UNIVERSITY OF CALIFORNIA,
IRVINE

Learning to Control Schedules of Reinforcement

DISSERTATION

submitted in partial satisfaction of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

in Cognitive Science

by

Mac Strelioff

Dissertation Committee:
Mimi Liljeholm, Chair
Sean Ostlund
Joachim Vandekerckhove

2019

Contents

List of Figures

ACKNOWLEDGMENTS

The research presented here would not have been possible without the guidance, supervision, and contributions of Dr. Mimi Liljeholm. I would also like to think Dr. Liljeholm, Dr. Lee, and the Associate Dean Fellowship for financial support; and the research assistants who were involved in the projects described below, namely: Stephanie Ponce, Brenda Vasquez, Samia Temueri, and Erick Garcia.

CURRICULUM VITAE

Mac Strelioff

EDUCATION

Doctor of Philosophy in Cognitive Science **2019**
University of California, Irvine *Irvine, CA*

Masters of Science in Statistics **2018**
University of California, Irvine *Irvine, CA*

Masters of Science in Cognitive Neuroscience **2016**
University of California, Irvine *Irvine, CA*

Bachelor of Science in Psychology **2014**
University of California, Davis *Davis, CA*

TEACHING EXPERIENCE

Teaching Assistant **2014–2019**
University of California, Irvine *Irvine, CA*

ABSTRACT OF THE DISSERTATION

Learning to Control Schedules of Reinforcement

By

Mac Strelioff

Doctor of Philosophy in Cognitive Science

University of California, Irvine, 2019

Mimi Liljeholm, Chair

An ability to quickly learn about relationships between actions and outcomes is essential for adaptive behavior. Such learning can be complicated when the action-outcome relationship depends on the latency with which the action is performed. Inferences about such latency-modified contingencies can greatly improve an agent's performance by allowing for a timing of responses that optimizes the probability of an outcome given an action. Here we specify a Bayesian reinforcement learner that infers the functional and causal form of the relationship between response latencies and response contingent outcome probabilities. The performance of this Bayesian learner is contrasted with that of a model-free actor-critic algorithm, using behavioral data from three latency-modified schedules of reinforcement. Results suggest that a model-based characterization of latency-modified schedules provides a superior account of free operant behavior. Next the test of latency-modified schedules is extended to multi-action contexts where the latency of performing one action modifies the probability of an outcome given performance of a different action. Participants quickly discovered optimal rates of responding on both the modified and modifying actions, even when rewards contingent on performing the modifying action provided local incentives for deviating from the optimal rate. A final set of experiments tested the hypothesized explicit inference about latency-modified schedules. In summary, these models and experiments highlight the advantages of incorporating latency into a causal structure of actions and goals.

Chapter 1

Introduction

Many everyday activities, such as cooking a new recipe, mastering a new musical instrument, playing a new video game, or driving a new car, can require careful use of one action to modify the effects of another. For example, holding strings on the neck of a guitar does nothing by itself, but profoundly modifies the sounds produces by strumming. The primary goal of this work was to develop and validate a new model of instrumental learning, particularly in free operant contexts with action-modified contingencies. Although our approach was designed to address complex instrumental behavior, it is also applicable to simpler contingencies defined over a single action that can be cast as action-modified contingencies.

1.1 Schedules of Reinforcement: Behavior and Theory

Reward contingencies, or schedules of reinforcement, are functions that relate features of actions and the environment to outcomes. Contingencies involving only one action offer simple methods for investigating behavior. The most common of these are: ratio schedules, where delivery of reward depends on the number of actions performed; interval schedules,

where reward delivery depends on the time elapsed since the last rewarded action; and differential rate schedules, where reward delivery depends on latencies between actions. These qualitatively distinct schedules produce different response profiles, with interval schedules generating much lower levels of responding (Ferster & Skinner, 1957). Over the last century, many theoretical accounts have been proposed to address the differential influence of reward contingencies on behavior but none acknowledge the functional dependence between response contingent reward probability and other features of behavior imposed by different schedules.

One framework proposed to explain differences in response profiles associated with interval and ratio schedules was inter-response time (IRT) theory (Reynolds & McLeod, 1970; Alleman & Platt, 1973). IRT theory is based on the observation that responses typically occur in bursts such that there are many responses with short IRTs within a burst and few responses with long IRTs between bursts. On ratio schedules where reward ratios match response ratios, short, within-burst, IRTs are frequently performed and followed by reward. Hence the propensity for short IRTs is strengthened by ratio schedules. Alternatively, on interval schedules where response contingent reward probabilities increase with longer intervals between rewards, rewards are more likely to follow long, between-burst, IRTs. Hence the propensity for long IRTs is strengthened by interval schedules. Recent research with animals (Tanno, Silberberg, & Sakagami, 2009, 2012) and humans (Silberberg, Goto, Hachiga, & Tanno, 2008) has attributed the discriminability of ratio and interval schedules to the differential reinforcement of short and long IRTs, without a theoretical account of why IRTs would be related to an ability to explicitly differentiate between schedules. This framework alone neglects any cognitive representation of different schedules.

At a cognitive level, people may use perceived casual relationships to determine schedule structures. Perceived causal effectiveness, and response rates, increased with the difference between the probability of a reward given an action and the probability of a reward given no

action (Allan, 1980, 1993; Liljeholm, Tricomi, O'Doherty, & Balleine, 2011; Tanaka, Balleine, & O'Doherty, 2008); that is, with $\Delta P = P(r|a) - P(r|\ a)$. However, perceived causal effectiveness and response rates also increased when short IRTs were selectively reinforced (Reed, 2001, 2003). Hence it is unclear whether higher response rates are attributable to local mechanisms, such as rewarded IRTs, or global statistics, such as ΔP. The model developed in section 2.3 formalizes a relationship between local variables and global statistics where locally observed IRTs and rewards are used to infer functions relating IRTs to the probability of a reward given a response and infer whether rewards are more likely to be generated by the inferred function of IRTs (as on an interval schedule) compared to constant function of IRTs (as is specified by a ratio schedule).

1.2 Reinforcement Learning Theories

Similar distinctions between directly enhancing the propensity for a particular response and developing a cognitive representation of global statistics permeate modern reinforcement learning algorithms (Sutton & Barto, 1998). A central goal of reinforcement learning methods is to accurately estimate expected reward for each action in each of a set of possible states. These state-action values, originally referred to as action qualities by Sutton and Barto (1998) or Q-values, are often represented with a matrix, or look-up table, with a row for each action and a column for each state. Common model-free methods (Watkins, 1989) use reward prediction errors to associate observed rewards with the actions and states that lead to them, and a learning rate parameter to control the relative weight of more recent experiences in these computations. In contrast to the model-free approach, a model-based approach (Doya, Samejima, Katagiri, & Kawato, 2002) would maintain representations of the probability of transitioning into a possible subsequent state given that a particular action is performed in the current state, and the reward associated with that subsequent state.

Transition probabilities and rewards are dynamically combined, at each choice point, to yield action qualities. Both model-free and model-based algorithms obtain a policy, or probabilistic mapping from states to actions, as a function of action qualities. Common policies include a greedy policy where the action with the highest Q-value is chosen, or a softmax function where actions with higher Q-values are chosen more often. While these methods work for trial-based tasks with discrete state and action spaces, free operant behavior is complicated by continuous action spaces (because a learner can choose any latency) and potentially a continuous state space (e.g. on interval schedules, the state depends on continuous time). A model-free account of free operant behavior, discussed in detail in section 2.2, has been developed by Niv (2007). Currently, no model-based accounts of free operant behavior exist in the literature. Model-based algorithms afford a great deal of flexibility, including sensitivity to outcome devaluation and contingency degradation (Dickinson, 1985). Yet, more complex environments may require learners to infer causal structures that explicitly represent the latent, dynamic functions that produce rewards, as discussed below. One novel contribution of the current work is the development of a model-based account of free operant behavior, described in section 2.3, that is based on inferring the functional and causal structure of arbitrarily complex latency-dependent contingencies.

1.3 Reinforcement Schedules as Causal Structures

Structure learning refers to the acquisition of a model that relates stimuli, actions, or outcomes using more abstract representations. Reversal learning tasks and bandit problems have been used to investigate structure learning. Reversal learning tasks generally include a discrimination phase and reversal phases. In the discrimination phase, one stimulus, or action, is paired with reward (S_1+) and another with punishment or loss (S_2-). In the reversal phase, the contingencies reverse, such that the stimulus previously paired with reward

signals a punishment or loss (S_1-) and the stimulus previously paired with punishment or loss signals a reward (S_2+). Because multiple reversals occur throughout a reversal learning task, a participant must continually judge whether or not a reversal has occurred and update their mental representation of the schedule accordingly. Reversal errors occur when a participant is punished for making a response that was previously correct but has since become incorrect due to a reversal. Participants tend to shift responding after 2 to 3 reversal errors (Cools, Clark, Owen, & Robbins, 2002). Shifts in responding may involve the ventral striatum (Cools et al., 2002), orbitofrontal cortex (O'Doherty, Critchley, Deichmann, & Dolan, 2003; Remijnse, Nielen, Uylings, & Veltman, 2005), anterior cingulate (O'Doherty et al., 2003; Remijnse et al., 2005; O'Doherty, Buchanan, Seymour, & Dolan, 2006), anterior insula (O'Doherty et al., 2006), and prefrontal cortex (Cools et al., 2002; Remijnse et al., 2005). Thus, these are regions of interest for future investigations of behavior in an environment where the consequence of one response provides information about the schedule over another. Hampton, Bossaerts, and O'Doherty (2006) developed a Bayesian hidden state Markov model that accounted for the hidden structure of the forced choice reversal learning task. This model predicted human neural activity better than reinforcement learning models that did not account for the reversal structure. Specifically, activity in the medial prefrontal cortex reflected increases in expected value of the alternative response that were inferred from the reversal structure. Thus, the prefrontal cortex is likely involved in modeling abstract stimulus-outcome contingency structures (Hampton et al., 2006; O'Doherty, Hampton, & Kim, 2007) and may also be involved in representing abstract relationships necessary for learning about action-modified structures.

Structure learning, albeit in a simplistic form, has also been studied in bandit problems with Gaussian (Reverdy, Srivastava, & Leonard, 2014; Speekenbrink & Konstantinidis, 2015) and Bernoulli (Acuna & Schrater, 2009; Steyvers, Lee, & Wagenmakers, 2009; Lee, Zhang, Munro, & Steyvers, 2011; Yi, Steyvers, & Lee, 2009; Speekenbrink & Konstantinidis, 2015) reward distributions. Bandit problems with Gaussian reward distributions are concurrent ra-

tio schedules with reward magnitudes distributed as Gaussian random variables. Bernoulli bandit problems are concurrent variable ratio schedules. Human behavior reflects online learning of nonstationary reward magnitudes in Gaussian bandit problems (Speekenbrink & Konstantinidis, 2015) and nonstationary reward probabilities in Bernoulli bandit problems (Yi et al., 2009). This prior research has examined structures where reward probabilities depend on latent events (e.g. reversals) or drift stochastically across time (e.g. in Bernoulli bandit problems), but has not considered cases where actions influence the relationship between responding and outcomes. The approach developed under Aim 1 specifies a computational mechanism for relating latencies to the relationship between an action and its outcome. Summary and conclusion

1.4 Summary

Behavioral research has implicated numerous variables, including reinforced IRTs (Reynolds & McLeod, 1970; Alleman & Platt, 1973) and delta-P (Liljeholm et al., 2011), as influencers of response rate. These variables in particular are also likely to be involved in the process of inferring schedule structures (Reed, 2001, 2003; Silberberg et al., 2008; Tanno et al., 2009, 2012). Various knowledge, specified by reinforcement learning algorithms (Sutton & Barto, 1998), is necessary for behavior that maximizes reward, and various neural substrates have been associated with the computational mechanisms employed by these algorithms (O'Doherty et al., 2003, 2006, 2007). These findings encourage further exploration of reinforcement learning algorithms as cognitive models of instrumental behavior. In dynamic environments, human behavior reflects online tracking of reward probabilities (Lee et al., 2011; Steyvers et al., 2009; Yi et al., 2009) and probabilities of reversals between latent structures (Hampton et al., 2006). This project unifies and extends this work by specifying a model for learning about arbitrarily complex latency-modified contingencies, demonstrat-

ing the ability of this model to approximate functional forms and infer causal structures, and developing an experimental paradigm to arbitrate between model-based and model-free control of behavior on latency-modified schedules.

Part I

Model Derivations

Chapter 2

A Structure Inference Account of Free Operant Behavior

2.1 Semi-Markov Decision Processes

Here we use a Markov Decision Processes (MDP) framework to model an agent in a free operant environment. We generally use notation from Sutton and Barto (1998), adapted to the context of free operant behavior. Some exceptions, described in sections below, were made to be consistent with the notation of other authors who influenced the work developed here. A visual summary of the key components of an MDP using the notation described here is shown in Figure 2.1.

An MDP requires a set of all possible states $\mathcal{S}$, a set of all possible actions $\mathcal{A}$, and a set of all possible reward values $\mathcal{R}$. At each timestep t, an agent is in one state $s_t \in \mathcal{S}$, chooses

an action $a_t \in \mathcal{A}$, obtains reward[1] $r_t \in \mathcal{R}$, and then transitions to a new state[2] $s_{t+1} \in \mathcal{S}$. A reward function[3] $p(r_t = r|s_t, a_t)$ specifies the probability of obtaining reward r after taking action a_t in state s_t, and a transition function $p(s_{t+1} = s|s_t, a_t)$ specifies the probability of arriving in state s after taking action a_t in state s_t. The final component is a distribution over actions commonly referred to as a policy[4], $p(a|s_t)$.

A semi-Markov Decision Process (sMDP) extends MDPs to these continuous settings by introducing a *dwell time* spent within each episode. In free operant settings, actions a_t are a pair of the discrete response alternative a_i and a positive real valued inter-response interval $\tau \in (0, \infty)$ that determines the dwell time, that is; $a_t = (a_i, \tau_t)$ with a_i indicating which action was taken and τ_t representing the time since the last response.

