AMERICAN PSYCHOLOGICAL ASSOCIATION

Psychological Review

ISSN: 0033-295X

2024, Vol. 131, No. 2, 578–597
https://doi.org/10.1037/rev0000414

THEORETICAL NOTE

A Probabilistic Successor Representation for Context-Dependent Learning

Jesse P. Geerts[1, 2], Samuel J. Gershman[3, 4], Neil Burgess[1, 2], and Kimberly L. Stachenfeld[5]
[1] Institute of Cognitive Neuroscience, University College London
[2] Sainsbury Wellcome Centre for Neural Circuits and Behaviour, University College London
[3] Department of Psychology, Harvard University
[4] Center for Brain Science, Harvard University
[5] DeepMind, London, United Kingdom

Two of the main impediments to learning complex tasks are that relationships between different stimuli, including rewards, can be uncertain and context-dependent. Reinforcement learning (RL) provides a framework for learning, by predicting total future reward directly (model-free RL), or via predictions of future states (model-based RL). Within this framework, "successor representation" (SR) predicts total future occupancy of all states. A recent theoretical proposal suggests that the hippocampus encodes the SR in order to facilitate prediction of future reward. However, this proposal does not take into account how learning should adapt under uncertainty and switches of context. Here, we introduce a theory of learning SRs using prediction errors which includes optimally balancing uncertainty in new observations versus existing knowledge. We then generalize that approach to a multicontext setting, allowing the model to learn and maintain multiple task-specific SRs and infer which one to use at any moment based on the accuracy of its predictions. Thus, the context used for predictions can be determined by both the contents of the states themselves and the distribution of transitions between them. This probabilistic SR model captures animal behavior in tasks which require contextual memory and generalization, and unifies previous SR theory with hippocampal-dependent contextual decision-making.

Keywords: reinforcement learning, successor representation, uncertainty, context

Learning to predict future rewards is critical to the survival of humans and other animals. It is generally assumed that animals achieve this by learning associations between stimuli and future rewards. However, this is a hard learning problem for a number of reasons. First, there is the challenge of flexibly updating learned patterns in a changing world: stimulus–reward associations might be highly context-dependent, such that a stimulus might predict different stimuli or a rewards in different settings, and changing goals can drastically alter where and when reward is anticipated. Furthermore, it is difficult to deal with uncertainty: Since the full state of the world cannot be observed, humans and animals must reason under uncertainty about how rewarding outcomes are, how outcomes relate to each other, and how context is mitigating these associations.

There is plentiful evidence that animals can deal flexibly with change, and understanding how context modulates behavior is a long-standing topic of interest in psychology and neuroscience (Bouton, 2004; Bouton & Bolles, 1979; Gershman, 2017a; Gershman et al., 2010; Heald et al., 2021). In this literature, researchers have investigated contextual effects by changing the animal's environment or by changing the task demands in the same environment. Interestingly, context can also refer to the animal's internal belief about the environment. In fear conditioning studies, for example, animals are often observed to freeze more in the context in which they received the shock. This does not have to be the same environment as long as the animal *infers* it to be the same (Chang & Liang, 2017): For instance, when uncertainty is high, the animal might falsely infer

This article was published Online First May 11, 2023.
Jesse P. Geerts https://orcid.org/0000-0003-2960-5727
The authors thank Athena Akrami, Peter Dayan, and Steven Hansen for their helpful comments. Jesse P. Geerts received a PhD studentship from the Gatsby Charitable Foundation and the Wellcome Trust. Samuel J. Gershman received funding from Air Force Office of Scientific Research (Grant FA9550-20-1-0413). Neil Burgess was supported by a Wellcome Principal Research Fellowship, European Research Council (ERC) Advanced Grant NEUROMEM, Wellcome Collaborative Award "Organising knowledge for flexible behavior in the prefrontal hippocampal circuitry." Jesse P. Geerts and Neil Burgess received support from the ERC advanced Grant NEUROMEM. This research was funded in whole, or in part, by the Wellcome Trust (222457/Z/21/Z and 214314/Z/18/Z to Neil Burgess). For the purpose of Open Access, the author has applied a CC BY public copyright licence to any author accepted manuscript version arising from this submission.

Analysis code for this study will be made available upon publication. This study was not preregistered. An earlier version of this article was posted on bioRxiv: https://doi.org/10.1101/2022.06.03.494671.

Correspondence concerning this article should be addressed to Jesse P. Geerts, Sainsbury Wellcome Centre for Neural Circuits and Behaviour, University College London, 25 Howland Street, London W1T 4JG, United Kingdom. Email: jesse.geerts.14@ucl.ac.uk

that it is in a familiar environment. This inferential aspect of context, which is dependent on the hippocampus, can be captured by models in which context is treated as a latent variable which the animal must infer (Gershman et al., 2010, 2014; Sanders et al., 2020). According to these models, animals perform clustering to organize their experience into discrete categories.

Recently, researchers have also tried understanding flexible behavior through the lens of reinforcement learning (RL). In this framework, flexibility is thought to derive from an internal model of the world, used to predict the future and plan actions accordingly. Such "model-based planning" is flexible in that it allows new information about which states are related and which are rewarding to be incorporated rapidly into the model. This approach can be contrasted with model-free RL (Sutton & Barto, 1998), which supports rapid decision-making but is inflexible, as it forgoes learning explicit knowledge about the environment's dynamics in favor of learning and caching simple associations between stimuli and cumulative future reward. The dichotomy between model-free and model-based decision-making has inspired a rich body of work in psychology and neuroscience suggesting that multiple learning systems operate in parallel, competing for control of behavior (Daw et al., 2005; Geerts et al., 2020) and that there are alternative approaches that combine aspects of model-based and model-free learning (Gershman, 2018; Russek et al., 2017). One such alternative approach to both model-based and model-free RL takes the form of a representation of long-run state expectancies. This successor representation (SR; Dayan, 1993) combines the aspects of model-based and model-free RL. Unlike model-based methods, reward prediction with the SR can be done without explicit forward simulation and is therefore relatively computationally efficient. Unlike model-free methods, the SR captures information about the environment structure that is not directly related to rewards, which means that value can be flexibly recomputed under different reward functions, as long as the associated behavior policy does not change. The effect on prediction of the behavior or task performed by the animal (its "policy") is also captured by the SR. Signatures of decision-making strategies based on the SR have been used to explain the aspects of human and animal behavior (De Cothi et al., 2022; Momennejad, Russek, et al., 2017) and neural activity (Gardner et al., 2018; Stachenfeld et al., 2017), particularly how predictions are modified under changing goals and structure and how generalization is affected by prior experience. In this work, we consider contributions of both predictive representations and context to flexible behavior and consider the problem of how to perform optimal inference over these quantities.

In order to incorporate uncertainty into the inference process, we adopt the Bayesian approach to learning which suggests that animals not only learn to predict rewards or stimuli but also estimate their uncertainty about these predictions. For example, a wide range of animal learning phenomena can be explained by probabilistic generalizations of simple model-free learning algorithms, such as the Kalman filter (Dayan & Kakade, 2001; Dayan & Yu, 2003; Gershman, 2015). These theories posit that, rather than learning a single-point estimate for the weights used to approximate future reward, animals track a distribution over these weights. Such distributions contain additional information about the variance of the value-function weights, which reflect uncertainty, as well as information about the covariance between interdependent parameters. These uncertainty and interdependency terms can explain why animals often learn more slowly in situations of low uncertainty (i.e., latent inhibition) and why they can learn about stimuli that are not currently present (i.e., backward blocking; Dayan & Kakade, 2001; Gershman, 2015). Because SR learning and value-function learning are mathematically similar, this approach can be easily adapted to the SR setting, permitting optimal about uncertainty not just about expected reward but about expected predictive structure as well.

To model the multiple context setting, we build on the literature that casts context as a latent state that the animal needs to infer from observations (Gershman et al., 2010; Sanders et al., 2020). Our model takes the form of a switching Kalman filter (Gershman et al., 2014), which can switch between and update multiple predictive maps which describe the expected future states in a given context. Since the number of contexts is a priori unknown, we use a nonparametric (infinite capacity) model in which the number of contexts can grow with the observations. This allows the model to learn and maintain multiple task-specific SR maps and infer which one to use at any moment based on its sensory observations. Switches between contexts are driven by prediction errors associated with the SR as well as uncertainty. In this definition, context thus reflects both the contents of sensory states and the transitions between them. This means that a new set of transition rules or even a new policy in the same environment can lead to a change of the context by which predictions are being made.

Our goal in this work is to incorporate accurate reasoning under uncertainty into a flexible learning framework that includes context-dependent learning of the SR and to demonstrate that this captures several apparently contradictory behaviors under a common model. To achieve this, we first generalize the most common algorithm for learning the SR online to take into account uncertainty in the predictions. Second, we use these uncertainty estimates to give our model the ability to switch between different, context-specific predictive maps based on the prediction errors associated with previously stored contexts. This probabilistic SR model allows us to explain several apparently contradictory animal behaviors in tasks which require contextual memory and generalization, such as rapid reevaluation, the conditions under which context preexposure facilitates or inhibits learning and how it mediates contextual generalization, among other phenomena.

Model Description

This article addresses the problem of how to deal with uncertainty when learning predictive maps and then uses these maps to model context-dependent learning. The predictive map we consider takes the form of an SR (Dayan, 1993), a representation of states in terms of the expected discounted future occupancy of each state, from the current state. The SR has been proposed to explain the aspects of neural data in the rodent and human hippocampus (Brunec & Momennejad, 2022; de Cothi & Barry, 2020; Gershman, 2018; Momennejad, Otto, et al., 2017; Russek et al., 2017, 2021; Stachenfeld et al., 2017) and the aspects of human and rodent behavior (De Cothi et al., 2022; Momennejad, Russek, et al., 2017; Russek et al., 2017). Our first contribution, which we initially described in Geerts et al. (2019), is to introduce a probabilistic SR, in which the agent's belief about the parameters of the SR is expressed in terms of a distribution over possible SRs. This enables efficient learning by making use of the second-order statistics of predictions about future states or features and can be used to understand a range of animal learning phenomena.

Our second novel contribution is to extend the probabilistic SR to a probabilistic *hierarchical* SR in which the agent can switch between multiple SR maps when the environment, task, or context changes. This allows us to represent uncertainty within a context as well as over contexts. In this section, we describe the pieces of the model in sequence. First, we describe how to handle uncertainty when learning the SR in a single environment using Kalman temporal differences (KTD; Geist & Pietquin, 2010; Gershman, 2015) applied to the SR. Next, we describe how to generalize this for multiple environments and context-dependent SR maps by using a nonparametric switching linear dynamical system (Fox et al., 2011; Gershman et al., 2014; Murphy, 1998) that infers new maps when observations change drastically over time. Simulations using these models are presented in Kalman SR Simulations and Switching Context Model Simulations sections, respectively.

Background

We define an RL environment to be a Markov decision process consisting of *states* s the agent can occupy, *transition probabilities* $T_\pi(s'|s)$ of moving from state s to state s' given the agent's policy $\pi(a|s)$ over actions a, and the reward available at each state, for which $R(s)$ denotes the expectation. An RL agent is tasked with finding a policy that maximizes its expected discounted total future reward, or *value*:

$$V(S) = \mathbb{E}_\pi\left[\sum_{t=0}^{\infty}\gamma^t R(s_t)|s_0 = s\right], \tag{1}$$

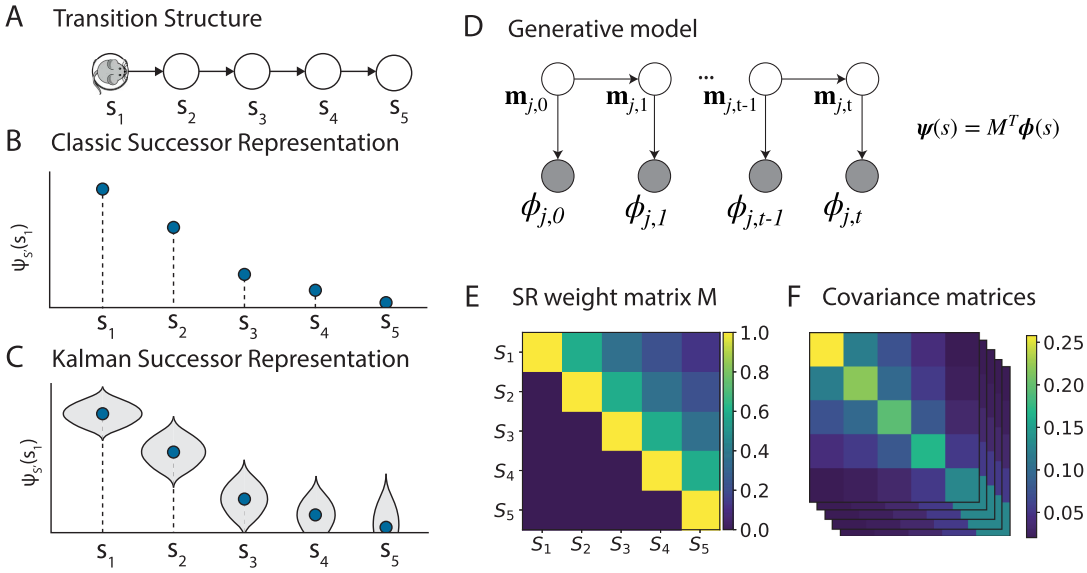
where t indexes time step and γ, where $0 \leq \gamma < 1$, is a discount factor that downweights distal rewards. In classical model-free learning algorithms called temporal difference (TD) learning (Sutton & Barto, 1998), V is learned directly through trial and error: each time a new state is encountered, V is updated proportionally to the difference between the expected and observed reward, the TD reward prediction error. However, such algorithms suffer from a lack of flexibility: When the reward function changes, model-free learners are slow to relearn the new value function. Dayan (1993) proposed one solution to this problem, made possible by the fact that V decomposes into a dot product of the direct rewards R and a predictive representation ψ:

$$V(s) = \sum_{s'}\psi_{s'}(s)R(s'), \tag{2}$$

where $\psi(s)$ is a vector with entries $\psi_{s'}(s)$ containing the expected discounted future occupancy of state s' along trajectories started in state s (see Figure 1A and B, for a simple example):

$$\psi_{s'}(s) = \mathbb{E}\left[\sum_{t=0}^{\infty}\gamma^t \mathbb{I}(s_t = s')|s_0 = s\right]. \tag{3}$$

