10.1002/hipo.22523

Stachenfeld, K. L., Botvinick, M. M., & Gershman, S. J. (2017, November). The hippocampus as a predictive map. *Nature Neuroscience*, *20*(11), 1643–1653. doi: 10.1038/nn.4650