As an example, let's assume an experimenter chooses to administer a random interval (RIλ) schedule. Interval schedules queue rewards to be obtained by the next response. A RIλ schedule is defined so that at each point in time there is a constant probability of a reward being queued so long as the average interval between the last rewarded response and the time at which a reward becomes queued is the value λ chosen by the experimenter. Here the action space $\mathcal{A}$ would be the single action that the schedule is defined over, a_1, and the latency with which this action is taken τ. The state space $\mathcal{S}$ would consist of the time that has elapsed since the last response, and the transition function would deterministically track time since the last response. The reward space $\mathcal{R}$ could consist of a 1 to indicate that a reward had

[1]While Sutton and Barto (1998) end an episode with the action a_t and then deliver rewards at the start of the next episode – so that a_t is paired with r_{t+1} – Niv (2007) begins the episode with the choice of action a_t and the observation of reward. We adopt the structure and notation of Niv (2007) and hence pair a_t with r_t rather than with r_{t+1}.

[2]Sutton and Barto (1998) use notation s and s' to denote current and future states. Here we use s_t and s_{t+1}, which is both more explicit and consistent with notation used by Niv (2007).

[3]Whereas Sutton and Barto (1998) define a transition function as the joint distribution $p(s_{t+1} = s, r_t = r|s_t, a_t)$, Niv (2007, pp. 46) separately defines a marginal transition function $p(s_{t+1} = s|s_t, a_t)$ and a marginal reward function $p(s_{t+1} = s|s_t, a_t)$. The definitions used by Niv (2007) are easily applied in free operant contexts because schedules of reinforcement define the reward function.

[4]Sutton and Barto (1998) use π, Niv (2007) uses $p(a, \tau|s)$. I use notation from Niv (2007), but simplify it by defining the action to be the pair of the action and latency, $a = (a_i, \tau)$, so that the policy can be compactly presented as $p(a|s_t)$

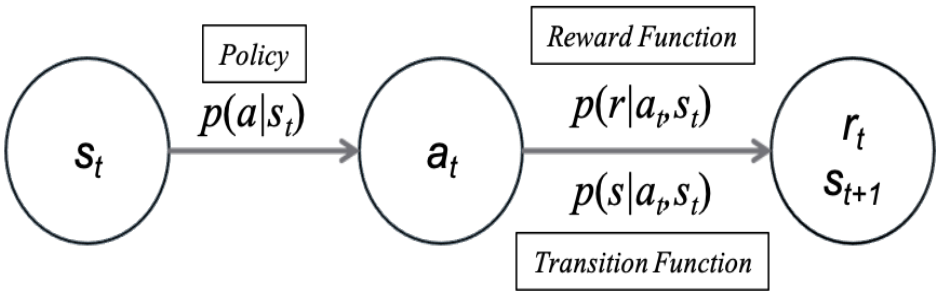

Figure 2.1: Graphical depiction one episode of a Markov Decision Process. Here an agent is in state s_t at time t, they draw an action a_t from the policy $p(a|s_t)$, and then observe reward r_t drawn from the reward function $p(r|a_t, s_t)$ and transition to state s_{t+1} drawn from the transition function $p(s|a_t, s_t)$.

been queued and a 0 to indicate otherwise. The RIλ schedule implies that the probability of a reward having been queued since the last response[5] is the exponential cumulative density function with scale parameter λ, hence the reward function is; $p(r_t = 1|a_1, \tau_t) = 1 - e^{-\frac{\tau_t}{\lambda}}$, where τ_t represents the latency with which the response was made which is the time since the last response in a single action setting.

We assume throughout that the agent's goal is to learn a policy that maximizes reward rates[6]. The agent can pursue this goal in a model-free or model-based way. In the remaining sections I first describe the model-free approach taken by Niv (2007), and then the novel model-based approach developed here.

2.2 Niv's (2007) Model-Free Account

Model-free methods learn a policy from reward prediction errors, without explicitly representing the reward or transition functions. One such method, used by Niv (2007), is an

[5]In contrast to variable interval schedules that depend on the time that has elapsed since the last rewarded response, since the process specified by a random interval schedule is memoryless, response contingent reward probabilities on random interval schedules depend only on the time since the last response rather than the time since the last rewarded response.

[6]A conventional goal in trial-based settings is to maximize total expected discounted future rewards (Sutton & Barto, 1998), however maximizing reward rates is a better objective in free operant settings where response rates can influence dwell times (Niv, 2007; Daw, 2003).

Actor-Critic architecture.

2.2.1 Setup and Overview

Niv (2007) considered a rat in a Skinner box with that could perform three possible actions; lever press (LP), nose poke (NP), or other. Hence the rat can perform actions $a_i \in \{LP, NP, Other\}$, where $i \in \{1, 2, 3\}$ indexes these discrete actions. The rat also chose the latency or duration before taking the action, $\tau_t \in (0, \infty)$. The action space consisted of pairs of a discrete action a_i and the latency τ_t after which that action was performed. A reward with value U_r could be queued by the LP action and obtained if a rat chose the NP action when a reward was queued. The state space contained an indicator for whether reward was queued ($I_{rew} = 1$) or not ($I_{rew} = 0$), as well as other features relevant to the reward function. The reward function governed when a LP action would queue reward and was specified to represent either a random ratio or random interval schedule of reinforcement.

In an actor critic model, the agent is comprised of two modules; a critic and an actor. The critic module uses reward signals to compute prediction errors that are used to estimate state values and, in this context, reward rates. The critic passes prediction errors to the actor module which uses them to update the policy.

2.2.2 The Critic

The reward signal used by Niv (2007) depended on three parameters; the utility of reward U_r, the response cost C_u, and a vigor cost C_v, and was defined as the net reward from the previous action, $U_r - C_u$, less a term inversely proportional to latencies $\frac{C_v}{\tau_t}$;

$$r_t = U_r - C_u - \frac{C_v}{\tau_t}$$

The critic used the reward signal and averaging rate $\eta_{\bar{R}}$ to update the estimated reward rate as[7];

$$\bar{R}_{t+1} \leftarrow (1-\eta_{\bar{R}})^{\tau_t}\bar{R}_t + \eta_{\bar{R}} r_t$$

The critic also used the reward signal along with the estimate of the reward rate $\bar{R}_t$ and state values, $v(s)$ to compute prediction errors as[8];

$$\delta_t = r_t - \bar{R}_t \int_0^{\tau_t} e^{-t\eta_{\bar{R}}} dt + v_t(s_{t+1}) - v_t(s_t)$$

These prediction errors were used by the Critic to update state values using;

$$v_{t+1}(s_t) \leftarrow v_t(s_t) + \eta_v \delta_t$$

where η_v is a learning rate. Prediction errors were also passed to the Actor module for use in policy updates.

2.2.3 The Actor

Niv (2007, pp. 112) chooses a mixture of gamma distributions as the policy;

[7]This rule is the extension of the discrete update rules, $\bar{R}_{t+1} \leftarrow \bar{R}_t + \eta_{\bar{R}}(r_t - \bar{R}_t)$, to continuous time. For application, Niv (2007) noted the equality $(1-\eta_{\bar{R}})^{\tau_t}\bar{R}_t + \eta_{\bar{R}} r_t = e^{log(1-\eta_{\bar{R}})\tau_t}\bar{R}_t + \eta_{\bar{R}} r_t$ and then used the approximation $e^{log(1-\eta_{\bar{R}})\tau_t}\bar{R}_t + \eta_{\bar{R}} r_t \approx e^{-\eta_{\bar{R}}\tau_t}\bar{R}_t + \eta_{\bar{R}} r_t$. This approximation works for small values of $\eta_{\bar{R}}$, however, small values were not obtained when fitting this model to behavioral data. Because of this, when fitting to data I used the form: $e^{log(1-\eta_{\bar{R}})\tau_t}\bar{R}_t + \eta_{\bar{R}} r_t$

[8]In application I used the closed form solution to the integral, resulting in; $\delta_t = r_t - \bar{R}_t \frac{(1-e^{-\eta_{\bar{R}}\tau_t})}{\eta_{\bar{R}}} + v_t(s_{t+1}) - v_t(s_t)$

$$p(a_t, \tau_t | s_t, \Theta) = \frac{m_i}{\Sigma_j m_j} \frac{\tau^{\alpha_i - 1} e^{-\frac{\tau}{\theta_i}}}{\theta_i^{\alpha_i} \Gamma(\alpha_i)}$$

Here $\Theta = \{\vec{m}, \vec{\alpha}, \vec{\theta}\}$ represents all policy parameters. The mixing proportions are determined by m_i and represent the probability of choosing action i. The latency for each action is gamma distributed with action-specific shape (α_i) and scale (θ_i) parameters.

In deriving the update rules for the actor, Niv (2007, pp. 112) specifies the algorithm with conventional update rules then later described modifications for numerical stability. The unmodified update rules, derived from gradient ascent, are;

$$\begin{aligned}
m_j &\leftarrow m_j + \eta_m \delta_t \frac{1}{m_j} \left(I_{a_t = a_j} - \frac{m_j}{\Sigma_k m_j} \right), \forall j \\
\alpha_i &\leftarrow \alpha_i + \eta_\alpha \delta_t \frac{\delta p(\tau_t | a_i, s_t)}{\delta \alpha_i} = \alpha_i + \eta_\alpha \delta_t \left(log \left(\frac{\tau}{\theta_i} \right) - \frac{\Gamma'(\alpha_i)}{\Gamma(\alpha_i)} \right) \\
\theta_i &\leftarrow \theta_i + \eta_\theta \delta_t \frac{\delta p(\tau_t | a_i, s_t)}{\delta \theta_i} = \theta_i + \eta_\theta \delta_t \left(\frac{\tau - \alpha_i \theta_i}{\theta_i^2} \right)
\end{aligned}$$

Where $I_{a_i = a_j}$ is an indicator function that takes the value 1 if the current action is the same as the previous action ($a_i = a_j$), and 0 otherwise. To incorporate reward rates into the maximum of the latency distributions Niv (2007, pp.115-116) defined $\hat{\theta}_i = \theta_i \sqrt{\overline{R}}$, then derived the update rules based on $\hat{\theta}_i$;

$$\begin{aligned}
\alpha_i &\leftarrow \alpha_i + \eta_\alpha \delta_t \left(log \left(\frac{\tau \sqrt{\overline{R}}}{\hat{\theta}_i} \right) - \frac{\Gamma'(\alpha_i)}{\Gamma(\alpha_i)} \right) \\
\hat{\theta}_i &\leftarrow \hat{\theta}_i + \eta_\theta \delta_t \left(\frac{\tau \sqrt{\overline{R}} - \alpha_i \hat{\theta}_i}{\hat{\theta}_i^2} \right)
\end{aligned}$$

Niv (2007, pp. 116-117) also introduced transformations so that the policy parameters would be on a constrained space that ensured the gamma distributions would be defined and first increasing then decreasing. Specifically, Niv defined $m := e^{\tilde{m}} \geq 0$, $\alpha := e^{\tilde{\alpha}} + 1 < 1$, and $\hat{\theta} := e^{\tilde{\theta}} \geq 0$. Finally, to avoid numerical instabilities, Niv (2007, pp. 117) forced $\bar{R}$ to be no smaller than 0.1, and scaled updates to m by $e^{-\tilde{m}}$, and updates to $\tilde{\alpha}$ and $\tilde{\theta}$ by $e^{-|\tilde{\alpha}|-|\tilde{\theta}|}$. The update rules below were derived from these constraints and were used when applying this model;

$$\tilde{m}_j \leftarrow \tilde{m}_j + e^{-\tilde{m}} \eta_m \delta_t \left(I_{a_t=a_j} - \frac{e^{\tilde{m}_j}}{\sum_k e^{\tilde{m}_k}} \right)$$
$$\tilde{\alpha}_i \leftarrow \tilde{\alpha}_i + e^{-|\tilde{\alpha}|-|\tilde{\theta}|} \eta_\alpha \delta_t \left(log\left(\frac{\tau\sqrt{\bar{R}}}{e^{\tilde{\theta}_i}} \right) - \frac{\Gamma'(e^{\tilde{\alpha}_i}+1)}{\Gamma(e^{\tilde{\alpha}_i}+1)} \right) e^{\tilde{\alpha}_i}$$
$$\tilde{\theta}_i \leftarrow \tilde{\theta}_i + e^{-|\tilde{\alpha}|-|\tilde{\theta}|} \eta_\theta \delta_t \left(\frac{\tau\sqrt{\bar{R}}}{e^{\hat{\theta}_i}} - e^{\tilde{\alpha}_i} - 1 \right)$$

Then actions are based on $\Theta = \{\vec{m} = e^{\tilde{m}}, \vec{\alpha} = e^{\tilde{\alpha}} + 1, \vec{\theta} = \frac{e^{\tilde{\theta}}}{\sqrt{\bar{R}}}\}$.

For numerical stability in simulations, Niv (2007, pp. 117) further constrained the parameters of the gamma distributions so that the shape parameter was greater than 1.05 and the scale parameter was greater than 0.05, and chose $\eta_{\bar{R}} = 0.0005$, $\eta_v = 0.05$, and $\eta_m = \eta_\alpha = \eta_\theta = 0.01$. These additional constraints were not used when fitting the model to data in later sections.

2.2.4 Free Parameters

When more than one action is available, this model has at least 4 free parameters; $\eta_{\bar{R}}$, η_m, η_α, and η_θ. In the single action context updates for m_j are eliminated[9]. This reduces the number of parameters by one since η_m is no longer estimable nor necessary. Fully parameterized,

[9]This is because $I_{a_t=a_1} = 1 = \frac{m_1}{m_1} = \frac{m_j}{\sum_j m_j} \Rightarrow I_{a_t=a_j} - \frac{m_j}{\Sigma_k m_j} = 0$

this model has additional parameters for C_u and C_v that scale with the number of states and actions considered. When implemented here, C_u was defined based on the experimental contingency. This left 5 free parameters; $\eta_{\bar{R}}$, η_m, η_α, η_θ, and C_v.

2.3 A Bayesian Structure Inference Account

Model-based methods directly represent the reward and transition functions, then construct a policy based on these representations. We desired an approach that enabled the agent to quickly differentiate between reward functions where the response-contingent reward probability was a constant and reward functions where the response-contingent reward probability depended on latencies of either the same or a different action. The approach also had to make this arbitration quickly in free operant settings, and use this knowledge to formulate an optimal policy.