Figure 1
Model Overview



Note. (A) Transition structure of a sequence of five states. (B) Successor representation of state s_1, corresponding to the expected discounted future occupancy given starting state s_1 (blue dots). (C) In the Kalman SR model, a distribution over feature predictions is estimated: In addition to the mean (blue dots), the variance of each successor feature is estimated (gray shaded region shows distribution). This reflects the agent's uncertainty about the values of the SR, which arises here because of forgetting. Note that the distribution includes estimates below zero, outside the limits of this figure. (D) The Kalman SR generative model's graphical structure. ϕ_j, t denotes the j^{th} feature of state s_t, $\mathbf{m}_j, t$ denotes the j^{th} column of the (latent) SR weight matrix shown in (E), at time t. $\psi(s)$ denotes the vector successor representation at state s. (E) The SR weight matrix corresponding to the transition structure in (A). The relation between the weight matrix M and the current representation ψ is given in the inset equation in (D). (F) In addition to a mean estimate of M, Kalman SR represents the uncertainty over SR weights with a set of covariance matrices corresponding to the columns of M. SR = successor representation. See the online article for the color version of this figure.

Factorizing value into an SR term and a reward term permits greater flexibility because if one term changes, it can be relearned, while the other remains intact (Barreto et al., 2016; Dayan, 1993; Gershman, 2018; Russek et al., 2017; Tomov et al., 2021). Since long-term expectations about state occupancy can be slow to estimate, this lends particular robustness when reward is changing and transition dynamics are not.

The SR can be generalized to continuous states by representing states using a set of feature functions $\phi_j(s)$. In this case, the SR is referred to as successor features (SFs; Barreto et al., 2016) and encodes the expected feature values:

$$\psi_j(s) = \mathbb{E}_\pi\left[\sum_{t=0}^{\infty} \gamma^t \phi_j(s_t) | s_0 = s\right]. \tag{4}$$

Specifically, $\psi_j(s)$ denotes the expected discounted future occurrence strength of feature j from state s. In this linear function approximation case, the reward is given by the dot product:

$$R(s) = \sum_j \phi_j(s) w_j, \tag{5}$$

where $\phi(s)$ are the state features and $\mathbf{w}$ are weights parameterizing the reward function. The decomposition of value (Equation 2) is then rewritten as:

$$V(s) = \sum_j \psi_j(s) w_j. \tag{6}$$

In the special case where the state space is finite and ϕ is a tabular representation of the state (i.e., Figure 1A, where states are discrete and represented as one-hot vectors), Equations 4 and 6 reduce to Equations 2 and 3. The contents of the feature vector can be arbitrary and will in this article depend on the particular task being modeled. We model this feature-based SR as $\hat{\psi}(s) = M^T\phi(s)$, where M is a weight matrix in which each entry M_{ij} indicates the extent to which feature i predicts feature j (Figure 1E). We thus assume for now that the agent has access to a state representation ϕ such that a linear mapping exists from features of each state to the discounted future occurrence of those features, given the start state. Seen as a single layer of a biological neural network, each column of M comprises a vector of input weights of one SR-encoding neuron ψ_j, and the vector $\psi(s)$ gives the population activity of SR-encoding neurons. To avoid cluttered notation, we will denote a column by $\mathbf{m}_j = M_{:,j}$, so that

$$\hat{\psi}_j(s) = \mathbf{m}_j^T \phi(s). \tag{7}$$

The factorized representation of the value function in Equation 6 means that two quantities have to be learned: the reward weights $\mathbf{w}$ and the SR ψ (note that, although we do not model this here, ϕ can be learned too; see, e.g., Hansen et al., 2019). We track the reward weights with a Kalman filter (Gershman, 2015). The SR ψ can be learned with TD learning, in which the SR is updated according to a TD state prediction error reflecting the difference in estimates of $\psi(s)$ and estimates of $\phi(s) + \gamma\mathbb{E}_{s'}[\psi(s')]$ (Dayan, 1993; Gardner et al., 2018). TD learning of the value function (using reward rather than state prediction errors) is a popular model of learning in the striatum (Schultz et al., 1997), and TD algorithms can be implemented in biologically realistic spiking networks (Bono et al., 2021; Brea et al., 2016; Frémaux et al., 2013).

Probabilistic SFs

In its original formulation, the SR computes a point estimate of ψ_j from the values of the SR weight matrix M. Our first step is to replace this with a probabilistic description of the SR. This probabilistic SR explicitly represents uncertainty. We model each column $\mathbf{m}_j$ of the SR weight matrix M as a set of random variables. Under this interpretation, the animal implicitly assumes there is a true, hidden set of SR parameters $\mathbf{m}_j$, which predict each new noisy observation via a generative model. The animal's goal is to invert this generative model in order to infer a distribution over the SR weights from observations. More precisely, from a sequence of observations $\phi_{1:t}$, the agent can infer information about the hidden SR weights using Bayes' rule:

$$p(m_{j,t}|\phi_{1:t}) \propto p(\phi_{1:t}|m_{j,t})p(m_{j,t}). \tag{8}$$

This idea, in the form of a Kalman filter (Kalman, 1960), has previously been applied to learning value functions (Geist & Pietquin, 2010; Gershman, 2015) and readily applies to the SR. In our SR model, each column $\mathbf{m}_j$ of M is modeled as a vector-valued random variable, with dimensionality N_ϕ.

Generative Model

Performing inference requires specifying a probabilistic generative model relating the hidden SR weights to the animal's observations (Figure 1D). It consists of a *prior* on each column of the SR matrix $\mathbf{m}_{j,0}$, an *evolution* equation describing how these hidden SR vectors evolve over time, and an *observation* equation describing how the hidden SR relates to observations. The observation equation follows directly from the Bellman equation, with additive Gaussian observation noise $\nu \sim \mathcal{N}(0, \sigma_\phi^2)$:

$$\begin{aligned}\phi_j(s_t) &= \psi_j(s_t) - \gamma\psi_j(s_{t+1}) + \nu \\ &= \mathbf{m}_j^T\phi(s_t) - \gamma\mathbf{m}_j^T\phi(s_{t+1}) + \nu \\ &= \mathbf{m}_j^T\mathbf{h}_t + \nu,\end{aligned} \tag{9}$$

where we have defined $\mathbf{h}_t = \phi(s_t) - \gamma\phi(s_{t+1})$ to be the discounted temporal difference in feature observations. We assume, in other words, that each successor feature $\psi_j(s_t)$ is a noisy linear function of the current features (see Appendix B, for additional analysis). For the evolution equation, our generative model follows a Gaussian random walk allowing the weights to change incrementally over time. We also assume a Gaussian prior on the weights. Together, these form the following probabilistic generative model (shown in Figure 1D):

$$\mathbf{m}_{j,0} \sim \mathcal{N}(\mu_0, \Sigma_0), \tag{10}$$

$$\mathbf{m}_{j,t} \sim \mathcal{N}(\mathbf{m}_{j,t-1}, Q), \tag{11}$$

$$\phi_{j,t} \sim \mathcal{N}(\mathbf{m}_{j,t}^T\mathbf{h}_t, \sigma_\phi^2), \tag{12}$$

where μ_0 is the prior mean, $\Sigma_0 = \sigma_0^2 I$ is the prior covariance matrix with prior variance σ_0^2, $Q = \sigma_\nu^2 I$ is the (diagonal) transition noise covariance matrix with transition noise variance σ_ν^2, and σ_ϕ^2 is the observation noise variance. Parameter values used in the simulations are given in Table 1.

Table 1
Parameter Settings Used in the Simulations, Except Where Indicated Otherwise

Symbol	Value
α	1.0
β	1.5
γ	0.9
σ_ϕ^2	1.0
σ_ν^2	0.001
σ_0^2	5

Inference

Our goal is to estimate the parameters $\mathbf{m}_j$ such that they satisfy $\psi_j = \mathbf{m}_j^T \phi$, for each successor feature j given the observation ϕ. Since the generative model described above is a linear-Gaussian dynamical system (LDS), we can perform exact inference on these SR weights by combining the Kalman filter equations with TD learning. Estimating a distribution over SR weights involves adjusting the mean estimate $\mathbf{m}_j$ using a temporal difference learning rule, but now taking into account the covariances of the weights, matrix Σ, via the Kalman gain $\boldsymbol{\kappa}$, an adaptive, feature-specific learning rate (Figure 1E, F). This allows for a closed-form update of a *posterior distribution* over the weights (Figure 1C):

$$\mathbf{m}_{j,t+1} = \mathbf{m}_{j,t} + \boldsymbol{\kappa}_t \delta_{j,t}, \tag{13}$$

$$\Sigma_{t+1} = \Sigma_t + Q - \lambda_t \boldsymbol{\kappa}_t \boldsymbol{\kappa}_t^T, \tag{14}$$

where $\delta_{j,t} = \phi_j(s_t) - \hat{\phi}_j(s_t) = \phi_j(s_t) + \gamma \hat{\psi}_j(s_{t+1}) - \hat{\psi}_j(s_t)$ is the successor prediction error for feature j, $\lambda_t = \mathbf{h}_t^T(\Sigma_t + Q)\mathbf{h}_t + \sigma_\phi^2$ is the residual variance, and κ_t is the Kalman gain given by:

$$\boldsymbol{\kappa}_t = \frac{(\Sigma_t + Q)\mathbf{h}_t}{\lambda_t}, \tag{15}$$

Importantly, this learning rate is feature-specific and dependent on the covariance.

The Kalman filter's covariance-dependent learning rate gives rise to several learning phenomena that have been previously explored in the literature. When the uncertainty about the hidden weights $\mathbf{m}_j$ is low compared to the uncertainty of the observation (low variance, in the diagonals of Σ), the posterior should be close to the prior, resulting in a lower learning rate. Under high uncertainty (high variance), new incoming observations should be weighted as more informative and the learning rate should be high. When there is nonzero covariance between a set of weights, these weights are updated together because they share the same $\boldsymbol{\kappa}_t$. This permits nonlocal updating of parameters; that is, parameters for features not present in the current observation may be updated if these parameters have an established covariance with parameters in the current observation. In standard TD learning, the updated equation would have the prediction error multiplied by the activity of the feature neuron and a scalar learning rate η:

$$\Delta \mathbf{m}_j = \eta \phi(s_t) \delta_j. \tag{16}$$

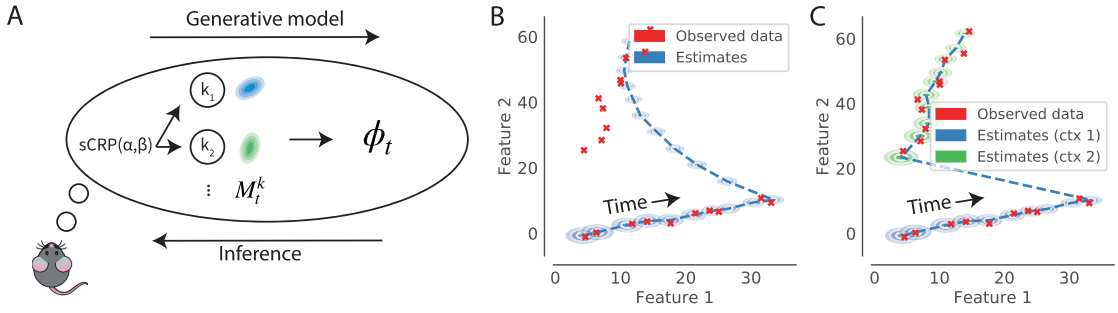
Replacing the first two terms with $\boldsymbol{\kappa}$ means that learning can occur without feature neuron activity (cf. Equation 13).

Inferring SR and Context Simultaneously

By design, Kalman filter models evolve smoothly and do not capture situations where the hidden variable undergoes large sudden changes or jumps (Figure 2). However, such sudden changes in the environment might occur when the animal switches to a different context or returns to an old one.

We can account for these jumps with a "switching LDS," which posit that there is a collection of different modes or contexts, in which each context is associated with its own LDS. This means that

Figure 2
Switching Kalman Filter Model Illustration



Note. (A) In the infinite switching Kalman filter generative model, a context k is drawn from a sticky Chinese restaurant process (sCRP) prior. The currently active context selects one of infinitely many possible linear-Gaussian models to pass through to the observations, ϕ_t. Given this generative model, the animal's goal is to infer both the SR parameters, M and the discrete context variable, k. (B) A single Kalman filter does not account for large jumps in the hidden variable that is tracked (ellipses show the posterior distribution at each time step). (C) A switching Kalman filter deals with large prediction errors by assigning them to a new context (posterior distributions are color-coded with the inferred context). (B) and (C) show a series of observations of a pair of features over time (red crosses, feature two increases monotonically with time) and their estimated posterior distribution (shaded ellipses) under Context 1 in (B) and with a model that can switch to a new Context 2 when prediction errors are large. SR = successor representation. See the online article for the color version of this figure.

$$p(\phi_{j,t}|\phi_{1:t-1}, z_t = k) = \begin{cases} \mathcal{N}(\phi_{j,t}; \mathbf{h}_t^T \mathbf{m}_{j,t}^k, \lambda_t), & \text{if } k \text{ is previously sampled} \\ \mathcal{N}(\phi_{j,t}; \mathbf{h}_t^T \boldsymbol{\mu}_0, \mathbf{h}_t^T \Sigma_0 \mathbf{h}_t + \sigma_\phi^2), & \text{otherwise.} \end{cases} . \quad (20)$$

our model switches between different SR maps M^κ that correspond to different contexts k. Since there are infinitely many possible contexts, we use a nonparametric switching LDS (Fox et al., 2011), which allows the number of inferred contexts to grow as more observations are made. This generative model corresponds to that used in Gershman et al. (2014) to model memory updating, with the difference that here the continuous hidden state will be the SR. Thus, an SR-context is chosen if it correctly predicts observations. If there are no such predictive contexts, a new SR is created.

Generative Model

In the generative process, this model assumes that a context z_t is first drawn from a sticky Chinese restaurant process (sCRP) prior. A CRP prior allows for a potentially infinite number of contexts but tends toward fewer contexts by proportionately assigning observations to contexts that already explain more observations. The "sticky" CRP has an additional bias to remain in the current context. The sCRP prior is written as:

$$p(z_t = k|\mathbf{z}_{1:t-1}) = \begin{cases} \frac{N_k + \beta\delta[z_{t-1}, k]}{\alpha + \beta + t - 1}, & \text{if } k \text{ is previously sampled context} \\ \frac{\alpha}{\alpha + \beta + t - 1}, & \text{otherwise} \end{cases} , \quad (17)$$

where N_k is the number of observations previously assigned to context k and $\delta[x, y] = 1$ if $x = y$ and 0 otherwise. The concentration parameter α controls the propensity to create new modes, and the "stickiness" parameter β determines how likely the model will stay with the current context.

After choosing a context, the generative model proceeds by evolving the state variable for each previously active context k according to the evolution equation of the LDS: $\mathbf{m}_{j,t}^k \sim \mathcal{N}(\mathbf{m}_{j,t-1}^k, Q)$. If z_t is a new context, a new SR is first drawn with columns $\mathbf{m}_{j,0}^{z_t}$ drawn from a Gaussian prior: $\mathbf{m}_{j,0}^{z_t} \sim \mathcal{N}(\boldsymbol{\mu}_0, \sigma_\mu^2 I)$. Finally, a sensory observation ϕ_t is emitted from the currently active context z_t using the observation equation: $\phi_{j,t} \sim \mathcal{N}((\mathbf{m}_{j,t}^{z_t})^T \mathbf{h}_t, \sigma_\phi^2)$.

Inference

When there is uncertainty about the context, inference requires marginalizing over all possible context histories $\mathbf{z}_{1:t}$:

$$p(\mathbf{m}_{j,t}^k|\phi_{1:t+1}) = \sum_{z_{1:t}} p(\mathbf{m}_{j,t}^k|\phi_{1:t+1}, \mathbf{z}_{1:t}) p(\mathbf{z}_{1:t}). \quad (18)$$

If there are K modes, the posterior at time t will be a mixture of K^t Gaussians, one for every possible history $z_1, \ldots z_t$. Exact inference is intractable under this exponentially growing number of modes. We therefore use the Gaussian-sum filter (Barber, 2012), which approximates the exponentially growing number of components with a smaller number of I components.

At each time step, the Gaussian-sum filter first uses the deterministic Kalman filter equations (see above) to compute the posterior for each possible current and next context and for each component of the mixture, with Kalman gain:

$$\kappa_t = \begin{cases} \frac{(\Sigma_t + Q)\mathbf{h}_t}{\lambda_t}, & \text{if } k \text{ is previously sampled context} \\ 0, & \text{otherwise} \end{cases} . \quad (19)$$

Given K previous context assignments at time t, each mixture component branches into another K components such that the resulting joint approximation contains $K^2 I$ components. This larger mixture then collapsed back to I components by merging them into I Gaussians, weighted by their likelihood:

(See above)

This corresponds to choosing contexts according to how well they predict observations: each previously used context k makes a prediction centered on $\mathbf{h}_t^T \mathbf{m}_{j,t}^k$. Thus, the log-likelihood of any existing context will be inversely proportional to the magnitude of the SR prediction error for that context, $\|\phi_t - M^T \mathbf{h}\|^2$, meaning that very large prediction errors will likely lead to the inference of a new context. Furthermore, since the variance of a context grows with the amount of time since its last occurrence, older modes will be more tolerant of prediction errors. The intuitive explanation for this is that if the animal has not seen a context for a long time, its certainty about the details of the events will have deteriorated. For simplicity, we used $I = 1$ components here. Note that the collapsing step is similar to the resampling step in particle filtering (Fearnhead & Clifford, 2003).

In summary, the Kalman SR algorithm operates by predicting the occurrence of a feature using the weights $\hat{\phi}$ and updating its prediction based from observation. Updates are sensitive to the covariance between predicted features, such that the magnitude of the update is informed by uncertainty, and covarying features are simultaneously updated (permitting learning about features not currently experienced). We model context-dependent SR learning using a nonparametric switching Kalman filter, in which the SR diffuses gradually until it jumps to either a previously activated context or to a new one. This allows us to model how uncertainty over the active context modulates inference using the SR. The Bayesian view of the SR outlined here allows us to reconcile and reinterpret some results in the animal learning literature, which we will describe in the following section.

Code Availability and Preregistration

Analysis code for this study will be made available upon publication. This study was not preregistered.

Results

In the first part of this section, we will discuss results that follow from the single Kalman filter SR described in Probabilistic SFs section. In the second part, we will discuss experimental predictions relating to the switching context model described in Inferring SR and Context Simultaneously section.

Kalman SR Simulations

Context Preexposure: Facilitation and Latent Inhibition

Prior work on contextual fear conditioning in rodents looks at the conditioned response (freezing) following an aversive stimulus (a small foot shock) received in some new environment. In these experiments, the animal receives a single shock, after varying amounts of preexposure to the environment. It has been shown that a stronger conditioned response is evoked if animals are able to explore the new environment for several minutes before the first shock, a finding known as the "context preexposure facilitation effect" (Fanselow, 2010). As shown by Stachenfeld et al. (2017), a predictive model such as the SR can account for this: During preexposure, the animal explores and learns a predictive representation of the context such that subsequent value learning is rapidly propagated across the environment. By contrast, model-free learning does not predict this because learning only occurs after the reward or punishment arrives. Hence, the reward signal cannot propagate through the environment prior to the time of shock.

Context preexposure facilitation stands in apparent contrast to "latent inhibition," which refers to the finding that preexposure to a conditioned stimulus (CS) typically *impairs* the acquisition of a conditioned response. This latent inhibition effect has been taken as evidence for the assertion that animals are Bayesian learners: As the preexposed cue is presented repeatedly, the animal's uncertainty about the expected reward associated with that cue decreases, resulting in slower subsequent learning (Gershman, 2015). Latent inhibition and facilitation are thus opposing effects that could both be driven by context preexposure. Kiernan and Westbrook (1993) showed that there is an intriguing time course to these phenomena: A brief amount of preexposure facilitates subsequent learning, whereas extended preexposure to the environment inhibits further associative learning (Figure 3A). The authors also varied the interval between entering the chamber and the shock and found that this U-shaped effect arose for both short (7 s, shown in Figure 3A) and long (60 s) intervals.

To simulate this study using our probabilistic SR model, we followed Stachenfeld et al. (2017) by setting up the fear conditioning experiment as a grid world environment, where features $\phi(s_t)$ correspond to locations or local cues, and the agent follows a random walk through this environment for varying amounts of time until it receives a single, negative reward. The Kalman SR model naturally captures the nonmonotonic relationship found by Kiernan and Westbrook (1993), as both facilitation (driven by the SR) and inhibition (driven by the Kalman filter) occur during preexposure. As the animal explores the environment, there should be facilitation early on as the SR is learned and value is generalized across the environment. However, this should be followed by inhibition after extensive training because reduced variance in the estimates of the reward weights $\mathbf{w}$ results in a decrease in Kalman gain, as shown in Figure 3B. This U-shaped curve is shown by the model as long as the SR generalizes sufficiently and as long as the reduction in variance with incoming observations outweighs the increase in variance from forgetting (see Table 2, for a parameter range). The standard, temporal difference (TD) SR (employed by Stachenfeld et al., 2017, to model fear conditioning) shows facilitation, but not inhibition (Figure 3C). In contrast, the Bayesian generalization of TD learning for value (Gershman, 2015) shows inhibition, but not facilitation (Figure 3D). Thus, both the predictive map and the uncertainty components of this model are required to capture this finding.

Transition Revaluation

A key prediction of standard temporal difference SR learning is that "reward revaluation" (changes in the reward at each state) should be easier to acquire than "transition revaluation" (changes in the transition probabilities between states), since the latter requires propagating state occupancy predictions to distal states. This is because temporal difference learning updates the SR only for experienced states, even though predictions from previous states are also ultimately affected by the change. Further experience is needed to update all affected states (note that eligibility traces could address this issue, but only for situations where the affected states are experienced in the same episode). Momennejad, Russek, et al. (2017) tested whether or not this is the case in human learning. In their experiment, participants learned about two sequences of states leading up to a reward (see Figure 4A). In the next phase, half of the participants were exposed to a transition revaluation condition, observing novel transitions leading up to the reward. The other half experienced "reward revaluation" in the form of novel reward amounts at the final states. Importantly, the novel experiences in Phase 2 did not include starting States 1 or 2, meaning that under a classical SR model, the SR for States 1 and 2 would not be updated.

While participants were significantly better at reward revaluation than transition revaluation, they were capable of some transition revaluation as well (Figure 4B). Accordingly, the authors proposed a hybrid SR model: an SR-TD agent that is also endowed with capacity for replaying experienced transitions (Figure 4E). This permits updating of the SR vectors of states 1 and 2 through simulated experience. Note that this pattern of results cannot be explained by a simple model-free (MF) or model-based (MB) strategy, or by a simple hybrid (MF–MB), as MF methods are equally unable to do either form of revaluation and MB methods are equally able to do both (Figure 4E).