2.3.1 Setup

We assume that the agent represents possible reward functions with a set of candidate models m_l in a model space $\mathcal{M} = \{m_1, ...m_L\}$, where L is the number of candidate models or structures. Models here represent possible conditional dependencies between reward probabilities and features of an action or the environment. A simple example for a case where reward probabilities are a constant (m_1) or a function of latencies (m_2) is displayed graphically in Figure 2.2. The models considered here are intended to be agnostic with respect to the functional form. A separate function approximation method is used to estimate functional forms of the dependencies within each model. Obtaining model probabilities involved integrating over parameters of the function approximation method, as described in subsection 2.3.3.

We represent reward with an indicator r_t that takes value 1 if an action is rewarded at time

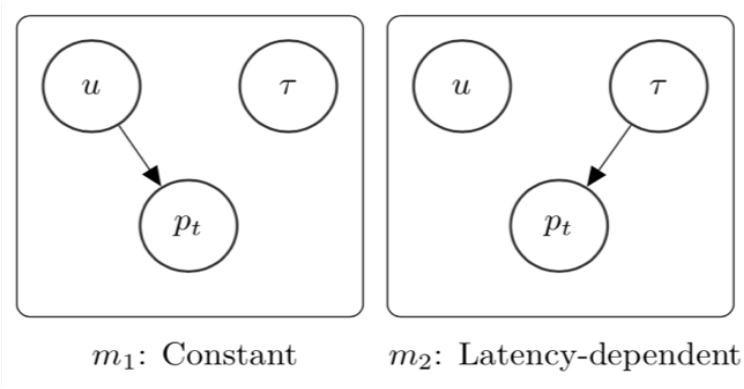

Figure 2.2: Graphical representations of possible models. Left: A model where response contingent reward probability (p_t) is determined by a constant u and uninfluenced by latency τ. Right: A model where response contingent reward probability is fully determined by latencies τ.

t, and value 0 otherwise. For a given model, this reward is Bernoulli distributed with a probability $p_t^{(m)}$ that depends on the model (m) and changes over time (t) depending on the parameters of the function approximation method described below;

$$r_t|m \sim Bernoulli(p_t^{(m)})$$

To maximize reward the agent must estimate the response contingent reward probabilities under each model, $p_t^{(m)}$, as well as the probability of each model.

2.3.2 Function Approximation

We used logistic functions as the functional form of the link from model-specific feature vectors $\vec{x}_t^{(m)}$ and weights $\vec{\theta}_t^{(m)}$ to model-based beliefs about the probability of a response contingent reward, $p_t^{(m)}$;

$$p_t^{(m)} = \frac{1}{1 + e^{-(\vec{x}_t^{(m)} \vec{\theta}_t^{(m)})}}$$

For the model that depended only on a constant, $x_t^{(m)}$ was a scaler that always took the value 1. For models where response-contingent reward probabilities depended on latencies, we represented latencies with a set of Gaussian basis functions with means μ_i and a shared variance σ, which were normalized to have a maximum value of 1. An example set of Gaussian basis functions and values of $\vec{x}_t^{(m)}$ for a particular latency, are shown in the top-left subplot of Figure 2.3. Gaussian basis functions have three main advantages. First, observations would generalize to nearby latencies but not latencies far away. This is consistent with results on stimulus generalization (Boneau & Cole, 1967; Honig & Urcuioli, 1981). Second, Gaussian basis functions are commonly used in the machine learning literature (Park & Sandberg, 1991; Sutton & Barto, 1998; Geramifard et al., 2013) thus providing a normative grounding for their use here. Third, these representations enabled our model to learn arbitrarily complex functions of latencies. Examples of inferred functions based on weights drawn from a uniform distribution are shown in the top-right subplot of Figure 2.3.

Weights for each model were learned using a Bayesian procedure adapted from McCormick, Raftery, Madigan, and Burd (2012). First we assumed a normal prior with mean vector of zeros and an identity covariance matrix; $\theta_0^{(m)} \sim N(\vec{0}, I)$. After each action, we computed the posterior over these weights according to Bayes Rule;

$$p(\theta_t^{(m)}|r_t) \propto p(r_t|\theta_{t-1}^{(m)})p(\theta_{t-1}^{(m)})$$

where $p(r_t|\theta_{t-1}^{(m)})$ is the Bernoulli likelihood of the reward outcome at time t (r_t) under model m and $p(\theta_{t-1}^{(m)})$ is the Normal prior, which was itself the posterior over weights after the last observation at time $t-1$. We used Laplace approximation to quickly estimate this posterior after each outcome (Lewis & Raftery, 1997). This resulted in the following approximations

to the first and second derivatives of the log posterior;

$$\frac{\partial}{\partial\theta}\ell(\hat{\theta}_{t-1}^{(m)}) = (r_t - \hat{p}_t^{(m)})x_t^{(m)}$$
$$\frac{\partial^2}{\partial\theta^2}\ell(\hat{\theta}_{t-1}^{(m)}) = \left(\hat{\Sigma}_{t-1}^{(m)}\right)^{-1} + x_t^{(m)}\hat{p}_t^{(m)}(1-\hat{p}_t^{(m)})(x_t^{(m)})^T$$

Where $\ell(\hat{\theta}_{t-1}^{(m)})$ represents the log likelihood of $\hat{\theta}_{t-1}^{(m)}$, and $\hat{p}_t^{(m)}$ represents the predicted probability of a response-contingent reward under model m. Then the parameters of the distribution over model weights were updated according to;

$$\hat{\theta}_t^{(m)} \leftarrow \hat{\theta}_{t-1}^{(m)} - \left(\frac{\partial^2}{\partial\theta^2}\ell(\hat{\theta}_{t-1}^{(m)})\right)^{-1}\frac{\partial}{\partial\theta}\ell(\hat{\theta}_{t-1}^{(m)})$$
$$\hat{\Sigma}_t^{(m)} \leftarrow \left(-\frac{\partial^2}{\partial\theta^2}\ell(\hat{\theta}_{t-1}^{(m)})\right)^{-1}$$

which gave the posterior as $\theta_t^{(m)} \sim N(\hat{\theta}_t^{(m)}, \hat{\Sigma}_t^{(m)})$.

2.3.3 Model Probabilities and Inference

We initialize the agent with a uniform prior over models like those specified in Figure 2.2. The agent then uses Bayes' Rule to update model probabilities after each observed outcome; $p(m_l|r_t) \propto p(r_t|m_l)p(m_l)$. Here l is indexing models, and this updating happens for all models. To find the model-specific likelihood $p(r_t|m_l)$ we start with the likelihood of r_t conditional on the model and model weights; $p(r_t|\hat{\theta}_t^{(m_l)}, m_l) = r_t\hat{p}_t^{(m_l)} + (1-r_t)(1-\hat{p}_t^{(m_l)})$. Then integrate over the posterior over model weights; $p(r_t|m_l) = \int_\theta p(r_t|\theta_t^{(m_l)}, m_l)p(\theta_t^{(m_l)})d\theta_t^{(m_l)}$. Since there is no closed form solution, we use a Laplace approximation such that in practice we compute the likelihood with;

$$p(r_t|m_l) \approx (2\pi)^{\frac{d}{2}}\left|\frac{\partial^2}{\partial\theta^2}\ell(\hat{\theta}_{t-1}^{(m_l)})^{-1}\right|^{\frac{1}{2}} p(r_t|\hat{\theta}_t^{(m_l)})p(\hat{\theta}_t^{(m)})$$

Where d is the dimension of $\hat{\theta}_t^{(m_l)}$, and $p(\hat{\theta}_t^{(m_l)})$ is the probability density function of a normal distribution evaluated at $\hat{\theta}_t^{(m_l)}$ with mean vector $\hat{\theta}_{t-1}^{(m_l)}$ and covariance matrix $\hat{\Sigma}_{t-1}^{(m_l)}$. Unnormalized posterior model probabilities, $\omega_t^{(m_l)}$ were computed as; $\omega_t^{(m)} = p(r_t|m_l)p(m_l)$, and normalized posterior model probabilities were updated with;

$$p(m_l|r_t) \leftarrow \frac{\omega_t^{(m_l)}}{\sum_k \omega_t^{(m_k)}}$$

2.3.4 Policy

The agent computes expected reward across latencies under each model for given reward values U_r and response costs C_u, the specific values of which were defined based on their values in an experiment, and latency costs C_v, which were fit to behavioral data;

$$E^{(m_l)}(v(\tau)) = \hat{p}_t^{(m_l)} U_r + (1 - \hat{p}_t^{(m_l)}) C_u - \frac{C_v}{\tau_t}$$

The agent then finds the latency that would maximize reward rates under each model, $\tau^{*(m)} = \underset{\tau}{argmax} \left(\frac{E^{(m)}(v(\tau))}{\tau} \right)$. An example of expected reward rate functions is shown in the bottom-left subplot of Figure 2.3, with the maximum represented with a star and a dashes line down to the corresponding latency. A latency is then drawn from a gamma distribution with shape α and scale θ defined such that the mean of this distribution is τ^* and the variance v left as a free parameter;

$$\alpha = \frac{\tau^*}{\theta}$$
$$\theta = \frac{v}{\tau^*}$$
$$\tau_t \sim Gamma(\alpha, \theta)$$

Example policies are shown in the bottom-right subplot of Figure 2.3.

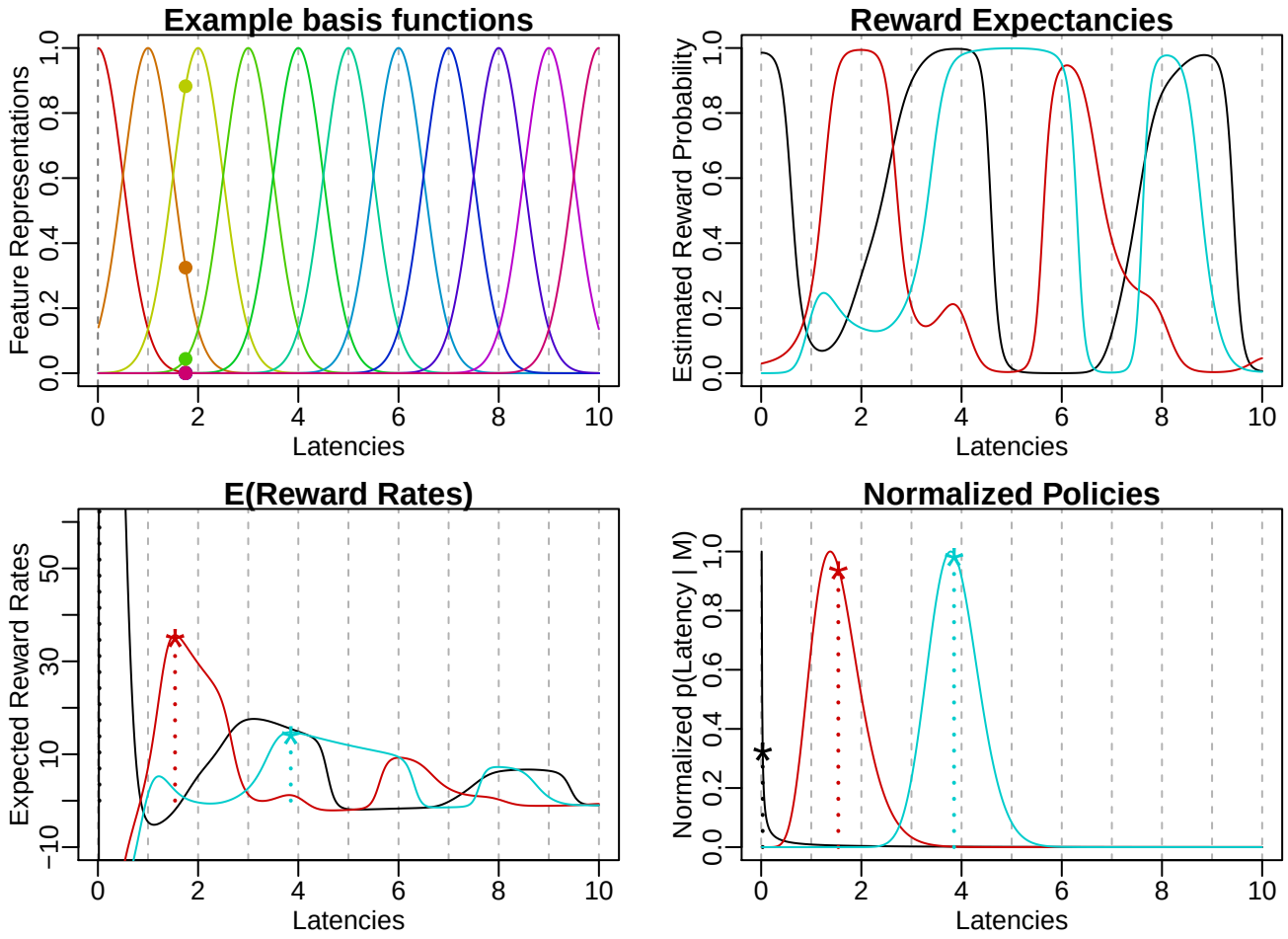

Figure 2.3: In this example, $\sigma = .25, U_r = 60, C_u = 10, C_v = 1, v = 0.25$. Top Left: An example of the Gaussian basis functions. Points represent the activation of different Gaussian functions, which make up the vector $x_t^{(m)}$ for a response with a latency of 1.75s. Top Right: Three possible inferred functions for $\hat{p}_t^{(m)}$ that were obtained using the Gaussian basis function from the top left figure and drawing values for the parameter vectors, θ, from a uniform distribution over the interval from -5 to 10. Bottom Left: Expected reward rates as a function of latency. Dashed lines and stars represent the maximum value of these functions and the corresponding τ^*. Bottom Right: Policies derived from the optimal latencies in the bottom left figure, normalized such that the maximum height is 1.

Part II

Learning To Control Schedules of Reinforcement

Chapter 3

Structural Inferences About Latency-Modified Schedules

3.1 Abstract

An ability to quickly learn relationships between actions and outcomes is essential for adaptive behavior and central to many everyday tasks. Such learning can be complicated, however, when the action-outcome contingency depends on the latency with which the action is performed. At the same time, discovery of and inferences about latency-modified schedules can greatly improve an agent's performance. We postulate that people explicitly represent these types of dependencies and use them to maximize reward rates. Formally, we specify a Bayesian reinforcement learner and contrast its performance with that of a model-free, actor-critic, algorithm using behavioral data from three latency-modified schedules of reinforcement. Our results suggest that a model-based characterization of latency-modified reinforcement schedules provides a superior account of behavior in free operant contexts.