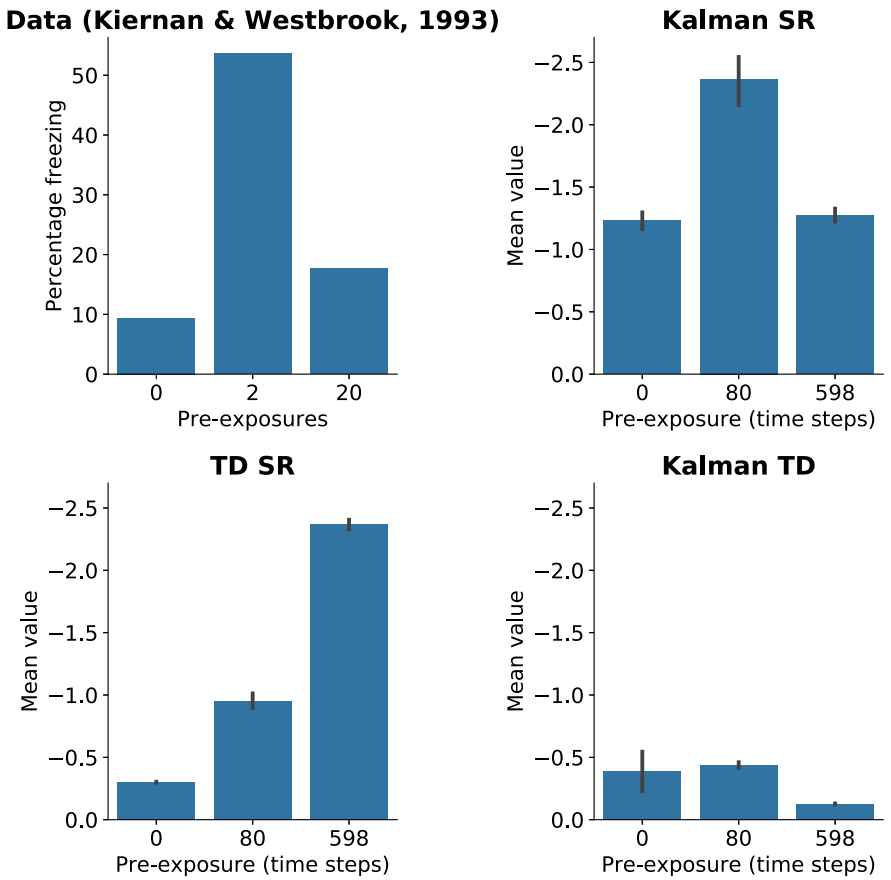
Simulating this experiment with Kalman SR shows that the model can account for the partial transition revaluation without explicit simulated experience. Kalman SR correctly learns the SR matrix after Phase 1 as well as an estimate of the covariance between features, Σ. Unlike standard temporal difference methods, Kalman TD uses the covariance matrix to estimate the Kalman gain and uses that to update the SR nonlocally. This means that after seeing $3 \rightarrow 6$, it updates not just $\psi(3)$ but also $\psi(1)$ because these entries have historically covaried, and similarly for $\psi(4)$ and $\psi(2)$.

Reward Devaluation of Preconditioned Cues

Our model similarly captures the findings of Hart et al. (2020), which show that responding to preconditioned cues is sensitive to devaluation, a hallmark of model-based learning. The particular experiment of interest here started with a preconditioning phase, during which associations were learned between pairs of neutral (i.e., nonrewarding) stimuli, followed by a conditioning phase during which a neutral stimulus was paired with a food reward (Figure 5A). After this conditioning phase, the food reward was devalued in one group but not another, by pairing it with sickness-inducing lithium chloride (LiCl). A key finding of this experiment was that healthy

Figure 3

A Brief Amount of Preexposure to an Environment Facilitates Subsequent Learning but Extended Preexposure Inhibits Learning



Note. (A) Behavioral data from Kiernan and Westbrook (1993). Mean percentage freezing scores in the shocked environment E1 for the groups receiving no, brief, or extended preexposure to E1. (B) Under the Kalman SR model interpretation, exploring the environment during preexposure allows a predictive representation to be learned. Since value is computed by multiplying the SR by the reward function, this means that longer preexposure initially facilitates learning the negative value in the environment. Prolonged preexposure, however, causes a decrease in uncertainty and therefore in Kalman gain, inhibiting further learning. Simulation results show the mean value estimated by the model. (C) A nonprobabilistic SR estimated using maximum-likelihood TD learning shows the facilitation but not the inhibition effect. (D) Applying Kalman TD to value-function learning shows the inhibition but not the facilitation effect. Error bars indicate *SEM* across 20 runs of the model. SR = successor representation; TD = temporal difference; *SEM* = standard error of the mean. See the online article for the color version of this figure.

animals' responding to the unblocked preconditioned cues (C) was sensitive to the subsequent devaluation of the food reward (Figure 5B, see also Hart et al., 2020). Thus, reward devaluation can alter stimulus–stimulus associations that were learned through the activation of dopamine neurons. Furthermore, the value inference depended on an intact orbitofrontal cortex (OFC) during the preconditioning phase (see Discussion section).

The SR naturally accommodates many preconditioning phenomena because the separate representations of stimulus–food predictions and their valence allow for flexible revaluation. Indeed, in a single-step devaluation paradigm where X is paired with food after which food is devalued, the SR does show sensitivity to devaluation (Gardner et al., 2018). However, in the experiment shown in (Figure 5A), the food reward was paired with illness in the absence of any of the neutral stimuli introduced in the preconditioning stage. This means that a standard SR agent would not be sensitive to the reward devaluation (Figure 5C). This is because in the TD SR, only stimuli that directly predict reward will change value after devaluation, and unlike X, C was never directly associated with food. Thus, even though C is associated with X and X with food, there is no devaluation sensitivity because there is no C → food association. This was also observed by Gardner et al. (2018), who simulated a very similar task with an SR model endowed with the ability to simulate offline experience (see Appendix A, Figure A1). This latter model showed the same sensitivity to devaluation that was shown by the animals.

Table 2
Parameter Range for Which Context Preexposure Leads to a U-Shaped Curve in Associability in Our Model

Symbol	Range
γ	(0.8, 1)
σ_ϕ^2	(1e−6, 0.004)
σ_v^2	(0.3, 2)
σ_0^2	(1.3, Inf)

Note. These values were obtained by simulating the experiment described in Figure 3, varying each parameter while holding all other parameters constant at the values described in Table 1.

Like the animals in Hart et al. (2020) and the replay-endowed model, Kalman SR was sensitive to the reward devaluation paradigm (Figure 5D). During the preconditioning phase, a positive covariance between C and X is learned, which means that during conditioning, C becomes directly associated to the food (Equation 13). Subsequent devaluation thus directly affects C as well as X. This permits long range temporal credit assignment without explicitly necessitating hand-engineered features or simulated sequential experience.

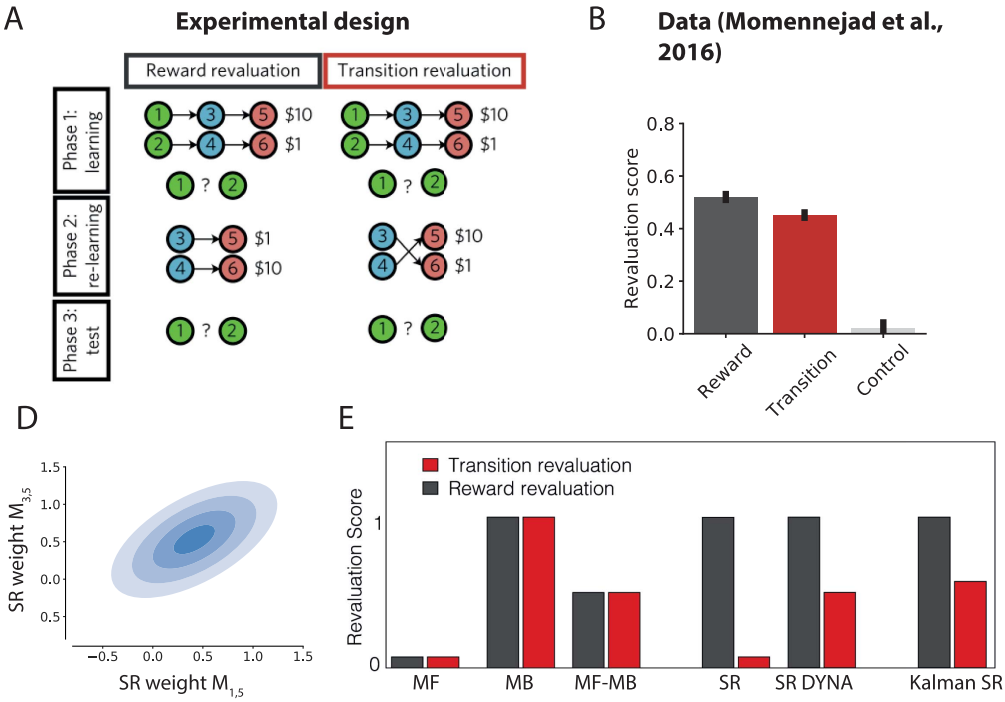
Switching Context Model Simulations

Contextual Memory

The context-switching model allows us to explain an intriguing finding in contextual fear conditioning (Figure 6). Winocur et al. (2009) exposed rodents to a conditioned stimulus–unconditioned stimulus (CS–US) pair (a tone and a foot shock) in Context A and subsequently tested for a fear response in either Context A or in a new, similar Context B, which differed from Context A in that it was smaller, had a different level of transparency of the walls and it was placed in a different room. Furthermore, the experimenters tested for a fear response after either short (24 hr) or long (28 day) delays (Figure 6A). They found that the animals learned the association in Context A but did not generalize their conditioned responding to Context B after the short delay (Figure 6D). However, the level of generalization increased with the delay interval. Furthermore, when animals were briefly exposed to the training context, as a reminder prior to test in the second context after the long delay, the generalization decreased again (Figure 6D). Thus, cross-context specificity decreases with time but can be restored with a reminder of the context.

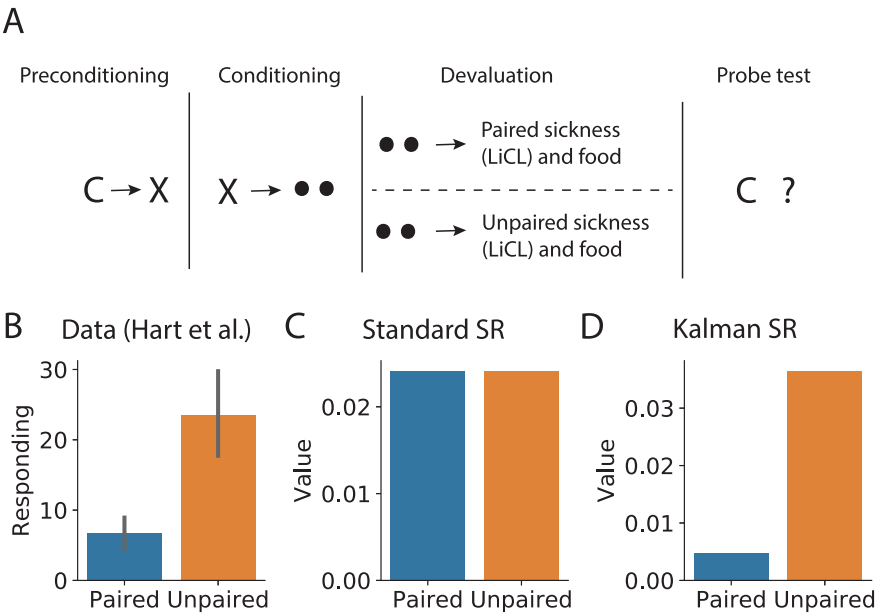
The switching Kalman SR model can explain this result in terms of switching between contextual representations. To demonstrate

Figure 4
Revaluation Experiment of Momennejad, Russek, et al. (2017)



Note. (A) Experimental design. In an initial learning phase, participants learned sequences of states, associated with high ($10) or low ($1) rewards. During a second relearning phase, either the rewards associated to the two terminal (red) states were swapped (reward revaluation) or the transitions from the middle (blue) to the terminal states were swapped (transition revaluation). (B) Human participants' revaluation scores (Momennejad, Russek, et al., 2017). (D) The joint distribution over weights $M_{1,5}$ and $M_{3,5}$ shows a positive covariance induced by the first phase of learning, which explains the revaluation from State 1 to State 5. (E) Predicted revaluation scores, change in rating $V(1)-V(2)$, between Phases 1 and 3 for different algorithms. SR = successor representation; MF = model-free; MB = model-based; SR DYNA = replay of experienced transitions to update the SR. See the online article for the color version of this figure.

Figure 5
Behaviour to Preconditioned Cue Is Sensitive to Devaluation



Note. (A) Experimental design (Hart et al., 2020). During the initial preconditioning phase, one neutral stimulus always precedes a second (*C X*). During the conditioning phase, the second stimulus is paired with a food reward. After the conditioning phase, the food is paired with lithium chloride (LiCl) to induce sickness in one group of animals. Letters denote different neutral stimuli, black circles indicate food reward. (B, D) Data and simulation results show that, like animals and unlike TD-SR, Kalman SR shows sensitivity to devaluation in this paradigm. Data in (B) replotted from Hart et al. (2020), error bars show *SEM*. SR = successor representation; TD = temporal difference; *SEM* = standard error of the mean. See the online article for the color version of this figure.

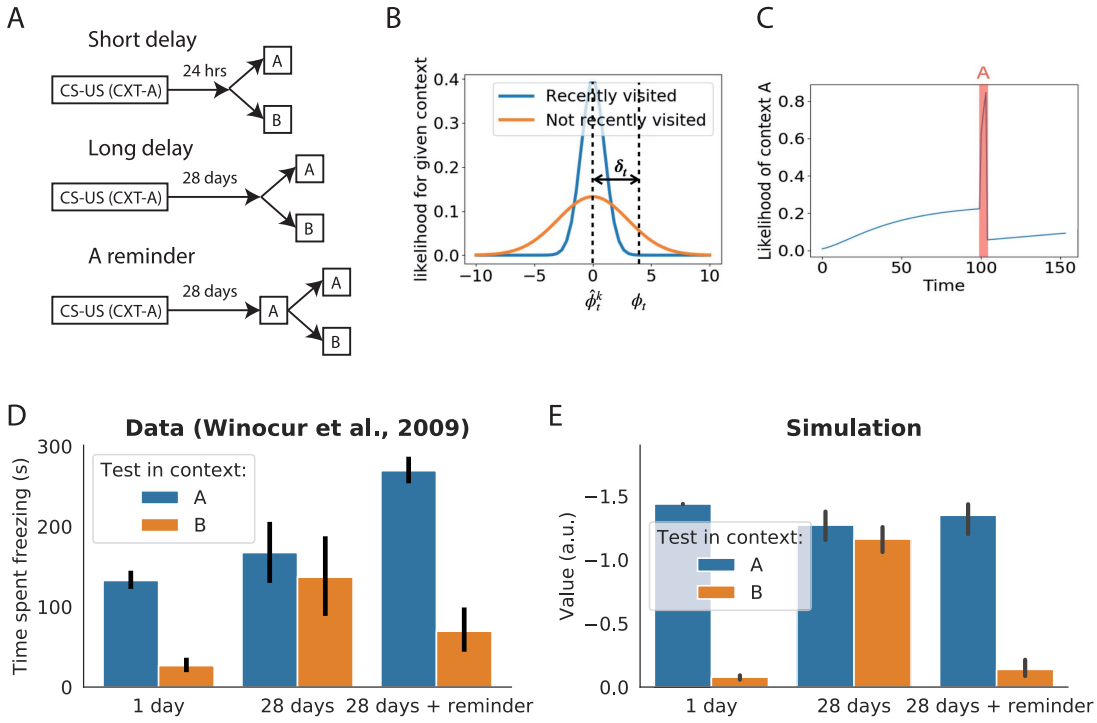
this, we simulated Contexts A and B as containing different sequences of observations. We generated observations as feature vectors drawn from two independent random transition matrices, reflecting the fact that in the experiment, different sequences of features would be observed by the animal in the different contexts. While this is not a detailed model of the complicated arrangement of stimuli observed by the animals (such as the texture of the walls of the chamber, the metal rod floor, the configuration of furniture, fixtures and lighting conditions, which were all different in the two contexts; Winocur et al., 2009), this setup is meant to demonstrate the key point: When the distance between the observation and the predicted observation given a context (successor prediction error: $\phi_{j,t} - \mathbf{h}_t^T \mathbf{m}_{j,t}^k$) is large, the model will assign a low likelihood to that context. If the posterior probability of every currently active context is low, the model will be likely to assign the observation to a new cluster, initiating the use of a new, separate predictive map. Furthermore, since the variance of clusters that have not been visited for a while keeps growing (as can be seen in the Kalman filter updates above), old clusters will be more "tolerant" to prediction errors, that is, their likelihood will be larger, even for larger distances (Figure 6B). A reexposure to the original context will reduce the variance again, restoring the sensitivity to prediction errors (Figure 6C). Thus, the model predicts that learning is highly context-specific early on but will lose context-specificity with time because of growing within-context uncertainty. Furthermore, because the Kalman filter's covariance updates do not depend on the reward outcomes, mere reexposure to the features of Context A should restore the context-specificity of the learned predictions, thus recapitulating the results observed by Winocur et al. (2009; Figure 6E).

Contextual Generalization

In addition to elapsed time, the amount of contextual generalization is also dependent on the amount of initial exposure to the original context. This was shown by Kiernan and Westbrook (1993) in a follow-up experiment similar to that described in the previous section (Figure 7). The amount of context preexposure was again varied, and the propensity of animals to show conditioned responding was now recorded both in the original environmental context and in a novel environmental context, to test whether animals would generalize their responding in the original environment to a novel context. Recall from the previous section that these authors showed a nonmonotonic effect of preexposure duration *within* a context, whereby preexposure to a context first facilitates, then inhibits learning (Figure 3). This is explained in our model because of the SR-driven facilitation and inference-driven inhibition. In contrast, increasing preexposure duration monotonically *decreases* the amount of generalization of the fear response to a second context (cf. blue and orange bars; Figure 7A). Under our model, this can be explained because increased exposure to the context results in a sharper posterior over the SR and reward weight parameters (Figure 7B).

Figure 6

Contextual Memory Experiment by Winocur et al. (2009)



Note. (A) Experimental design. In the short-delay condition, animals were conditioned in Context A and then tested in Context A and a different Context B, 24 hr later. In the long-delay condition, there was a 28-day delay between conditioning and testing. In the reminder condition, animals were briefly reintroduced to Context A, without administering the conditioned stimulus (CS) or unconditioned stimulus (US), before testing. (B) In the model, each context's likelihood is a Gaussian centered on that context's predicted observation ϕ. The larger the SR prediction error for that mode, the lower the likelihood. Submodels for contexts that have not been active for a long time will have higher variance around the predicted mean and be more tolerant to prediction errors. (C) A reintroduction to the original context (red shaded region) reduces that context model's variance, and hence it reduces the likelihood of inferring the context given a large prediction error. (D) Data replotted from "Changes in Context-Specificity During Memory Reconsolidation: Selective Effects of Hippocampal Lesions," by G. Winocur, P. W. Frankland, M. Sekeres, S. Fogel and M. Moscovitch, 2009, *Learning & Memory*, *16*(11), pp. 722–729 (https://doi.org/10.1101/lm.1447209) showing the time spent freezing in response to the CS in different conditions. (E) Simulation results showing the state value estimate when the CS is shown in different conditions. As in the data, value is increasingly generalized to Context B, but a reminder of the original context restores context-specificity. Error bars indicate *SEM* across 20 runs of the model. CS = conditioned stimulus; SR = successor representation; *SEM* = standard error of the mean. See the online article for the color version of this figure.

This *reduces* contextual generalization because the likelihood of the original Context 1 will be low in Context 2: since Context 1's SR is represented with a high precision, small differences in Context 2 will not cause it to be grouped with Context 1.

Discussion

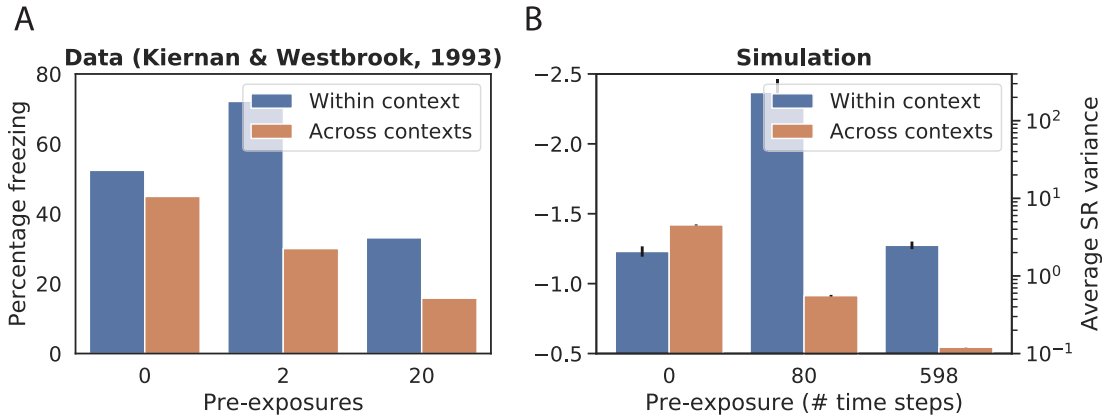
In this article we addressed the problem of learning in uncertain and context-dependent situations, by using probabilistic predictive maps within an RL framework. The SR constitutes an efficient, flexible middle ground between model-based and model-free RL algorithms by separating reward representations from cached long-run state predictions. Here, we introduce a probabilistic SR model using KTD that supports principled handling of uncertainty about state feature predictions and interdependencies between these predictions. This model is extended to a switching Kalman filter that switches between different modes or contexts. These models, described at the computational and algorithmic levels, capture human and animal behavior in settings of context preexposure, transition revaluation, and contextual generalization and memory.

Relationship to Previous Models

Our model builds on and synthesizes a set of computational principles from previous work: generalization using a predictive map and probabilistic updating using a nonparametric switching Kalman filter.

Several previous theories have given explanations for the facilitation, inhibition, and generalization effects observed by Kiernan and Westbrook (1993, Figure 3). In the argument of Honey and Hall (1989), increased discrimination between contexts can arise directly as a result of latent inhibition. Upon preexposure to Context A, when

Figure 7
Contextual Generalization and Uncertainty



Note. (A) Contextual discrimination data replotted from "Effects of Exposure to a to-Be-Shocked Environment Upon the Rat's Freezing Response: Evidence for Facilitation, Latent Inhibition, and Perceptual Learning," M. J. Kiernan and R. F. Westbrook, 1993, *The Quarterly Journal of Experimental Psychology Section B, 46*(3b), pp. 271–288 (https://doi.org/10.1080/14640749308401089; left). Blue bars show conditioned freezing responses after conditioning for animals preexposed 0, 2, and 20 times to environment 1. Orange bars show the conditioned response to the same cue in a different environmental context. (B) Model simulation results showing the negative value estimated by the model after conditioning as a function of preexposure time in blue. In orange, the average variance of the Kalman SR model is shown. Error bars indicate *SEM* across 20 runs of the model. SR = successor representation; *SEM* = standard error of the mean. See the online article for the color version of this figure.

features common to A and B are more salient than the features unique to A, the common features undergo latent inhibition. This means that subsequent conditioning primarily affects the features unique to A, thus decreasing generalization. However, as pointed out by Kiernan and Westbrook (1993), this theory cannot fully account for the results in Figure 7A. If the common features are indeed more salient, a brief preexposure should result in the bulk of associative strength being acquired by the common elements, which would not result in more discrimination. McLaren and Mackintosh (2000) presented a recurrent network model which could explain the findings of Kiernan and Westbrook (1993), although the experiment was not simulated. In the network model, generalization and latent inhibition both arise as a result of the construction through associative learning of an integrated representation of the features comprising the environment. They assume that a subset of these features are sampled during every given observation and that weights between simultaneously active features are adapted so as to minimize the error (the difference between the external input and the elements corresponding to the other features). If features are sampled together very consistently, this leads to latent inhibition, but if they are sampled variably this can lead to facilitation because one observed feature can retrieve others. The resulting process of "unitization" could explain the observed increase in learning and a decrease in generalization after a brief preexposure. For this to be true, the authors further assumed that in Kiernan and Westbrook's experiment, the common features to Contexts A and B were sampled more consistently than the unique features.

How does this relate to our model? In our model, the preexposure-induced facilitation also results from a more integrated representation, in this case represented by the SR. However, the decay in generalization is because of a reduction in uncertainty, which increases the probability of a separate context being inferred, rather than latent inhibition of the common elements between two contexts. This means that different predictions can be drawn from our model versus that of McLaren and Mackintosh (2000). In their model, the decrease in generalization is dependent on the assumption that there is biased sampling of common features over unique ones. This should not be observed if the features common between two contexts are less salient than the features unique to each context. In our model, the reduction in uncertainty about SR weights mediates the decay in generalization. If this uncertainty reduction is less strong, for example, by increasing the amount of time between preexposures, this should partially cancel the generalization decay effect.

Under partial reinforcement, the Pearce–Hall model predicts that the associability should be high, resulting in faster extinction compared to deterministic reinforcement (Pearce & Hall, 1980). The opposite is true; however, extinction is slower after partial reinforcement, a phenomenon known as the partial reinforcement extinction effect (PREE; Gibbon et al., 1980; Haselgrove et al., 2004; Rescorla, 1999). At a first glance, this effect seems counterintuitive from the point of view of Bayesian theories which predict faster learning under high uncertainty about the weights. However, the computational problem an animal faces during extinction is that of nonstationarity: learned CS–US associations might not be valid in the future. The PREE can therefore be reconciled with Bayesian theories by positing models that deal optimally with changes in the stochastic parameters (Gallistel, 2012). For these models, discriminating between the conditioning and extinction phase is more difficult when these have similar rates of reinforcement (in the case of partial reinforcement; Courville et al., 2006). Similarly, for models that, like the model presented here, infer discrete latent causes, the hypothesis that the conditioning and extinction phases are generated by the same latent cause is more likely in the partial reinforcement condition (Gershman & Niv, 2012).