3.2 Introduction

Schedules of reinforcement are functions that relate actions to valenced outcomes. The most extensively studied of these are ratio schedules – where delivery of response contingent rewards depends on the number of actions performed since the last rewarded response – and interval schedules – where response contingent reward delivery depends on the time elapsed since the last rewarded response. A substantial literature has addressed how these different schedules produce different rates and patterns of free operant responding (Ferster & Skinner, 1957), respectively induce goal-directed vs. habitual behavior (Dickinson, Nicholas, & Adams, 1983), and lead to distinct causal inferences about action-outcome relationships (Reed, 2001, 2003). A less studied aspect of ratio and interval schedules is that they differ with respect to the influence that an agent can have on the the probability of a response contingent reward. Specifically, on a random ratio schedule the probability of a response contingent reward is independent of the agents behavior; in contrast, on a random interval schedule, the probability of a response contingent reward depends on the latency with which the response is performed. A related class of schedules, differential rate schedules, likewise involve dependencies between latencies and response contingent reward probability – in this case the dependency is in terms of time between responses, rather than the time elapsed since the last rewarded response. Here, we use Bayesian structure inference to formally characterize a cognitive representation of the dependence (interval & differential rate), or lack of dependence (ratio), of the action-outcome contingency on response latencies.

As noted, over the last century, many theoretical accounts have been proposed to explain the influence of distinct schedules of reinforcement on operant behavior; however none, to our knowledge, have addressed the possibility of explicit inferences about the dependence between response latencies and response contingent reward probability. Rather, accounts have focused on the expected utility of an action, formalized as the sum over the products of the probabilities and utilities of its outcomes (Bernoulli, 1954), the causal relationship, or con-

tingency, between the action and reward, formalized as the difference between probabilities of reward given the presence and absence of the action respectively (Hammond, 1980), the error-derived cached value of an action based on its reinforcement history (Watkins, 1989), or the complex associative structure relating actions to stimuli and rewards (Dickinson & Balleine, 1993). In contrast, we postulate a functional relationship between action latencies and response contingent reward probabilities, arguing that the learner's goal is to explicitly infer this relationship. Specifically, we develop a Bayesian model of the discovery of latency-modified action-outcome contingencies based on structure inferences and estimated functional forms. We assess the performance of this model given behavioral data and compare its fit to an alternative, model-free, account.

3.3 The Models

The problem is formalized as a semi-Markov Decision Process (sMDP), as discussed in section 2.1. The reward functions were defined by the reward contingencies, as depicted in Figure 3.2. The contingencies used here were specified over only one action at a time, hence the action space consisted of only one action along with its latency τ. The task environment, depicted in Figure 3.1, consisted of only one state. Both our structure inference algorithm and the previously developed model-free algorithm (Niv, 2007) were based on this sMDP with reward value U_r, response costs C_u, and vigor costs C_v. Both models were designed to maximize reward rates, and both models specify the policy as a gamma distribution over latencies (τ). The key difference between the models is that the structure inference model obtains policy parameters from a model-based computation that depends on inferring the structure and functional form of the potential dependence of response contingent reward probabilities on response latencies, while the model-free account gradually nudges the distribution over latencies toward values associated with positive prediction errors.

3.3.1 The Structure Inference Model

Here I describe how our structure inference model developed in section 2.3 was adapted to the task used for this experiment. The primary problem for our structure inference algorithm was to infer from the data (observed actions and outcomes) whether or not the response contingent reward probability was a function of the latency of the action. The response contingent reward probability is shown as a function of latency for each condition in Figure 3.2. As described in subsection 2.3.2, and demonstrated in Figure 2.3, the structure inference account uses Gaussian basis functions to approximate the functions shown in Figure 3.2, and then infer whether the data are more likely under a model where these approximate functions are independent of latencies (m_1) or depend on latencies (m_2). The Gaussians were densely packed with a Gaussian centered at every $50ms$ between 0 and $250ms$, then less densely packed with a Gaussian at every $100ms$ between $300ms$ and 12 seconds. The variance of these Gaussians, σ^2, was fit to behavioral data. Posterior model probabilities were computed for a structure where response contingent reward probabilities were constant and a structure where they depended on latencies as described in subsection 2.3.3. Following Griffiths and Tenenbaum (2005), rather than using model probabilities directly, we used log Bayes factors as a cognitive model of structure inference.

3.3.2 The Actor-Critic Model

We also adapted the actor-critic model described in section 2.2 for the task used in this experiment. Because there is only one state, the prediction error computed by the critic reduces to;

$$\delta_t = r_t - \bar{R}_t \int_0^{\tau_t} e^{-s\eta_{\bar{R}}} ds - v_t(s_t)$$

where r_t, $\bar{R}$, and $v_t(s_t)$ are computed by the critic as described in subsection 2.2.2. Also, because there is only one action available on each block, the policy simplifies to a single Gamma distribution and the parameter η_m is no longer needed. Other than these simplifications, the model was applied as described in section 2.2.

3.4 The Experiment

The simplest setting in which to evaluate and test the two models is one where the latency of a single action can influence response contingent reward probability. In the current experiment, participants responded freely on four distinct actions, each available in a separate experimental block, where three of the four actions involved a latency-modified schedule of reinforcement.

3.4.1 Methods

Participants

21 participants (13 female, mean age=21, SD=3.47) at the University of California, Irvine participated for course credit. All participants gave informed consent and the study was approved by the Institutional Review Board at the University of California, Irvine.

Task & Stimuli

The code to run this experiment is avalable from my OSF repository (URL: https://osf.io/cbu6x/). The task stimuli are illustrated in Figure 3.1. Participants freely responded across four blocks, each of which lasted for seven minutes. While in a block, participants could press

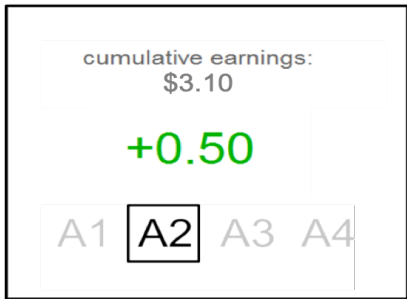

Figure 3.1: Interface for the experiment. Participants saw the avalable action in black text (here A2), and unavailable actions in gray text. When a response button was pressed, a black square appeared around the action and the net outcome value was shown in green text (for gains) or red text (for losses). Cumulative earnings were shown in gray in the top right (shown larger here for clarity).

the number keys at the top of the keyboard to earn rewards. The active key corresponded to the current block; the '1' key was active on the first block, '2' key on the second, and so on. Different keys were used across blocks to mitigate any possible order effects or carryover effects across blocks. The active key was conveyed with black text on the screen and the inactive keys were indicated with gray text. When the active key was pressed, a black square appeared around that key and the net outcome value was shown either in green text with a plus sign (for gains) or red text with a minus sign (for losses). All key presses carried a cost of \$0.10, and a probability of yielding a \$0.60 reward, each in hypothetical monetary units. Cumulative earnings were shown in small gray font at the top right of the screen, and updated immediately whenever a key was pressed.

The response contingent reward probabilities varied across blocks according to four schedules of reinforcement, the assignment of which to blocks was counterbalanced across participants. The four schedules used were: (1) a differential reinforcement of low rates (DRL) schedule where only responses with latencies (or inter-response-times; IRTs) over 5s produced the \$0.60 outcome, (2) a quadratic schedule where the probability of the \$0.60 outcome given a key press was highest for IRTs of 2s, (3) a differential reinforcement of high rates (DRH) schedule where a key press produced the \$0.60 outcome whenever three presses had occurred within 2s, and (4) a variable ratio (VR) schedule where the probability of the \$0.60 outcome

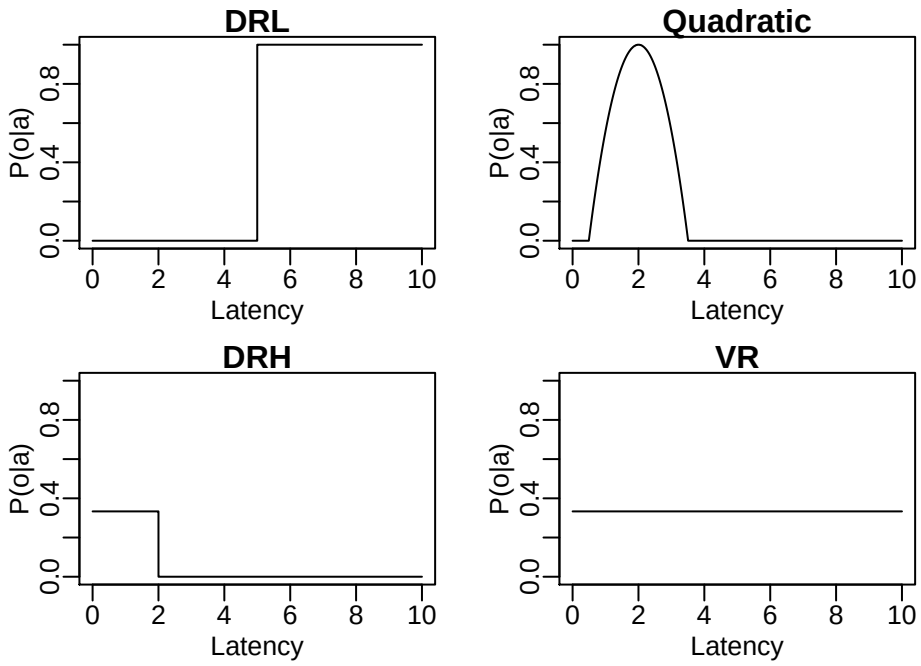

Figure 3.2: The schedules (reward functions) used in the single action experiment. Top Left: The differential reinforcement of low rates (DRL) schedule set the probability of a reward given a response to 0 for latencies less than 5 seconds, and 1 for latencies above 5 seconds. Top Right: The quadratic schedule set the probability of a response contingent reward to 1 for latencies of exactly 2 seconds, and reduced the probability quadratically for latencies farther away from 2 seconds. Bottom Left: The differential reinforcement of high rates (DRH) schedule set the response contingent reward probability to 1 for the 3rd response made in a 2 second window, and 0 for all other responses. Note that the probabilities shown in this figure average over three responses made within a 2 second window – that is, one out of three responses within a 2 second window will produce reward, implying a response contingent reward probability of $\frac{1}{3}$ among responses that occur within the 2 second window. Bottom Right: The variable ratio schedule fixed the probability of a response contingent reward to $\frac{1}{3}$ regardless of latency.

given a key press was set to 1/3. The functions relating latencies to response contingent reward probabilities defined by these schedules are shown in Figure 3.2.

3.4.2 Results

The data from this experiment can be downloaded from my OSF repository (URL: https://osf.io/cbu6x/).

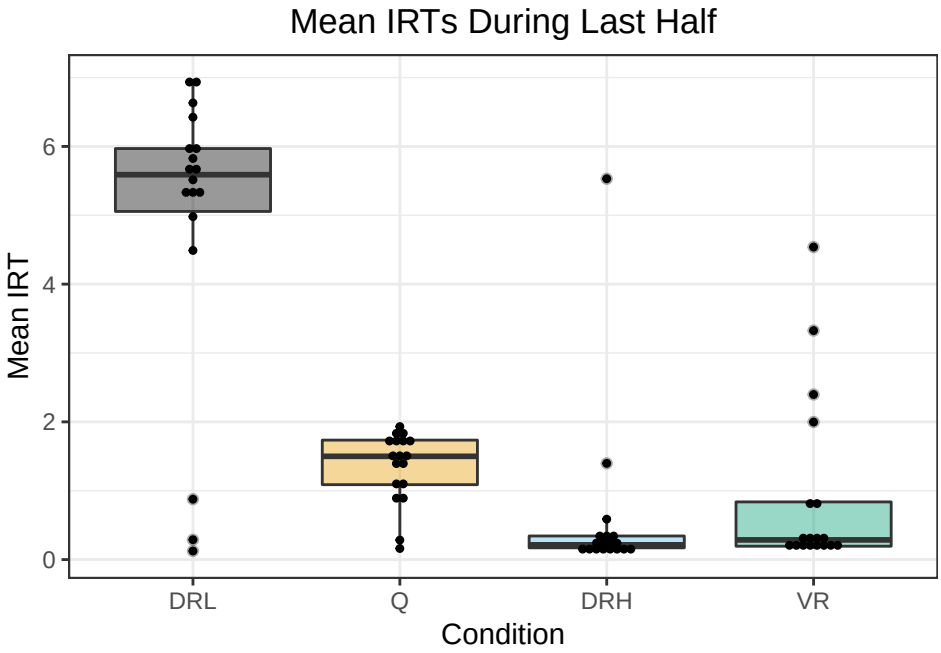

Figure 3.3: Mean latencies during the last half of each block.