Potential Roles for Replay

An attractive feature of models such as the Kalman filter that track the covariance between different weights is that this allows for retrospective revaluation. This feature has previously been used to explain learning phenomena such as backward blocking (Dayan & Yu, 2003; Gershman, 2015) and more sophisticated Kalman filter models in which stochasticity and volatility parameters are estimated from data can further extend explanations to effects like the robustness of partial reinforcement (Piray & Daw, 2021a). Applied to SR learning, we have shown that this can extend to reevaluating states after a change in the transition structure (Figures 4, 5, and A1). These effects have been explained in the past by positing that agents augment their SR learning with a replay buffer that can replay experienced transitions to update the SR offline (Gardner et al., 2018; Momennejad, Russek, et al., 2017). In fact, these two explanations might be closely related: In the neural network implementation of the Kalman filter introduced by Dayan and Kakade (2001) and applied to Kalman TD by Gershman (2017b), the covariance matrix is approximated by a recurrent layer. Given a feature vector, the network activates other features whose weights positively covary with the weights of the currently activated features and it deactivates features whose weights negatively covary. This process can be seen as a covariance-based memory retrieval process similar to an attractor network. The "replay" process in this model amounts to a covariance-based memory retrieval process, similar to an attractor network: given the current feature vector, features whose weights positively covary with the weights of the active features are activated, and features whose weights negatively covary are deactivated. Applying a prediction error update to this "replay vector" approximates the Kalman TD algorithm (Gershman, 2017b). This is different from experience replay, in which experienced sequences of states are replayed (Lin, 1992), and simulated experience, in which possible experienced sequences are generated from a transition model (Momennejad, Otto, et al., 2017; Sutton, 1991); rather, multiple states are potentially activated and updated simultaneously (like the form of reactivation described in Manning, 2021).

Thus, the Kalman filter model suggests a biologically plausible implementation of a rapid covariance-based replay mechanism that would capture these results. It is interesting to note that Momennejad, Russek, et al. (2017) also found that transition revaluation was associated with longer reaction times than reward revaluation. Under our interpretation, this could be attributable to either uncertainty leading to longer reaction times or the recurrent dynamics of settling into an attractor state under the biological implementation. Note that another related model achieves replanning after transition changes without replay by adding a low-rank correction matrix to the original representation (Piray & Daw, 2021b).

An additional reduction in uncertainty about the SR could be achieved using offline inference or smoothing. The Kalman filter's uncertainty estimates could be an interesting measure for determining which states should be replayed (Evans & Burgess, 2019). An alternative metric for the utility of replaying a specific state, suggested by Mattar and Daw (2017), is the product of a gain and need term, where the need term corresponds to the SR and the gain term quantifies the net increase in value expected after a policy change in a given state. This latter measure does not explicitly take into account uncertainty, but such a term might be approximated using the *value of information*, which can be computed from uncertainty estimates (Dearden et al., 1998). In addition to prioritizing sequences of replayed states, information about uncertainty may also be used to direct exploration to states with high uncertainty (see Malekzadeh et al., 2022, for an application of Kalman SR in active learning). Replay and exploration both depend on an ability to generate sequential samples and correspond to different optimal sampling regimes, possibly mediated by entorhinal grid cells (McNamee et al., 2021).

For the context model, another interesting avenue for further research is to investigate whether we can understand replay as offline inference (i.e., smoothing) in the case of multiple maps. In switching Kalman filters, smoothing does not only sharpen the posterior of the within-mode continuous latent variable, it also makes the posterior over modes more precise (Barber, 2012). In this context, replay could serve the function of better separating out different maps from each other, or alternatively, to merge maps where this is appropriate. Guo et al. (2020) found evidence that a single coherent map is being built during sleep. In Lever et al. (2002), maps for environments of different geometries differentiate within trials but they get more similar again between trials.

Limitations

We make several assumptions for simplicity and in order to make this model tractable. First, the observation noise is assumed to be white (i.e., independent per time step) and constant. Since the white noise assumption does not hold in many cases, we have included an analysis of the white noise assumption and an alternative model in Appendix B. Second, following the value estimation method described by Geist and Pietquin (2010), we chose a random walk model for describing the evolution process on the SR parameters. With this identity evolution model, all inference burden is put on the observation process. This means that Kalman TD is simply a reinterpretation of TD learning, that is, a model-free way to estimate the SR. Given this evolution model, and assuming independent noise, we could make the assumption that the parameters for each successor feature (i.e., each column of the weight matrix) were independent such that, effectively, the Kalman SR model consists of N independent filters. Furthermore, since the evolution of the covariance matrix is independent of the prediction errors, the covariance matrix corresponding to each column was the same. Of course, in reality, there do exist dependencies between the different columns of M. For example, in the tabular case, visiting any particular state more than expected means that all other states will be visited less than expected. A more sophisticated evolution model could exploit these dependencies; however, this would break the independence assumptions and thereby increase the computational burden. Inference in the switching Kalman filter is generally intractable, and therefore, we have adopted a Gaussian-sum filter-based approximation in our simulations. The experiments we modeled here do not speak to one or another form of approximate inference, but this is an interesting avenue for further research.

Suggested Neural Information Processing Architecture

Although the model presented here is normative and agnostic of implementation, a complete account will of course need to incorporate the brain regions involved in different parts of the model. We hypothesize that the different SR maps are encoded by

the hippocampus, which shows many resemblances to the SR (Stachenfeld et al., 2017). For example, the firing fields of hippocampal place cells show experience-dependent skewing, consistent with a prediction of future locations (Mehta et al., 2000). Neuroimaging studies have furthermore shown predictive coding of nonspatial states (Garvert et al., 2017; Schapiro et al., 2016). Accordingly, we propose that the prediction error-mediated switching between contextual SR maps corresponds to hippocampal remapping (Sanders et al., 2020). It should be noted, however, that a recent direct neuroimaging test of the SR proved inconclusive about the role of the hippocampus (Russek et al., 2021) and that a study in rodents did not find evidence for SR coding in dorsal CA1 (Duvelle et al., 2021). As another possibility, the OFC has previously been shown to be involved in predicting both reward outcomes (Gottfried et al., 2003; Schoenbaum et al., 1998) and sensory events (Chaumon et al., 2014) and is crucial for learning the stimulus–stimulus associations in Hart et al. (2020; Figure 5). Accordingly, Wilson et al. (2014) have proposed that the OFC encodes a cognitive map of task space.

As for the prediction error used to update the SR, evidence from optogenetic studies in rodents suggests that these could be encoded by a population of dopamine (DA) neurons in the ventral tegmental area (see Figure A1; Gardner et al., 2018; Sharpe et al., 2017). Dopamine neurons are furthermore known to modulate the hippocampus, which in turn projects to the striatum (Lisman & Grace, 2005). Taken together, this suggests an information processing architecture in which SR maps are encoded in the hippocampus and/or OFC and updated by dopaminergic modulation. In this hypothesis, the striatum could compute values from the SR, feeding into action selection.

Conclusions

In this article, we introduced a model of reward prediction under uncertainty and context-dependent learning. To achieve this, we used a model based on the SR in which a distribution over SR weights is estimated using Kalman TD. In the model, the appropriate context is chosen based on how well a certain set of SR parameters serve for predicting the current observations. This model captures several learning phenomena, including the effects of context preexposure on learning and generalization, the effects of reward devaluation after preconditioning and the context-specificity of memories. This article demonstrates that these hitherto unconnected themes in animal learning can be unified under a single model that combines the principles of predictive maps and probabilistic updating. We believe that this type of model has broad explanatory scope within psychological science.

References

Barber, D. (2012). *Bayesian reasoning and machine learning*. Cambridge University Press.

Barreto, A., Munos, R., Schaul, T., & Silver, D. (2016). *Successor features for transfer in reinforcement learning*. PsyArXiv. http://arxiv.org/abs/1606.05312

Bono, J., Zannone, S., Pedrosa, V., & Clopath, C. (2021). *Learning predictive cognitive maps with spiking neurons during behaviour and replays*. PsyArXiv. https://doi.org/10.1101/2021.08.16.456545

Bouton, M. E. (2004). Context and behavioral processes in extinction. *Learning & Memory*, *11*(5), 485–494. https://doi.org/10.1101/lm.78804

Bouton, M. E., & Bolles, R. C. (1979). Role of conditioned contextual stimuli in reinstatement of extinguished fear. *Journal of Experimental Psychology: Animal Behavior Processes*, *5*(4), 368–378. https://doi.org/10.1037/0097-7403.5.4.368

Brea, J., Gaál, A. T., Urbanczik, R., & Senn, W. (2016). Prospective coding by spiking neurons. *PLOS Computational Biology*, *12*(6), Article e1005003. https://doi.org/10.1371/journal.pcbi.1005003

Brockman, G., Cheung, V., Pettersson, L., Schneider, J., Schulman, J., Tang, J., & Zaremba, W. (2016). *Openai gym*. PsyArXiv. https://doi.org/10.48550/arXiv.1606.01540

Brunec, I. K., & Momennejad, I. (2022). Predictive representations in hippocampal and prefrontal hierarchies. *Journal of Neuroscience*, *42*(2), 299–312. https://doi.org/10.1523/JNEUROSCI.1327-21.2021

Chang, S.-D., & Liang, K. (2017). The hippocampus integrates context and shock into a configural memory in contextual fear conditioning. *Hippocampus*, *27*(2), 145–155. https://doi.org/10.1002/hipo.22679

Chaumon, M., Kveraga, K., Barrett, L. F., & Bar, M. (2014). Visual predictions in the orbitofrontal cortex rely on associative content. *Cerebral Cortex*, *24*(11), 2899–2907. https://doi.org/10.1093/cercor/bht146

Courville, A. C., Daw, N. D., & Touretzky, D. S. (2006). Bayesian theories of conditioning in a changing world. *Trends in Cognitive Sciences*, *10*(7), 294–300. https://doi.org/10.1016/j.tics.2006.05.004

Daw, N. D., Niv, Y., & Dayan, P. (2005). Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nature Neuroscience*, *8*(12), 1704–1711. https://doi.org/10.1038/nn1560

Dayan, P. (1993). Improving generalisation for temporal difference learning: The successor representation. *Neural Computation*, *5*(4), 613–624. https://doi.org/10.1162/neco.1993.5.4.613

Dayan, P., & Kakade, S. (2001). Explaining away in weight space. In T. K. Leen, T. G. Dietterich, & V. Tresp (Eds.), *Advances in Neural Information Processing Systems* (Vol. 13, pp. 451–457). MIT Press.

Dayan, P., & Yu, A. (2003). Uncertainty and learning. *IETE Journal of Research*, *49*(2–3), 171–181. https://doi.org/10.1080/03772063.2003.11416335

de Cothi, W., & Barry, C. (2020). Neurobiological successor features for spatial navigation. *Hippocampus*, *30*(12), 1347–1355. https://doi.org/10.1002/hipo.23246

De Cothi, W., Nyberg, N., Griesbauer, E.-M., Ghanamé, C., Zisch, F., Lefort, J. M., Fletcher, L., Newton, C., Renaudineau, S., Bendor, D., Grieves, R., Duvelle, É., Barry, C., & Spiers, H. J. (2022). Predictive maps in rats and humans for spatial navigation. *Current Biology*, *32*(17), 3676–3689.e5. https://doi.org/10.1016/j.cub.2022.06.090

Dearden, R., Friedman, N., & Russell, S. (1998). Bayesian Q-learning. *Proceedings of the AAAI, 1998*, 761–768.

Duvelle, É., Grieves, R. M., Liu, A., Jedidi-Ayoub, S., Holeniewska, J., Harris, A., Nyberg, N., Donnarumma, F., Lefort, J. M., & Jeffery, K. J. (2021). Hippocampal place cells encode global location but not connectivity in a complex space. *Current Biology*, *31*(6), 1221–1233. https://doi.org/10.1016/j.cub.2021.01.005

Engel, Y., Mannor, S., & Meir, R. (2005). *Reinforcement learning with Gaussian processes* [Conference session]. Proceedings of the 22nd international conference on Machine learning (pp. 201–208).

Evans, T., & Burgess, N. (2019). Explaining away in weight space. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, & R. Garnett (Eds.), *Advances in neural information processing systems 32* (pp. 451–457). MIT Press.

Fanselow, M. S. (2010). From contextual fear to a dynamic view of memory systems. *Trends in Cognitive Sciences*, *14*(1), 7–15. https://doi.org/10.1016/j.tics.2009.10.008

Fearnhead, P., & Clifford, P. (2003). On-line inference for hidden Markov models via particle filters. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, *65*(4), 887–899. https://doi.org/10.1111/1467-9868.00421

Fox, E., Sudderth, E. B., Jordan, M. I., & Willsky, A. S. (2011). Bayesian nonparametric inference of switching dynamic linear models. *IEEE*

Transactions on Signal Processing, *59*(4), 1569–1585. https://doi.org/10.1109/TSP.2010.2102756

Frémaux, N., Sprekeler, H., & Gerstner, W. (2013). Reinforcement learning using a continuous time actor-critic framework with spiking neurons. *PLOS Computational Biology*, *9*(4), Article e1003024. https://doi.org/10.1371/journal.pcbi.1003024

Gallistel, C. R. (2012). Extinction from a rationalist perspective. *Behavioural Processes*, *90*(1), 66–80. https://doi.org/10.1016/j.beproc.2012.02.008

Gardner, M. P. H., Schoenbaum, G., & Gershman, S. J. (2018). Rethinking dopamine as generalized prediction error. *Proceedings of the Royal Society B: Biological Sciences*, *285*(1891), Article 20181645. https://doi.org/10.1098/rspb.2018.1645

Garvert, M. M., Dolan, R. J., & Behrens, T. E. (2017). A map of abstract relational knowledge in the human hippocampal–entorhinal cortex. *eLife*, *6*, 1–20. https://doi.org/10.7554/eLife.17086

Geerts, J. P., Chersi, F., Stachenfeld, K. L., & Burgess, N. (2020). A general model of hippocampal and dorsal striatal learning and decision making. *Proceedings of the National Academy of Sciences*, *117*(49), 31427–31437. https://doi.org/10.1073/pnas.2007981117

Geerts, J. P., Stachenfeld, K., & Burgess, N. (2019). *Probabilistic successor representations with Kalman temporal differences* [Conference session]. 2019 Conference on Cognitive Computational Neuroscience. https://doi.org/10.32470/CCN.2019.1323-0

Geist, M., & Pietquin, O. (2010). Kalman temporal differences. *Journal of Artificial Intelligence Research*, *39*, 483–532. https://doi.org/10.1613/jair.3077

Gershman, S. J. (2015). A unifying probabilistic view of associative learning. *PLOS Computational Biology*, *13*(11), Article e1004567. https://doi.org/10.1371/journal.pcbi.1004567

Gershman, S. J. (2017a). Context-dependent learning and causal structure. *Psychonomic Bulletin & Review*, *24*(2), 557–565. https://doi.org/10.3758/s13423-016-1110-x

Gershman, S. J. (2017b). Dopamine, inference, and uncertainty. *Neural Computation*, *29*(12), 3311–3326. https://doi.org/10.1162/neco_a_01023

Gershman, S. J. (2018). The successor representation: Its computational logic and neural substrates. *The Journal of Neuroscience*, *38*(33), 7193–7200. https://doi.org/10.1523/JNEUROSCI.0151-18.2018

Gershman, S. J., Blei, D. M., & Niv, Y. (2010). Context, learning, and extinction. *Psychological Review*, *117* (1), 197–209. https://doi.org/10.1037/a0017808

Gershman, S. J., & Niv, Y. (2012). Exploring a latent cause theory of classical conditioning. *Learning and Behavior*, *40*(3), 255–268. https://doi.org/10.3758/s13420-012-0080-8

Gershman, S. J., Radulescu, A., Norman, K. A., & Niv, Y. (2014). Statistical computations underlying the dynamics of memory updating. *PLOS Computational Biology*, *10*(11), Article e1003939. https://doi.org/10.1371/journal.pcbi.1003939

Gibbon, J., Farrell, L., Locurto, C. M., Duncan, H. J., & Terrace, H. S. (1980). Partial reinforcement in autoshaping with pigeons. *Animal Learning & Behavior*, *8*(1), 45–59. https://doi.org/10.3758/BF03209729

Gottfried, J. A., O'Doherty, J., & Dolan, R. J. (2003). Encoding predictive reward value in human amygdala and orbitofrontal cortex. *Science*, *301* (5636), 1104–1107.

Guo, W., Zhang, J., Newman, J., & Wilson, M. (2020). Latent learning drives sleepdependent plasticity in distinct CA1 subpopulations. bioRxiv. https://doi.org/10.1101/2020.02.27.967794

Hansen, S., Dabney, W., Barreto, A., Van de Wiele, T., Warde-Farley, D., & Mnih, V. (2019). *Fast task inference with variational intrinsic successor features*. PsyArXiv. http://arxiv.org/abs/1906.05030

Hart, E. E., Sharpe, M. J., Gardner, M. P., & Schoenbaum, G. (2020). Responding to preconditioned cues is devaluation sensitive and requires orbitofrontal cortex during cue-cue learning. *eLife*, *9*, 1–11. https://doi.org/10.7554/eLife.59998

Haselgrove, M., Aydin, A., & Pearce, J. M. (2004). A partial reinforcement extinction effect despite equal rates of reinforcement during Pavlovian conditioning. *Journal of Experimental Psychology: Animal Behavior Processes*, *30*(3), 240–250. https://doi.org/10.1037/0097-7403.30.3.240

Heald, J. B., Lengyel, M., & Wolpert, D. M. (2021). Contextual inference underlies the learning of sensorimotor repertoires. *Nature*, *600*(7889), 489–493. https://doi.org/10.1038/s41586-021-04129-3

Honey, R. C., & Hall, G. (1989). Enhanced discriminability and reduced associability following flavor preexposure. *Learning and Motivation*, *20*(3), 262–277. https://doi.org/10.1016/0023-9690(89)90008-8

Kalman, R. E. (1960). A new approach to linear filtering and prediction problems 1. *Journal of Basic Engineering*, *82*(1), 35–45. https://doi.org/10.1115/1.3662552

Kamin, L. J. (1969). Predictability. surprise, attention, and conditioning. In B. A. Campbell & R. M. Church (Eds.), *Punishment aversive behavior* (pp. 279–296). Appleton-Century-Crofts.

Kiernan, M. J., & Westbrook, R. F. (1993). Effects of exposure to a to-be-shocked environment upon the Rat's freezing response: Evidence for facilitation, latent inhibition, and perceptual learning. *The Quarterly Journal of Experimental Psychology Section B*, *46*(3b), 271–288. https://doi.org/10.1080/14640749308401089

Lever, C., Wills, T., Cacucci, F., Burgess, N., & O'Keefe, J. (2002). Long-term plasticity in hippocampal place-cell representation of environmental geometry. *Nature*, *416*(6876), 90–94. https://doi.org/10.1038/416090a

Lin, L.-J. (1992). Self-improving reactive agents based on reinforcement learning, planning and teaching. *Machine Learning*, *8*(3), 293–321. https://doi.org/10.1007/BF00992699

Lisman, J. E., & Grace, A. A. (2005). The hippocampal-VTA loop: Controlling the entry of information into long-term memory. *Neuron*, *46*(5), 703–713. https://doi.org/10.1016/J.NEURON.2005.05.002

Malekzadeh, P., Salimibeni, M., Hou, M., Mohammadi, A., & Plataniotis, K. N. (2022). AKF-SR: Adaptive Kalman filtering-based successor representation. *Neurocomputing*, *467*, 476–490. https://doi.org/10.1016/J.NEUCOM.2021.10.008

Manning, J. R. (2021). Episodic memory: Mental time travel or a quantum "memory wave" function? *Psychological Review*, *128*(4), 711–725. https://doi.org/10.1037/rev0000283

Mattar, M. G., & Daw, N. D. (2017). *A rational model of prioritized experience replay* [Conference session]. The 3rd Multidisciplinary Conference on Reinforcement Learning and Decision Making, The University of Michigan. https://rldm.org/wp-content/uploads/2017/06/RLDM17AbstractsBooklet.pdf

McLaren, P. L., & Mackintosh, N. J. (2000). An elemental model of associative learning: I. Latent inhibition and perceptual learning. *Animal Learning & Behavior*, *28*(3), 211–246. https://doi.org/10.3758/BF03200258

McNamee, D. C., Stachenfeld, K. L., Botvinick, M. M., & Gershman, S. J. (2021). Flexible modulation of sequence generation in the entorhinal–hippocampal system. *Nature Neuroscience*, *24*(6), 851–862. https://doi.org/10.1038/s41593-021-00831-7

Mehta, M. R., Quirk, M. C., & Wilson, M. A. (2000). Experience-dependent asymmetric shape of hippocampal receptive fields. *Neuron*, *25*, 707–715. https://doi.org/10.1016/s0896-6273(00)81072-7

Momennejad, I., Otto, A. R., Daw, N. D., & Norman, K. A. (2017). Offline replay supports planning: FMRI evidence from reward revaluation. *bioRxiv*, *100*, Article 196758. https://doi.org/10.1101/196758

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

Murphy, K. (1998). *Switching Kalman filters*. Department of Computer Science, University of California. https://www.cs.ubc.ca/~murphyk/Papers/skf.pdf

Pearce, J. M., & Hall, G. (1980). A model for Pavlovian learning: Variations in the effectiveness of conditioned but not of unconditioned stimuli. *Psychological Review*, *87*(6), 532–552. https://psycnet.apa.org/record/1981-02676-001

Piray, P., & Daw, N. D. (2021a). A model for learning based on the joint estimation of stochasticity and volatility. *Nature Communications*, *12*(1), Article 6587. https://doi.org/10.1038/s41467-021-26731-9

Piray, P., & Daw, N. D. (2021b). Linear reinforcement learning in planning, grid fields, and cognitive control. *Nature Communications*, *12*(1), Article 4942. https://doi.org/10.1038/s41467-021-25123-3

Rescorla, R. A. (1999). Within-subject partial reinforcement extinction effect in autoshaping. *The Quarterly Journal of Experimental Psychology Section B*, *52*(1), 75–87. https://doi.org/10.1080/713932693

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

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

Sanders, H., Wilson, M. A., & Gershman, S. J. (2020). Hippocampal remapping as hidden state inference. *eLife*, *9*, 1–31. https://doi.org/10.7554/eLife.51140

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. https://doi.org/10.1002/hipo.22523

Schoenbaum, G., Chiba, A. A., & Gallagher, M. (1998). Orbitofrontal cortex and basolateral amygdala encode expected outcomes during learning. *Nature Neuroscience*, *1*(2), 155–159. https://doi.org/10.1038/407

Schultz, W., Dayan, P., & Montague, P. R. (1997). A neural substrate of prediction and reward. *Science*, *275*(5306), 1593–1599. https://doi.org/10.1126/science.275.5306.1593

Sharpe, M. J., Chang, C. Y., Liu, M. A., Batchelor, H. M., Mueller, L. E., Jones, J. L., Niv, Y., & Schoenbaum, G. (2017). Dopamine transients are sufficient and necessary for acquisition of model-based associations. *Nature Neuroscience*, *20*(5), 735–742. https://doi.org/10.1038/nn.4538

Stachenfeld, K. L., Botvinick, M. M., & Gershman, S. J. (2017). The Hippocampus as a Predictive Map. *Nature Neuroscience*, *20*(11), 1643–1653. https://doi.org/10.1038/nn.4650

Sutton, R. S. (1991). Dyna, an integrated architecture for learning, planning, and reacting. *ACM SIGART Bulletin*, *2*(4), 160–163. https://doi.org/10.1145/122344.122377

Sutton, R. S., & Barto, A. (1998). *Reinforcement learning: An introduction* (Vol. 9). MIT Press. https://doi.org/10.1109/tnn.1998.712192

Tomov, M. S., Schulz, E., & Gershman, S. J. (2021). Multi-task reinforcement learning in humans. *Nature Human Behaviour*, *5*(6), 764–773. https://doi.org/10.1038/s41562-020-01035-y

Wilson, R. C., Takahashi, Y. K., Schoenbaum, G., & Niv, Y. (2014). Orbitofrontal cortex as a cognitive map of task space. *Neuron*, *81*(2), 267–278. https://doi.org/10.1016/j.neuron.2013.11.005

Winocur, G., Frankland, P. W., Sekeres, M., Fogel, S., & Moscovitch, M. (2009). Changes in context-specificity during memory reconsolidation: Selective effects of hippocampal lesions. *Learning & Memory*, *16*(11), 722–729. https://doi.org/10.1101/lm.1447209