Behavioral Results

To address extreme IRT values that were likely due to task disengagement, we removed IRTs that were more than 4 standard deviations from the mean of all IRTs across subjects and blocks. This led to the exclusion of 0.37% of responses from further behavioral analyses. However, this also excluded three participants from some of the analyses below due to their not having any recorded responses in the intervals subject to analysis. To contrast IRTs across the different schedules, we entered the mean IRTs from the second half (3.5 minutes) of each block for each participant into a mixed analysis of variance (ANOVA) with schedule and block order as within and between subjects factors respectively. Three participants were excluded from this analysis for failing to respond with a non-excluded IRT for the entire second half (3.5 minutes) of at least one 7-minute block. The analysis revealed a significant main effect of schedule, $F(3,15)=53.02$, $p<0.001$, but no significant effect of order, nor a significant interaction (smallest $p>0.05$). Planned comparisons (using two-tailed t-tests) revealed longer mean IRTs on the DRL schedule than the quadratic schedule ($t(17)=7.985$, $p<0.001$) and

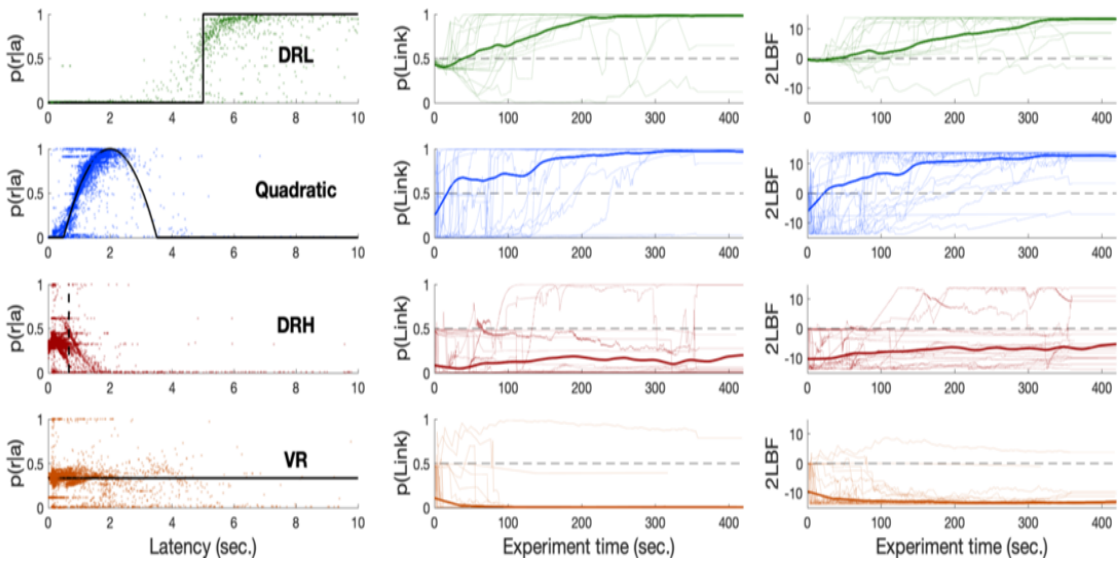

Figure 3.4: Left: Points represent every response made by any participant. The latency of the response is plotted on the x-axis, and the model-based predicted probability of response contingent reward is plotted on the y-axis. Black lines represent the true reward functions as in Figure 3.2. Middle: Posterior probability of a link model across time within each block. Faint lines represent individual participants and the thick line represents a windowed mean. The model quickly infers a low probability of a link in the DRH and VR conditions, and a high probability of a link in the quadratic and DRL conditions. Right: Twice the log Bayes factor shows strong evidence of the correct causal structure early on within each block.

longer mean IRTs on the quadratic schedule than DRH schedule (t(17)=2.386, p=0.029). In addition, comparing each latency modified scheduled to the latency independent variable ratio schedule, we found that mean IRTs were significantly longer on the DRL (t(17)=7.558, p<0.001) but not the quadratic schedule (t(17)=1.417, p=0.175). No difference was found between DRH and VR schedules (t(17)=1.237, p=0.233).

3.4.3 Model Results

Model-Based Function and Structure Inferences

Two important outputs from the model are the estimates of response contingent reward probabilities, $\hat{p}_t^{(m)}$, and model probabilities $p(Link)$. As demonstrated in Figure 2.3, the

model is capable of inferring a variety of functional forms. The left subfigure of Figure 3.4 demonstrates the model-based predicted probability of response contingent reward, $\hat{p}_t^{(m)}$, for latencies that were sampled by participants when those latencies were sampled. Qualitatively, the model-based estimates generally appear close to the true response contingent reward probabilities across latencies and conditions. This demonstrates that there is enough information in the data produced by people for our model to closely approximate the reward functions that are implied by a variety of common free operant contingencies, as well as a novel quadratic contingency.

Posterior probability of the link model $p(Link)$ is shown in the middle of Figure 3.4. Here the model infers the correct structure, on average, for all but the DRH condition. Notably, in the DRH condition, many participants concentrated responses at fast latencies where this schedule would be indistinguishable form a fixed ratio schedule. In off-policy simulation studies where latencies are uniformly sampled, the model does infer the link structure on DRH schedules. For ease of comparison with previous literature, we also show twice the log Bayes factor which is closely related to the log Bayes factor which has been dubbed causal support (Griffiths & Tenenbaum, 2005). Based on the criteria given by Kass and Raftery (1995), the model on average finds strong evidence of a link in the DRL and quadratic conditions, slight evidence against a link in the DRL condition, and strong evidence against a link in the VR condition, when making inferences based on behavioral data.

Model Comparisons

The model-based and model-free accounts were each calibrated using the first 6 minutes of data in a block. This allowed the model-based algorithm to approximate the reward function, and allowed the model-free algorithm to estimate state values and reward rates. We then obtained likelihoods of the observed latencies during the last minute of responding in each block while not allowing additional learning of model variables within that test period. These

likelihoods from the last minute of responding were used to estimate free parameters. The actor-critic model had 5 free parameters: η_v, $\eta_{\bar{R}}$, η_α, and η_θ were each fit within each of the 4 conditions, and C_v was fit across conditions. The structure inference model had 4 free parameters: v and σ were each fit within each of the 4 conditions, and C_v was fit across conditions. For both models, U_r and C_u were chosen to reflect the net outcome utility in this task[1].

We used the Bayesian information criterion (BIC) computed from the likelihoods obtained from the last minute of responding to formally compare the two models. BIC scores were computed for each participant and each schedule, and model performance across participants was assessed by contrasting median BIC scores with a sign rank test. If a participant failed to make any response in the final 1-minute test period of a block, that individual was excluded from the comparison involving the schedule in effect during that block. The model-based account provided a better fit in terms of BIC for 12 of 20 participants (p=0.627) on the DRL schedule, for 10 of 19 participants (p=0.546) on the quadratic schedule, for 14 of the 20 participants on the DRH schedule (p=0.014) and for 10 of 20 participants on the VR schedule (p = 0.681).

3.5 Discussion

The experiment presented here was a test of our model's ability to infer reward functions and discriminate between a structure where response contingent outcome probabilities depended on latencies from a structure where there was no such dependence. Behaviorally,

[1]This resulted in $U_r = 60$ in the model-free account and $U_r = 50$ in the model-based account, and $C_u = 10$ in both models. The differences in U_r are due to the different meanings of this variable within the models. In the model-free account, reward signals are given by $U_r - C_u - \frac{C_v}{\tau}$, such that the net outcome utilities are $50 - \frac{C_v}{\tau}$ for a rewarded response and $10 - \frac{C_v}{\tau}$ for an unrewarded response. In the model-based account, expected net reward is computed with $p_t U_r - (1 - p_t)C_u - \frac{C_v}{\tau}$, where U_r and C_u are the net outcome values for rewarded ($U_r = 50$) and unrewarded ($C_u = 10$) responses.

the schedules generally influenced participant behavior in the expected ways; mean latencies were slower on schedules that required slower latencies relative to those that required faster latencies. When fit to behavioral data, the model-based algorithm was able to learn an approximation of the reward functions, and make useful inferences regarding dependencies on latencies. Specifically, the model-based algorithm inferred a dependence in latencies on the DRL and quadratic schedules, which are the schedules where slower responses are necessary for a higher probability of reward. Such inferences could in principle be used to guide behavior, however, the model-based algorithm was not the best explanation for all participants across conditions.

In three of the four conditions, there was no detectable difference between the overall performance of the model-based and model-free algorithms when used to explain behavior. The only detectable difference in overall model fits was in the DRH condition, where the model-based algorithm was a better fit for 14 of 20 participants. In the absence of a dependence of response contingent outcome probabilities on response latencies, the model-based algorithm adapts a policy that can be interpreted as a heuristic to respond as fast as possible. Hence, the superior fit in this condition may be attributable to the 'respond as fast as possible' heuristic embodied by the model-based algorithm, rather than to it's ability to estimate functional forms and infer causal structures.

Overall, the experiment provided evidence that our model can approximate functional forms and infer causal structures from behavioral data. However, both the model-based and model-free algorithms can explain participant behavior in the settings investigated here. We address this limitation with experiments that were designed to more directly test the utility of inferring functional forms and causal structures.

Chapter 4

Structural Inferences About Latency Modification by Other Actions

4.1 Abstract

Many everyday activities involve the use of one action to modify the effects of another: When driving, shifting gears modifies the influence of pressing the gas pedal on acceleration; when cooking, the rate of adding a particular ingredient modifies the influence of stirring on viscosity. Here, we investigate a general ability to learn how to use actions to control schedules of reinforcement. In Experiments 2a and 2b, participants quickly discovered the optimal rate of responding on a *modifying* action that controlled the rate of reward contingent on performing a different, *modified* action. A group of participants who performed the task with a small probability of reward contingent on the modifying action, performed slightly worse when the contingency was specified in terms of response rates (Experiment 2a), and markedly worse when the contingency was specified in terms of inter-response times (Experiment 2b). Implications for formal theories of instrumental behavior are discussed.

4.2 Introduction

Since the early 20th century, researchers have investigated the influence of various reward schedules on the rate and selection of instrumental responses. For example, ratio schedules, in which reward delivery depends on the number of responses since the last reward, produce higher rates of responding than do interval schedules, in which reward delivery depends on the time elapsed since the last reward (Ferster & Skinner, 1957). When two or more action alternatives are available, that which yields the greatest, most immediate, or most certain reward is, all other things being equal, generally that most frequently selected (Rachlin, Raineri, & Cross, 1991). However, in the real world, many responses serve only to modulate the effects of other actions: The rate and pattern of pressing strings on a guitar does not itself yield music, but profoundly impacts the sounds produced by strumming. Here, we assess a domain-general capacity for learning about actions that control schedules of reinforcement on other actions.

Formally, the relationship between a particular action and its outcome has been modeled as a complex associative structure (Dickinson et al., 1983), as the difference between probabilities of reward given the presence versus absence of the action (Hammond, 1980), as the probability and subjective utility of the outcome given the action (Bernoulli, 1954), or as a cached value assigned to the action based on its reinforcement history (Watkins, 1989). What these diverse approaches have in common is that they address the identity and/or latency of a single action at a time, ignoring situations in which multiple, potentially interacting, actions are performed in concert. In our paradigm, an intermediate rate of responding on a *modifying* action maximizes the probability of reward contingent on performing a different, concurrently available *modified* action.

A well-studied phenomenon closely related to our query is that of "melioration" – a tendency to select an action alternative that produces a greater immediate pay-off, but that, when

selected repeatedly, lowers the overall rate of reward (Herrnstein, 1991). Such tendencies are commonly attributed to impulsivity (Herrnstein, 1991; Otto, Markman, & Love, 2012), but have also been described as rational choices under uncertainty (Gureckis & Love, 2009a, 2009b; Sims, Neth, Jacobs, & Gray, 2013). Other related paradigms, such as delay discounting (Ainslie, 1975; Johnson & Bickel, 2002) and differential reinforcement of low response rates (Wilson & Keller, 1953; Carter & MacGrady, 1966), have convincingly demonstrated the interfering influence of salient reward on rational decision-making (Ainslie, 1975; Van den Broek, Bradshaw, & Szabadi, 1987). In each experiment, we also assessed whether the lure of an immediate reward results in a failure to suppress responding on the modifying action, thus interfering with the ability to control the schedule of reinforcement on the modified action.

4.3 Experiment 2a: Response Rates

4.3.1 Methods

Participants

Forty-five undergraduates at the University of California, Irvine participated in the study for course credit. All participants gave informed consent and the study was approved by the Institutional Review Board of the University of California, Irvine.

Task & Procedure

The task interface is illustrated in Figure 4.1 and code to run this experiment can be downloaded from OSF (URL: https://osf.io/ebdfx/). We used a free operant paradigm in which participants were allowed to respond at will on two concurrently available actions, graphi-

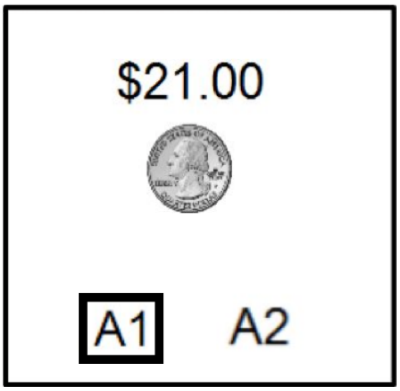

Figure 4.1: Depiction of task. The two actions were represented at the bottom of the screen. When an action was taken a black rectangle appeared around the representation of the key and, if rewarded, a quarter appeared at the center of the screen. Cumulative rewards were also shown at the center of the screen.

cally represented on the computer screen, by pressing the '1' key at the top of a keyboard, and the '2' key on the number pad on the right side of a keyboard. We intentionally chose keys on the opposite ends of the keyboard to enhance the distinctness of these actions – the keys were far enough away to require the use of both hands, rather than multiple fingers on one hand. A black rectangle appeared around the chosen action for 300ms whenever a key press was detected. If the response was rewarded, an image of a quarter appeared center screen for 500ms and a count of the cumulative monetary earnings, continuously displayed above the quarter image location, would increment by +$0.25. The task was comprised of ten 2-minute blocks separated by rest periods that the participant could terminate by pressing the space bar. All monetary earnings were fictitious.

In the "No Reward" group (n=15, 13 female, mean age = 22.13±4.47), the rate of responding on a "modifying" action influenced the probability that the concurrently available "modified" action would produce a reward. The "Reward" group (n=15, 11 female, mean age = 20.87 ± 2.42), was identical to the No Reward group except for an additional 20% chance of reward contingent on the modifying action. Note that, since this reward probability is much lower than the conditional, 0.9, probability of reward on the modified action, maintaining an optimal, intermediate, response rate on the modifying action dramatically increases the

average reward rate. When the modifying action was performed at an "optimal" rate of 1.25 to 2.75 presses per second, the probability of reward given a response on the modified action was 0.9. When response rates on the modifying action were outside of the 1.25 to 2.75 range, the probability of reward given the modified action was 0. The modifying action did not itself produce any reward. Response rates on the modifying action were tracked using a differential equation that increased by an impulse of 1 at the time of a response and decayed each impulse at a linear rate of 0.2 per second, so that each impulse from a response decayed to zero after 5 seconds. Specifically, for an impulse (a_i), which was 1 if an action were taken during the current iteration of the program and 0 otherwise, a decay rate of 0.2, a counter for the number of responses that occurred within the last 5 seconds (N_5) and the difference in time between the current iteration of the program and the previous iteration (Δt), the response rate variable (R) was updated on each iteration i by:

$$R_i \leftarrow R_{i-1} + a_i - 0.2N_5\Delta t$$

This method adjusts more quickly to changes in response rate than the commonly used approach of dividing the number of responses in a time window by the length of the window (e.g., Soto et al., 2006). The probability of reward on the modified action was set to 0.9 whenever the response rate variable, R, was in the optimal, 1.25 to 2.75, range and 0.0 otherwise. If melioration interfered with the ability to infer the dependency between reward probabilities given a modified response on latencies of modifying responses, we would see worse performance in the Reward group relative to the No Reward group.

Note that the optimal rate of responding on the modifying action was intermediate; this was done to rule out the contribution of systematic biases of either very high or very low responding. On the other hand, an intermediate rate might represent an average towards which most responders converge in free operant tasks. To address this possibility the "Yoked" group was included (n=15, 13 female, mean age = 23.07 ± 4.82), in which the rate of

responding on the modifying action had no influence, rather the probability of reward on the modified action was yoked to that of a participant in the No Reward group throughout time within a block. We predicted that, by the end of the session, participants in the No Reward group would respond on the modifying action at a rate falling within the optimal range, while those in yoked group would not.

4.3.2 Results

Behavioral

No data was discarded for cleaning, or any, purposes. We computed the mean distance of the RR variable from an edge of the optimal region when responding throughout the task. The mean and standard error for this statistic is shown by condition, across blocks in Figure 4.2. During the last block, mean distance from optimal was lower relative to the Yoked condition for both the No Reward ($t(28)=6.662$, $p<0.001$) and Reward ($t(28)=5.627$, $p<0.001$) conditions. Because differences were already apparent by the first block, we also computed mean distance from optimal within 10 second bins. In the first 10 second bin[1], there was no evidence of a difference between the No Reward and Reward group ($t(28)=1.546$, $p=0.133$), the No Reward and Yoked group ($t(27)=1.04$, $p=0.308$), nor the Reward and Yoked group ($t(27)=0.43$, $p=0.671$). Comparing the groups with a contingency to the Yoked group, both the No Reward ($t(26)=3.142$, $p=0.004$) and Reward ($t(27)=4.356$, $p<0.001$) groups showed evidence of a lower distance from optimal responding in the last 10 second bin. Evidence of learning in terms of a lower distance from optimal in the last bin relative to the first bin was observed for both the No Reward ($t(28)=4.384$, $p<0.001$) and Reward ($t(27)=5.738$, $p<0.001$) groups, and not the Yoked condition ($t(26)=0.15$, $p=0.882$).

[1]The different degrees of freedom reflect the absence of responses from some participants within the 10 second bins. Since these participants did not respond within the 10 second bin, they could not be included in these analyses.

Behavioral Results Figure-1.pdf

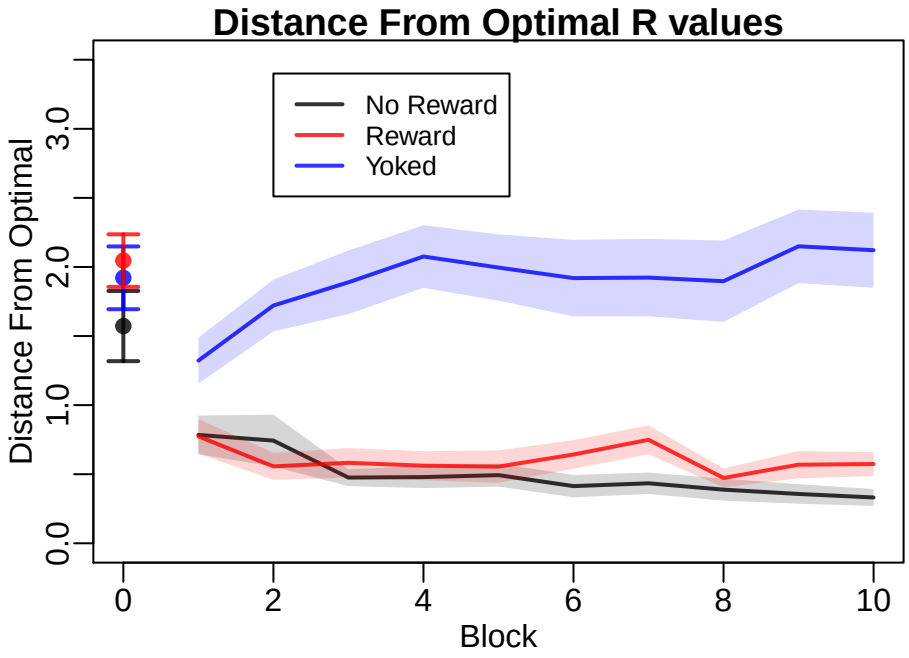

Figure 4.2: Absolute valued distance between R and an edge of the optimal region. Lines represent means of participant means across blocks, shaded regions represent standard errors. Points on the left are the mean of participant means during the first 10s bin, and bars represent standard errors.

4.4 Experiment 2b: Latencies

4.4.1 Methods

Participants

Fifty undergraduate students at the University of California, Irvine participated in the study for course credit. All participants gave informed consent and the study was approved by the Institutional Review Board of the University of California, Irvine.

Task & Procedures

The task used for Experiment 2b was very similar to that of Experiment 2a. The task interface was the same as that illustrated in Figure 4.1. We also used the same a free operant paradigm, the same keyboard keys, and the same stimuli for feedback as in Experiment 2a. The key difference was in the reward contingency. For Experiment 2b, we replaced the dependency on the response rate variable R with a simpler dependency on latencies. Here a response on the modified action was rewarded with probability .9 if the last latency on the modifying action had been between 2s to 5s, and also if less than 5s had elapsed since the last modifying response. As in Experiment 2a, participants were assigned to either a No Reward group (n=25, 21 female, mean age = 21.44 ± 3.04) where this contingency was the only way to earn rewards, or a Reward group (n=25, 20 female, mean age = 20.16 ± 2.21) which was the same as the No Reward group except for an added probability of .2 that performing the modifying action would yield a reward. Reward magnitudes were the same as in Experiment 2a. And again, if melioration interfered with the ability to infer the dependency between reward probabilities given a modified response on latencies of modifying responses, we would expect to see worse performance in the Reward group relative to the No Reward group.

4.4.2 Results

Behavioral

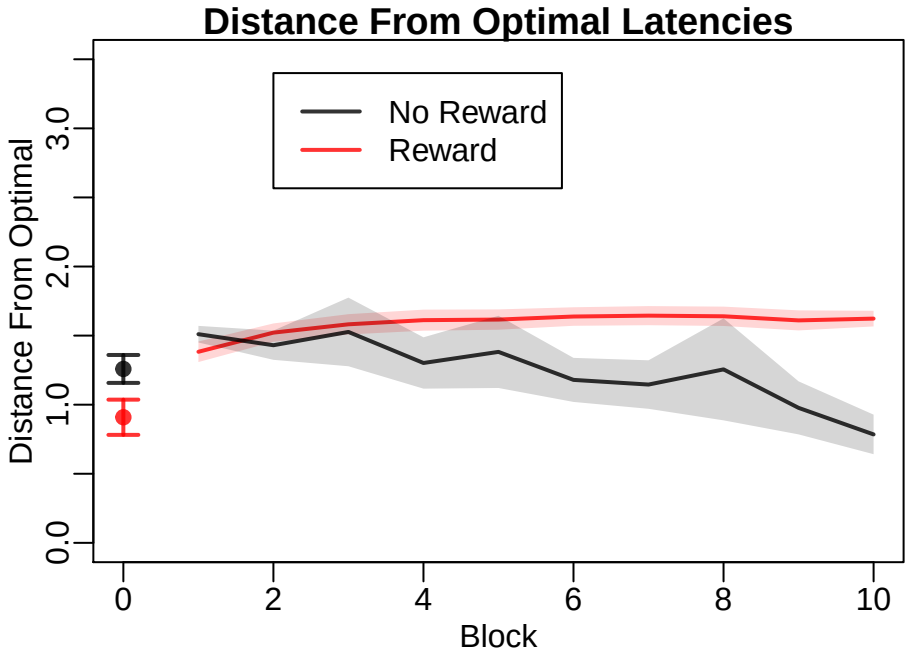

Figure 4.3: Absolute valued distance between observed latency and an edge of the optimal region. Lines represent means of participant means across blocks, shaded regions represent standard errors. Points on the left are the mean of participant means during the first 10s bin, and bars represent standard errors.

No data was discarded for cleaning, or any, purposes. We computed the average distance between observed latencies and optimal latencies on the modifying action for each participant in each block. Figure 4.3 shows the mean of these values across blocks. There was no evidence of a difference between groups in the first block ($t(48)=1.357$, $p=0.181$). Comparing the last and first blocks, the average distance from optimal decreased for the No Reward group ($t(24)=4.8$, $p<0.001$) and increased for the Reward group ($t(24)=3.005$, $p=0.006$). For consistency with the analysis done in Experiment 2a, we also checked these statistics in 10 second bins. While there was evidence that participants in the No Reward group had a higher mean distance from optimal in the first 10 second bin ($t(46)=2.149$, $p=0.037$), the

other results were consistent with those obtained from averaging over blocks in that the average distance from optimal decreased for the No Reward group ($t(45)=4.919$, $p<0.001$) and increased for the Reward group ($t(47)=5.301$, $p<0.001$) when comparing the first and last bins.

4.5 Discussion

In two experiments, we assessed the discovery and performance of an action that controlled the schedule of reinforcement on another, concurrently available, action. In both experiments, participants in the No Reward condition quickly discovered and implemented a near-optimal, intermediate, response rate on a modifying action that, while not producing any rewards itself, modulated the reward contingent on a distinct, concurrently available, action. In Experiment 2a, response rates in a yoked control group confirmed that convergence to the optimal rate was due to the influence of the modifying action on the reward schedule of the modified action. Consistent with a large literature on the failure to suppress inappropriate responding in the face of immediate reward (Ainslie, 1975; Carter & MacGrady, 1966; Van den Broek et al., 1987; Wilson & Keller, 1953), reinforcement of the modifying action apparently prevented discovery of the optimal response rate when the contingency was specified in terms of the most recent latency (Experiment 2b), but not when the contingency was in terms of the response rate variable (Experiment 2a). The focus in the existing literature on the disruptive effects of immediate reward has largely overshadowed the question raised here of whether, and how, agents learn about actions that modify schedules of reinforcement. Our results suggest that increasing levels of instrumental control can be achieved by incorporating information about dependencies between actions, particularly in the absence of competing reward contingencies.

Current work in our lab is focused on extending the models described above to the contin-

gencies developed here. To extend the model-free account developed by Niv (2007) to this context, we need a definition of the state space that captures the contexts in which participants should partake in different policies. One choice would be to defined the state space with respect to the relevant ranges of the variable used for the contingency – the response rate variable in Experiment 2a, and latencies in Experiment 2b. However, we did not want to hard-code exact contingency knowledge, as this is an unrealistic modeling assumption. Instead, we discretized the relevant variables in each experiment into 10 bins. Preliminary results suggest that the model is able to infer appropriate policies with this representation of the state space; Specifically the model prefers the modifying action when it has been used in the past to transition into states associated with a high response contingent reward probability on the modified action. For the model-based approach, earlier versions of the model were able to recover the reward functions and causal structures from behavioral data, however these versions were only early assessments of the function and structure inference abilities of this model, and did not include a policy or parameters that were fit from data. Fitting free parameters of the model-based algorithm requires extensive computational resources and may require hyperparameter tuning with respect to the number and location of Gaussian basis functions.

In conclusion, we have demonstrated a domain-general ability to learn about, and take advantage of, an action that modifies the schedule of reinforcement on a different action. We have also sketched the preliminary application of our algorithm, which makes inferences about dependencies between response latencies and conditional reward probabilities, and can account for behavior across a wide range of instrumental schedules.

Chapter 5

A Test of Structure Inference

5.1 Abstract

Since model-based algorithms maintain an abstract representation of the relationship between actions and outcomes, these algorithms can be particularly adaptive when faced with changing environments. For example, if a common path to reward in a maze is blocked, a model-based learner would query its model to quickly discover a new optimal route. However, a model-free learner would be stuck with actions that appear best based on action values learned before the path was blocked. Here we probe an ability to adapt to a truncation of the action space on an action-modified schedule. We trained participants on a reward function that produces a high probability of response contingent reward for fast and slow, but not intermediate, responding. We then blocked slow latencies, a manipulation that would immediately shift a model-based, and not model-free, learner towards faster responding. Participants quickly shift towards fast responding when feedback is provided (Experiment 3a), but not when feedback is masked (Experiment 3b).

5.2 Introduction

Model-based and model-free algorithms are thought to underlie, respectively, goal directed and habitual behavior. Goal directed behavior is sensitive to outcome devaluation and contingency degradation, while habitual behavior is maladaptively persistent despite changes in the environment. Outcome devaluation procedures (Singh, McDannald, Haney, Cerri, & Schoenbaum, 2010) involve a learning phase where a stimulus or action is associated with a valued outcome, followed by a phase where the outcome is devalued (e.g. through satiety or aversive conditioning). Model-based, but not model-free, algorithms update outcome values in response to devaluation and recompute action values based on a model of response contingent outcome probabilities, hence model-based algorithms uniquely predict a change in responding following outcome devaluation. Contingency degradation procedures involve a learning phase where an action is associated with a valued outcome, then a phase where that action-outcome contingency is degraded. Since model-based algorithms maintain an abstract representation of this relationship, they quickly update and react to the degraded contingency.

Model-based, but not model-free, algorithms can also use models of the environment to flexibly change behavior in response to changes in the available set of actions or changes in the domain of a continuous action space. Since latency modified schedules involve a choice of the latency with which to respond, we postulated that the model-based, but not model-free, algorithm would respond to a change in the available set of latencies by quickly shifting the latencies with which responses are made. Here we introduce a novel paradigm that tests responsiveness to a truncated action space, and hence model-based or model-free control of behavior on latency modified contingencies.