Appendix A

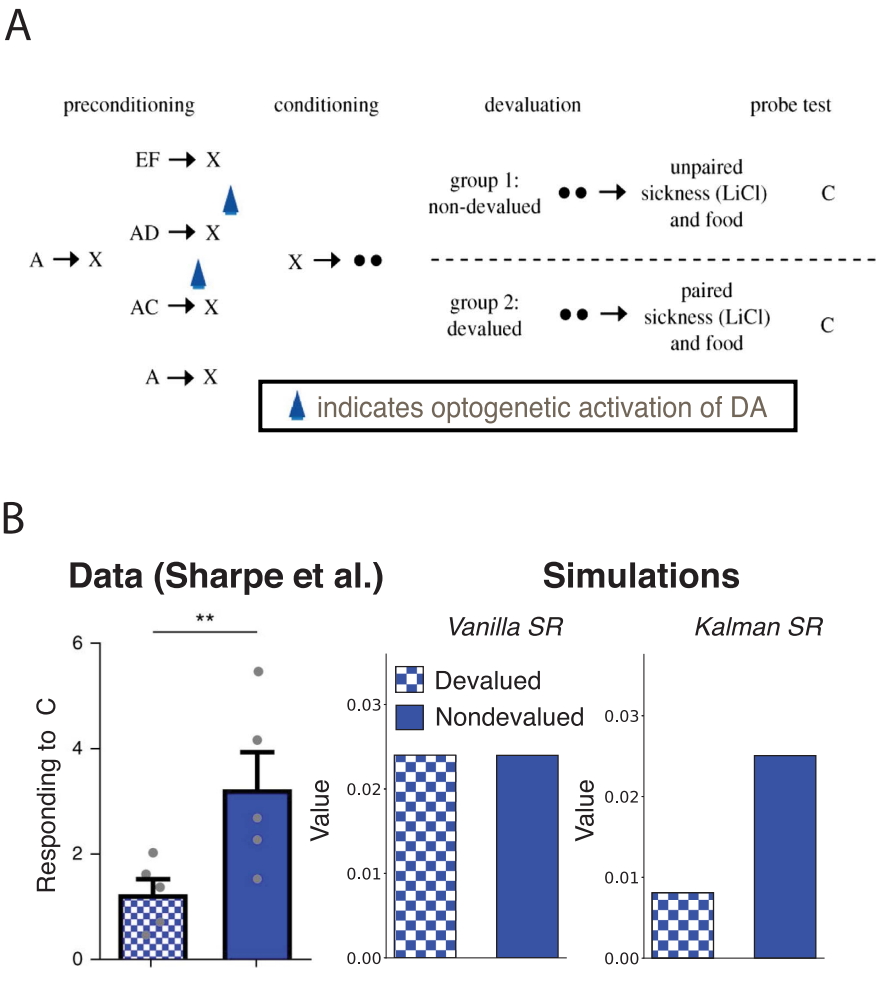
Dopamine-Dependent Devaluation

Similarly to the results of Hart et al. (2020) discussed in the main text (Figure 5), our model captures the findings of Sharpe et al. (2017), which show that the sensitivity to reward devaluation after preconditioning is dependent on dopamine transients. To show this, the authors designed an experiment similar to that of Hart et al. (2020). It started with a preconditioning phase, during which associations were learned between pairs of neutral stimuli, followed by a conditioning phase during which a neutral stimulus was paired with a food reward (Figure A1A). After this conditioning phase, the food reward was devalued in one group but not another. To show that learning the associations between neutral stimuli was mediated by dopamine, there were two preconditioning phases, during which the authors applied a "blocking" (Kamin, 1969) design. They established that the preconditioned associations where only learned when learning was unblocked by optogenetically stimulating dopamine neurons during learning. A key finding of this experiment was that animals' responding to the unblocked preconditioned cues (C) was sensitive to the subsequent devaluation of the food reward (Figure A1B, see also Hart et al., 2020). Thus, reward devaluation can alter stimulus–stimulus associations that were learned through the activation of dopamine neurons.

Gardner et al. (2018) simulated this experiment using an SR model and found that, while the SR accommodates many results found by Sharpe et al. (2017), in this particular experiment, a standard SR agent is not sensitive to the reward devaluation (Figure A1B). As in Figure 5, this is because in the TD SR, only stimuli that directly predict reward will change value after devaluation, and C was never directly associated with food. Gardner et al. (2018) therefore simulated the task with an SR model endowed with the ability to simulate offline experience, which allowed the model to be sensitive to devaluation. As with the Hart et al.'s (2020) experiment, Kalman SR is sensitive to this devaluation paradigm without the need for adding an offline replay mechanism.

(*Appendices continue*)

Figure A1
Devaluation Experiment by Sharpe et al. (2017)



Note. (A) Experimental design. During the initial preconditioning phase, one neutral stimulus always precedes a second ($A \rightarrow X$), after which the same preceding stimulus is compounded with a second predictor stimulus to precede the predicted stimulus (e.g., $AC \rightarrow X$). The initial $A \rightarrow X$ pairing blocks learning about the second association, but this blocking is prevented by optogenetic activation of dopamine neurons during learning (Sharpe et al., 2017). After the conditioning phase, the food is paired with lithium chloride (LiCl) to induce sickness. Letters denote different neutral stimuli, black circles indicate food reward. (B) Data and simulation results show that, like animals and unlike TD SR, Kalman SR shows sensitivity to devaluation in this paradigm. TD = temporal difference; SR = successor representation; DA = dopamine. See the online article for the color version of this figure.
** $p < .05$.

(*Appendices continue*)

Appendix B

The White Noise Assumption of Kalman TD

The basic version of the Kalman TD algorithm introduced in the main article was derived based on the simplifying assumption that the observation noise is white (independent per time step). In reality, this is only the case when the transitions are deterministic. In that deterministic case, optimal update of the weights can be derived, resulting in the Kalman TD algorithm used in the article (Geist & Pietquin, 2010). In most cases, however, the successive uncertainty terms cannot be treated as independent because they are related by the way in which the agent moves through the world. When the transitions are stochastic, the expectation over successor states given the current state (and action, if applicable) must be considered. When this is not done, and the original Kalman TD cost function is applied to tracking a state value function V, it can be analytically shown that this leads to the following bias (Geist & Pietquin, 2010):

$$\text{bias} = ||\kappa_t||^2 \mathbb{E}[cov_{s'|s_t,a_t}(r_t - g_t(\mathbf{w}))|r_{1:t-1}], \tag{B1}$$

with $g_t(w_t) = \hat{V}_{w_t}(s_t) - \gamma \hat{V}_{w_t}(s_{t+1})$, and κ denoting the Kalman gain.

This bias is inherent in applying Kalman TD in a stochastic setting to making predictions about any kind of cumulant, of which a reward function is only one example. Therefore, the same issue arises when estimating the SR, but for simplicity, we discuss the value-tracking case here. We discuss the issue of bias as well as a solution with an alternative noise model briefly here. For a more extensive discussion including derivations of the bias and the alternative noise model, we refer the reader to Geerts et al. (2019) and Geist and Pietquin (2010).

To alleviate the issue of bias in Kalman TD, Geist and Pietquin (2010) introduced a colored noise model that was first introduced by Engel et al. (2005) in Gaussian-process TD. The key idea is to replace the white observation noise in the generative model by a "colored" observation noise, that is, a noise that is not independent per time step. As shown by Geist and Pietquin (2010), this involves extending the parameter vector $\mathbf{w}$ to include the observation noise, such that the observation noise will be estimated from data in the inference process. The computational complexity of the resulting algorithm, extended Kalman TD (XKTD), is the same as for the original Kalman TD because the parameter vector is extended with two scalars. However, this colored noise estimation induces some memory effects which means that XKTD cannot be applied to off-policy evaluation.

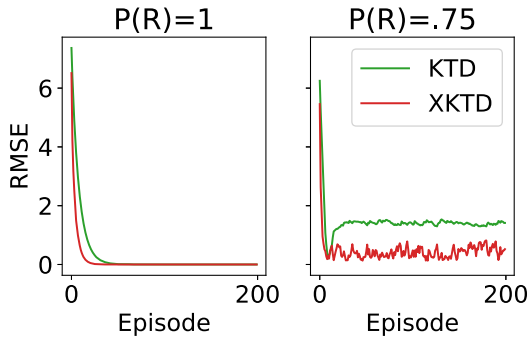
In order to empirically asses how damaging the white noise assumption is, we now compare KTD's value estimates to the true (unbiased) value, approximated by Monte Carlo sampling. In Monte Carlo sampling, value is estimated by simply averaging sample returns across episodes (Sutton & Barto, 1998). For completeness, we also compare these to XKTD estimates.

We evaluated both algorithms on a simple chain Markov Decision Process (adapted from Brockman et al., 2016). The Markov Decision Process has seven nonabsorbing states, arranged linearly from State 0 to State 6. Making a right move in the final state leads to an absorbing state. The agent can move to the left or right and receives a reward of −0.2 for every step except in the absorbing state, where it receives a reward of 10. The stochasticity in the state transitions will come from the policy, which can be defined by a single parameter $P(R)$, for the probability of making a step to the right, ($P(L)$=1 $P(R)$).

We ran KTD and XKTD on this domain for 200 episodes, with a deterministic optimal policy, ($P(R) = 1$), and with a stochastic policy, ($P(R) = 0.75$), computing after each episode the root-mean-square error (RMSE) between the true value function (as estimated by Monte Carlo) and the algorithm's value estimate (Figure B1). With

Figure B1

Root-Mean-Square Error After Each Episode for an Example Run of KTD and XKTD in a Deterministic (Left) and Stochastic (Right) MDP



Note. MDP = Markov Decision Process; KTD = Kalman temporal differences; XKTD = extended Kalman temporal difference. See the online article for the color version of this figure.

(*Appendices continue*)

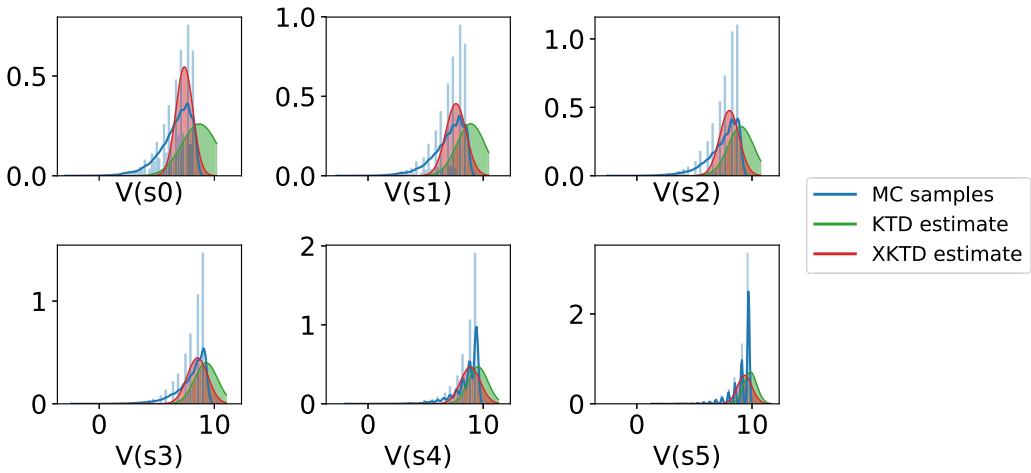
deterministic transitions, both algorithms converge to the same low error (left panel), but with stochastic transitions, KTD converges to a wrong value, maintaining higher error, consistent with the analytically derived bias (Equation B1). Figure B2 shows the posterior distribution over value after the example run of 200 episodes for both KTD and XKTD, overlaid with the actual, sampled returns. Indeed, the mean of the posterior for KTD is consistently off, while the XKTD posterior is closer to the true mean.

To quantify how damaging the deviations are as a function of stochasticity of the environment, we then varied $P(R)$ from 1 to 0.5 (completely random transitions), running KTD and XKTD for 200 episodes, repeated this 10 times for each value of $P(R)$ and computed the RMSE, which is shown in Figure B3. As could be seen theoretically, the bias grows as the environment is more stochastic. In addition, the bias is significantly reduced for XKTD, although even for the latter algorithm, the bias grows with higher stochasticity.

Thus, the Kalman TD algorithm incorrectly treats successive observation noise terms as "white" (independent from each other), while they are related because of the way the agent moves through the world. Any stochasticity in the transitions will therefore bring about a bias in the KTD estimates. This problem can be alleviated using XKTD, in which the hidden parameter vector is extended such that the observation noise, which is now assumed to be colored, can be estimated online. However, as shown in Figure B3, even XKTD

Figure B2

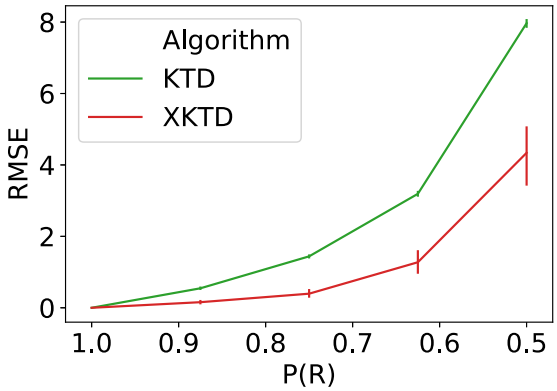
Posterior Distributions Over Value After an Example Run of 200 Episodes for KTD and XKTD, Overlaid With the True, Sampled Distribution of Returns



Note. MC = Monte Carlo; KTD = Kalman temporal differences; XKTD = extended Kalman temporal difference. See the online article for the color version of this figure.

(*Appendices continue*)

Figure B3
Comparing KTD and XKTD on the Linear Track Environment



Note. RMSE after 200 episodes is plotted as a function of stochasticity of the environment, $P(R) = 0.5$ corresponds to maximum entropy/randomness. Error bars show 95% confidence intervals. KTD = Kalman temporal differences; XKTD = extended Kalman temporal difference; RMSE = root-mean-square error. See the online article for the color version of this figure.

leads to biased estimates under high stochasticity. This is because, while the assumptions are less strong than for KTD, XKTD still incorrectly assumes that the successive residuals are independent from each other.

In conclusion, KTD's uncertainty estimation is incorrect for many realistic Markov Decision Processes, but this can be remedied by extending KTD with colored noise estimation, without adding significant computational or memory complexity.

Received March 7, 2022
Revision received September 29, 2022
Accepted December 2, 2022 ■