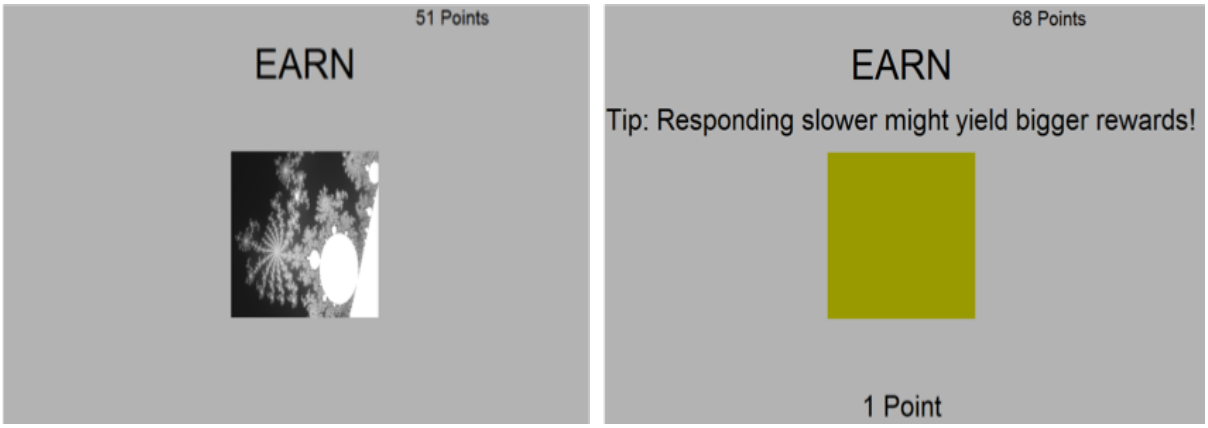

Figure 5.1: Interface for the experiment. Participants saw a fractal response stimulus at the center of the screen and the text 'EARN'. Upon responding the points earned or lost were shown at the bottom of the screen in black text, and the response stimulus was covered by a square for 50ms as feedback. In the Blocking phase, the yellow rectangle disappeared at a rate such that it was gone after the blocking period of 1 second.

5.3 Experiment 3a: No Baseline

5.3.1 Methods

Participants

We recruited 54 undergraduate students at the University of California, Irvine and paid them based on their performance as described below. All participants gave informed consent and the study was approved by the Institutional Review Board of the University of California, Irvine.

5.3.2 Task and Stimuli

Participants responded across three phases to earn points that were exchangeable for money at a rate of $0.01 per point. In all phases, latencies faster than 500ms produced a gain of 1 point, latencies between 500ms and 2 seconds resulted in a loss of 1 point, and latencies

slower than 2 seconds resulted in a gain of 3 points. This contingency ensured that reward rates were maximized by responding with latencies below 500ms. Participants began the task in a Training phase where they were free to respond, by pressing the space bar, at any time. The task stimuli are shown in Figure 5.1. Participants stayed in this phase until they made five responses slower than 2 seconds. After 20 seconds in this phase, prompts occasionally appeared to encourage participants to sample faster or slower latencies if less than 10% of the participant's responses faster than 500ms or slower than 2 seconds, respectively. An example prompt to sample slower latencies is shown in Figure 5.1. If a participant was in the training phase for more than 12 minutes, the task automatically terminated and the participant was dismissed. After five responses slower than 2 seconds, a Free Response phase began where the prompts were disabled. During this phase participants were expected to shift back to responding quickly, as such responding would maximize the reward rate. After one minute in this phase, the proportion of fast responses was computed throughout the remainder of this phase. Once 80% of responses were made with fast latencies, the Free response phase ended and participants began the Blocked phase. During the Blocked phase, responding was blocked for 1 second after each response. This blocking was communicated to participants through an instruction screen and represented during the task with a square that covered the response stimulus for 1 second, but gradually uncovered the response stimulus such that the response stimulus was fully visible after 1 second. This blocked participants from earning rewards from responses faster than 500ms. The Blocked phase lasted for 1 minute, and feedback was shown throughout the blocking phase.

5.3.3 Results

Behavioral Results

Of the 54 recruited participants, 15 could not be included in analysis for various reasons[1]. All analyses are based on the 39 participants (30 female, mean age = 21.15 ± 2.89) who were able to complete the task.

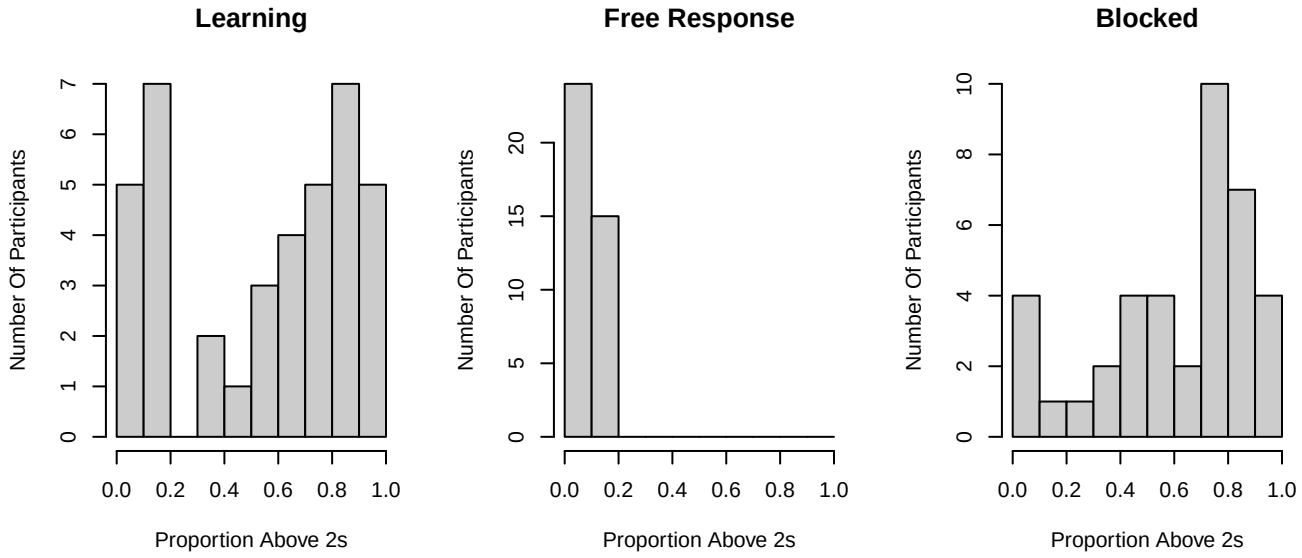

Figure 5.2: Distribution of the proportion of latencies above 2s during the blocking phase

Since the block durations differed across participants, we used the proportion of latencies over 2 seconds in each block as our outcome variable. This measure is shown in Figure 5.2. We found evidence that the proportion of latencies over 2 seconds was higher on average in the blocked phase relative to the free response phase ($t(38)=12.158$, $p<0.001$).

[1]These participants were all missing data from the Blocked block. Some did not complete a concurrently run task fast enough, others did not meet the criteria to pass the Training or Free Responses phases. See subject log for more details

5.3.4 Discussion

Although, as predicted, participants shifted to fast responding in the Free Response phase and then shifted to slower responding in the Blocked phase, there was a key limitation in the design of this experiment. Our primary hypothesis was that participants would shift back because of a model-based representation of the effect of latencies on response contingent outcome probabilities. While this result is obtained when averaging over behavior throughout the blocked phase, many participants begin the blocked phase with responses that were too fast to be rewarded. They then shifted to slower responses after experience with fast, unrewarded responses. Hence it is possible that the shift in this last phase could be due to learning from feedback during the blocked phase, rather than a model-based representation of the task. In Experiment 3b we mask all feedback during the blocked phase, so that behavior in this phase cannot be attributed to subsequent learning after fast rewarded latencies are blocked.

5.4 Experiment 3b: Baseline

5.4.1 Methods

Participants

We recruited 14 undergraduate students at the University of California, Irvine and paid them based on their performance as described below. All participants gave informed consent and the study was approved by the Institutional Review Board of the University of California, Irvine.

5.4.2 Task and Stimuli

Experiment 3b added an initial Baseline phase and a final Test phase to Experiment 3a. Both the Baseline and Test phases were one minute long and in both the Baseline and Test phases, feedback and cumulative points were masked by a blue rectangle. Participants were exposed to the baseline phase before any experience with the task. This was followed by the Learning and Free Response phases, which were the same as in Experiment 3a, and then the Test phase. The Test phase was identical to the Blocking phase in Experiment 3a with the exception that feedback and cumulative points earned were masked by the blue rectangle.

5.4.3 Results

Behavioral Results

Of the 14 participants recruited, 7 did not complete the task and hence could not be included in the analysis. All reported analyses are based on the 7 participants (5 female, mean age $= 21.57 \pm 6$) who completed the task.

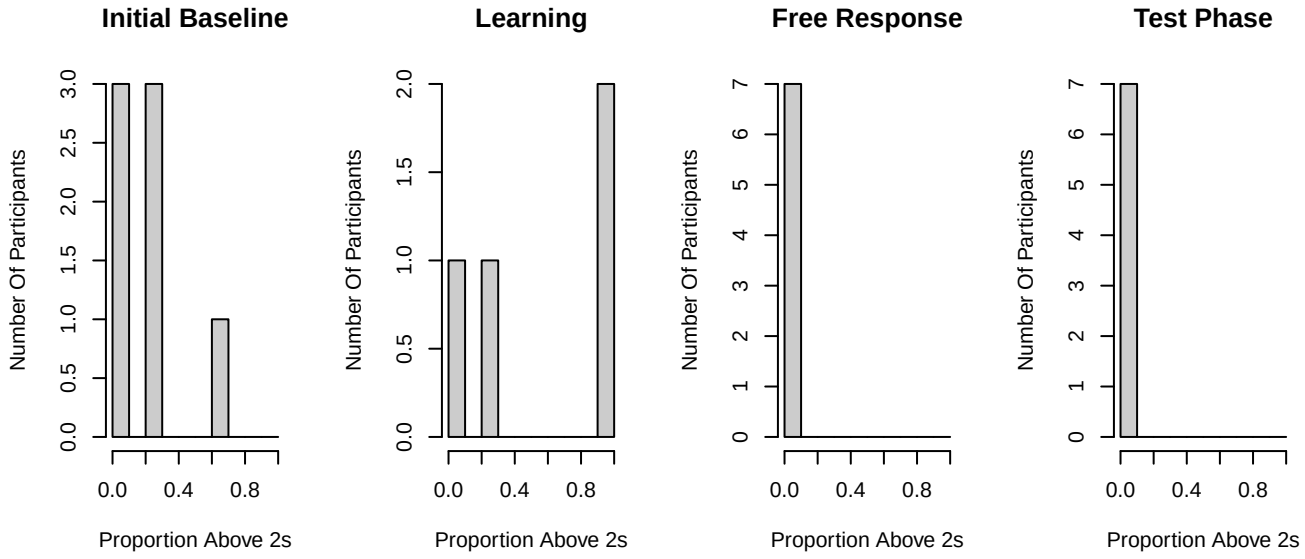

Figure 5.3: Distribution of the proportion of latencies above 2s during the blocking phase

Evidence at the $p = .1$ but not $p < .05$ level for a higher proportion of responses over 2 seconds in the Baseline phase relative to the Test phase (t(6)=2.171, p=0.073). Performing the same analysis as in Experiment 2a, there was evidence that average proportion of latencies above 2 seconds was higher during the Test phase relative to the Free Response phase (t(6)=3.731, p=0.01) but this effect could have been driven by the blocking itself disallowing bursts of responses at latencies under 1 second. Because of this limitation, we also tested for a difference in the number of responses with latencies above 2 seconds – this statistic is not directly influenced by the number of responses less than 2 seconds. We found evidence of fewer responses on average over 2 seconds in the Test phase relative to the Baseline phase (t(6)=5.756, p=0.001).

5.5 Discussion

Here we introduced a novel experimental paradigm to arbitrate between model-based and model-free control of behavior on latency modified schedules. In Experiment 1, we had shown that the model-based approach developed here was able to approximate reward functions and infer an influence of latency on response contingent reward probabilities when such a relationship was necessary for reward maximization. In light of those results, it is reasonable to assume that such an approach would also be able to infer the relationship between latencies and response contingent reward probabilities used in this experiment and use this knowledge to shift to slower latencies here when fast latencies are blocked. Indeed, in Experiment 3a, we observed a higher proportion of slow latency responses in the Blocked phase, which was indicative of a shift to slower latencies when fast latencies were blocked. However, the feedback presented during the blocked phase precludes attribution of this result entirely to model-based behavior. Experiment 3b was designed to test a shift attributable to model-based behavior in the absence of feedback.

While we have only a small sample of pilot data, we did not find evidence of a shift to slow latencies when fast latencies were blocked in Experiment 3b. One interpretation is that the lack of a shift here indicates that the results of Experiment 3a were due to learning from the feedback in the Blocked phase. However, the lack of an immediate shift in Experiment 3b may also be attributable to failures in learning or in discriminating. Due to the large number of participants who had failed to reach the criterion for progressing past the training phase, we ran these experiments with a relatively low criterion of only 5 slow responses before prompts were disabled. Perhaps this criterion is too low for participants to adequately approximate the reward function, hence a future study with a higher criterion may find more participants shift to slower latencies based on their knowledge of the schedule. Additionally, the difference of 1.5 seconds between the boundary for slow and fast rewarded responses may have been too small to have been discriminated, particularly combined with limited observations of

slow rewarded responses. Overall, it is very possible that a failure to shift to slower latencies in Experiment 3b reflects a failure to learn, or discriminate the boundary of, the response contingent reward probabilities for slow responses. Future studies could implement a higher criterion for disabling prompts, and/or a wider temporal distance between the boundaries of rewarded fast and slow responses.

Chapter 6

General Discussion

6.1 Summary and Conclusions

Inspired by the breadth of factors that may influence behavior, particularly rewarded inter-response times (Reynolds & McLeod, 1970; Alleman & Platt, 1973) and perceived causality (Reed, 2001, 2003), we developed a model-based reinforcement learning algorithm that uses observed response latencies and outcomes to infer a reward function and causal structure. To the best of our knowledge, this is the first model-based account of free-operant behavior. While the primary goal was to develop a model capable of detecting the causal influence of latencies on response contingent reward probabilities, our approach can also be used to explain and analyze free operant behavior on a plethora of contingencies.

In the first of several experiments, we demonstrated that the model could approximate reward functions and make the correct causal inferences on common contingencies, as well as a novel contingency that involved a quadratically shaped reward function. We fit this model to behavioral data and compared the fit to a model-free algorithm inspired by (Niv, 2007). Importantly, this was the fist time that reinforcement learning algorithms had been

fit to free-operant behavior. In the case of the model-free algorithm, fitting this model to behavioral data extended the simulation and qualitative work presented by Niv (2007); and in the case of the model-based algorithm this provides a strong empirical foundation from which the model can be further developed. Both the model-free and model-based algorithms qualitatively captured behavior on common contingencies, and the only case where there was a statistically significant difference in median fit was behavior on the DRH contingency. Though the model-based algorithm provided a better fit to behavior on a DRH contingency, this was likely driven by the choice rule for cases where there is no influence of latency on response contingent reward probability, which reduces to a "respond as fast as possible" heuristic. Subsequent experiments were conducted to arbitrate between model-based and model-free control of behavior on latency-modified contingencies.

In Experiment 2a and Experiment 2b, we introduced a novel contingency where the latencies on one action influenced the response contingent reward probability on another action. This contingency was used to test the inference of a latency-modified schedule when the latencies of one, modifying, action influenced the response contingent reward probability for another, modified, action. In both experiments, participants in the central No Reward condition quickly learned this contingency, evidenced by a quick reduction in the distance between their response rate (Experiment 2a) or latencies (Experiment 2b) and the optimal response rate (Experiment 2a) or latencies (Experiment 2b) on the modifying action. Prior research had indicated that smaller rewards can interfere with an ability to acquire optimal behavior on complex contingencies (Herrnstein, 1991). To test for such an interference here, we included a Reward condition where the modifying action had a small probability of producing reward itself. This impeded learning when the contingency was specified in terms of latencies (Experiment 2b), but not response rates (Experiment 2a). Work to extend the models to the contingency employed here is ongoing, though a crucial and interesting choice involved in model development is the representation of latencies when defining the state space. While the experiments here were not designed to test this, the difference between Experiment 2a and

Experiment 2b in the interfering effect of rewards from the modifying action on the discovery of the latency-modified contingency here might be related to the way participants represent latencies. That is, if participants naturally represent response rates rather than latencies of a single response, then it may be easier to detect a latency-modified contingency when such a contingency is specified in terms of response rates. Investigating the environmental features that facilitate or impede the efficacy of smaller rewards in interfering with the ability to detect more lucrative reward contingencies, and how this may relate to representations of a task, is an interesting direction for future research.

Experiment 3a and Experiment 3b employed another novel contingency where response contingent rewards were delivered for fast and slow, but not intermediate, latencies. After experience with this contingency during a learning phase, fast latencies were blocked. This was used to test the combined function and causal inference components of the model-based algorithm. If participants approximated the function and inferred a link, then they should have immediately shifted to slow latencies when the fast latencies were blocked. Such a shift was observed in Experiment 3a, when feedback was provided during the crucial test phase, but not in a pilot version of Experiment 3b, where no feedback was provided. These immediate results suggest that continued learning during the test phase underlied the shift to slow latencies. However, further adjustments to the design are needed to ensure that the criterion used for learning the contingency was sufficient, and that the boundaries for slow and fast latencies are discernible. Possible adjustments include: higher reward magnitudes for the slower latencies, which may incentivize more natural exploration of these latencies that is required to approximate the reward function; a wider gap between the boundaries for rewarded fast and slow latencies, which may help in making these boundaries more discernible; and a higher learning criterion so that more experience with rewarded slower latencies is acquired, which would facilitate inferring and remembering that these slower latencies also produce high response contingent reward probabilities.

In summary, we developed a model-based reinforcement learning algorithm that uses latency reinforcement to infer reward functions and causal structures. We demonstrated that this model can recover approximation of reward functions and make correct causal inferences when fit to behavioral data. We also found that behavior conforms to such inferences when shaped by a contingency where latencies on one action control the response contingent reward probabilities on another action, though such inferences may be mitigated in nuanced ways by interfering rewards. Finally, we found evidence that inferences about reward functions and causal structures may be used to produce behavior that is robust against limitations in the continuous action space comprised of latencies. However, in a pilot study that blocked feedback, behavior did not seem to be guided by knowledge of the underlying contingency. Future experiments are needed to elucidate the features of complex latency-modified contingencies that facilitate explicit knowledge of their functional forms.

6.2 Complex Empirical Patterns of Responding and Their Implications

The algorithms described here perform well in terms of finding an optimal time to wait before responding. Both algorithms can fit a steady rapid pattern of responding that characterizes variable ratio schedules, and a steady slow pattern of responding that characterized variable interval schedules. However, fixed ratio and fixed interval schedules of reinforcement often produce more complex patterns in responding (Ferster & Skinner, 1957) which have important implications for the models described here.

Fixed ratio schedules generally produce a post-reinforcement pause, characterized by a long latency following a rewarded response despite short latencies while in pursuit of the reward (Ferster & Skinner, 1957). This pause has been attributed to fatigue, because it increases

with the ratio requirement of the proceeding contingency, and also procrastination, because it also increases with the ratio requirement of the upcoming contingency (Schlinger, Derenne, & Baron, 2008). Since this pause is unique to ratio, and not latency-modified, schedules, it could indicate an inference of the independence between response latencies and response contingent outcome probabilities – the key feature of the model-based algorithm developed here. Furthermore, post-reinforcement pauses contradict predictions of the model-free algorithm. This is because response vigor correlates with reward rates under this model, implying that latencies would slow down during unrewarded responses because of the corresponding drop in reward rate, and speed up following a rewarded response because the new reward would increase reward rates. Overall, post-reinforcement pausing indicates an ability to detect a difference between ratio schedules and latency-modified schedules, and contradicts predictions of the model-based account described here.

Fixed interval schedules generally produce a scalloping pattern, where latencies immediately after a rewarded response are longer and then become shorter as the interval requirement approaches. Such scalloping is eliminated by previous experience on a ratio schedule (Weiner, 1969). The exclusivity of a scalloped response pattern to fixed interval schedules, along with the disruption of this pattern when participants first experience a ratio schedule, indicates that the scalloping behavior depends on an inference about the dependence between response contingent outcome probability and time since the last rewarded response. Such an inference could be made by a version of our model-based algorithm that includes features of time since the last rewarded response in the function approximation component. And again, such a response pattern contradicts predictions of the model-free algorithm discussed here because the decreasing reward rate between rewarded responses would imply a decelerating response profile, instead of the observed accelerating response profile.

Overall, both post-reinforcement pausing on ratio schedules and scalloped responding on interval schedules strengthen the case for a structure inference ability, which is the foundation

of the novel model developed here. Both response profiles also contradict the predictions of the model-free algorithm described here.

6.3 Response Bursting and Future Directions

Both the model-free algorithm and model-based algorithm described here employ one latency distribution around a latency that maximizes reward rates. In contrast, empirical data are ripe with examples of response bursting (Reed, 2001, 2003, 2011; Tanno et al., 2009, 2012; Schneider, 1969; Eldar, Morris, & Niv, 2011) which would require at least two latency distributions – a faster within-burst distribution, and a slower between-burst distribution. To remedy this discrepancy, the models described here can be seen as models of the slower, between-burst latencies and can be extended to incorporate an additional within-burst distribution.

Bursting may be important because of the type of evidence it provides about the environment. Imagine an agent that only samples one, slow latency at which there is a high response contingent reward probability. Without sampling other latencies, the agent would not be able to infer whether response contingent reward probabilities depend on latencies or do not depend on latencies. Indeed prior research has emphasized the role of sampling different latencies in differentiating between ratio and interval schedules (Tanno et al., 2009, 2012). Particularly, an agent would benefit if they discover that response contingent reward probabilities are also high at faster latencies. Evidence that aids in this discovery could be provided by scalloped responding, where latencies slightly faster than the optimal latency are sampled, or by response bursting, where fast latencies are often sampled. An important future direction would be to further develop this normative role of response bursting and scalloping by modeling optimal sampling within the context of our Bayesian structure-inference account of free operant behavior.

Bibliography

Acuna, D., & Schrater, P. R. (2009). Structure learning in human sequential decision-making. In *Advances in neural information processing systems* (pp. 1–8).

Ainslie, G. (1975). Specious reward: a behavioral theory of impulsiveness and impulse control. *Psychological bulletin*, *82*(4), 463.

Allan, L. G. (1980). A note on measurement of contingency between two binary variables in judgment tasks. *Bulletin of the Psychonomic Society*, *15*(3), 147–149.

Allan, L. G. (1993). Human contingency judgments: Rule based or associative? *Psychological Bulletin*, *114*(3), 435.

Alleman, H. D., & Platt, J. R. (1973). Differential reinforcement of interresponse times with controlled probability of reinforcement per response. *Learning and Motivation*, *4*(1), 40–73.

Bernoulli, D. (1954). Exposition of a new theory on the measurement of risk (L. Sommer, Trans.). *Econometrica*, *22*(1), 23–36. (Original work published 1738)

Boneau, C. A., & Cole, J. L. (1967). Decision theory, the pigeon, and the psychophysical function. *Psychological Review*, *74*(2), 123.

Carter, D. E., & MacGrady, G. J. (1966). Acquisition of a temporal discrimination by human subjects. *Psychonomic Science*, *5*(8), 309–310.

Cools, R., Clark, L., Owen, A. M., & Robbins, T. W. (2002). Defining the neural mechanisms of probabilistic reversal learning using event-related functional magnetic resonance imaging. *Journal of Neuroscience*, *22*(11), 4563–4567.

Daw, N. D. (2003). *Reinforcement learning models of the dopamine system and their behavioral implications* (Unpublished doctoral dissertation). Citeseer.

Dickinson, A. (1985). Actions and habits: the development of behavioural autonomy. *Philosophical Transactions of the Royal Society of London. B, Biological Sciences*, *308*(1135), 67–78.

Dickinson, A., & Balleine, B. (1993). Actions and responses: The dual psychology of behaviour.

Dickinson, A., Nicholas, D., & Adams, C. D. (1983). The effect of the instrumental training contingency on susceptibility to reinforcer devaluation. *The Quarterly Journal of Experimental Psychology*, *35*(1), 35–51.

Doya, K., Samejima, K., Katagiri, K.-i., & Kawato, M. (2002). Multiple model-based reinforcement learning. *Neural computation*, *14*(6), 1347–1369.

Eldar, E., Morris, G., & Niv, Y. (2011). The effects of motivation on response rate: A hidden semi-markov model analysis of behavioral dynamics. *Journal of neuroscience*

methods, *201*(1), 251–261.

Ferster, C. B., & Skinner, B. F. (1957). Schedules of reinforcement.

Geramifard, A., Walsh, T. J., Tellex, S., Chowdhary, G., Roy, N., How, J. P., et al. (2013). A tutorial on linear function approximators for dynamic programming and reinforcement learning. *Foundations and Trends® in Machine Learning*, *6*(4), 375–451.

Griffiths, T. L., & Tenenbaum, J. B. (2005). Structure and strength in causal induction. *Cognitive psychology*, *51*(4), 334–384.

Gureckis, T. M., & Love, B. C. (2009a). Learning in noise: Dynamic decision-making in a variable environment. *Journal of Mathematical Psychology*, *53*(3), 180–193.

Gureckis, T. M., & Love, B. C. (2009b). Short-term gains, long-term pains: How cues about state aid learning in dynamic environments. *Cognition*, *113*(3), 293–313.

Hammond, L. J. (1980). The effect of contingency upon the appetitive conditioning of free-operant behavior. *Journal of the experimental analysis of behavior*, *34*(3), 297–304.

Hampton, A. N., Bossaerts, P., & O'Doherty, J. P. (2006). The role of the ventromedial prefrontal cortex in abstract state-based inference during decision making in humans. *Journal of Neuroscience*, *26*(32), 8360–8367.

Herrnstein, R. J. (1991). Experiments on stable suboptimality in individual behavior. *The American Economic Review*, *81*(2), 360–364.

Honig, W. K., & Urcuioli, P. J. (1981). The legacy of guttman and kalish (1956): 25 years of research on stimulus generalization. *Journal of the experimental analysis of behavior*, *36*(3), 405–445.

Johnson, M. W., & Bickel, W. K. (2002). Within-subject comparison of real and hypothetical money rewards in delay discounting. *Journal of the experimental analysis of behavior*, *77*(2), 129–146.

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the american statistical association*, *90*(430), 773–795.

Lee, M. D., Zhang, S., Munro, M., & Steyvers, M. (2011). Psychological models of human and optimal performance in bandit problems. *Cognitive Systems Research*, *12*(2), 164–174.

Lewis, S. M., & Raftery, A. E. (1997). Estimating bayes factors via posterior simulation with the laplace?metropolis estimator. *Journal of the American Statistical Association*, *92*(438), 648–655.

Liljeholm, M., Tricomi, E., O'Doherty, J. P., & Balleine, B. W. (2011). Neural correlates of instrumental contingency learning: differential effects of action–reward conjunction and disjunction. *Journal of Neuroscience*, *31*(7), 2474–2480.

McCormick, T. H., Raftery, A. E., Madigan, D., & Burd, R. S. (2012). Dynamic logistic regression and dynamic model averaging for binary classification. *Biometrics*, *68*(1), 23–30.

Niv, Y. (2007). *The effects of motivation on habitual instrumental behavior* (Doctoral dissertation). Hebrew University.

O'Doherty, J. P., Buchanan, T. W., Seymour, B., & Dolan, R. J. (2006). Predictive neural coding of reward preference involves dissociable responses in human ventral midbrain and ventral striatum. *Neuron*, *49*(1), 157–166.

O'Doherty, J. P., Critchley, H., Deichmann, R., & Dolan, R. J. (2003). Dissociating valence of outcome from behavioral control in human orbital and ventral prefrontal cortices.

Journal of neuroscience, *23*(21), 7931–7939.

O'Doherty, J. P., Hampton, A., & Kim, H. (2007). Model-based fmri and its application to reward learning and decision making. *Annals of the New York Academy of sciences*, *1104*(1), 35–53.

Otto, A. R., Markman, A. B., & Love, B. C. (2012). Taking more, now: The optimality of impulsive choice hinges on environment structure. *Social Psychological and Personality Science*, *3*(2), 131–138.

Park, J., & Sandberg, I. W. (1991). Universal approximation using radial-basis-function networks. *Neural computation*, *3*(2), 246–257.

Rachlin, H., Raineri, A., & Cross, D. (1991). Subjective probability and delay. *Journal of the experimental analysis of behavior*, *55*(2), 233–244.

Reed, P. (2001). Schedules of reinforcement as determinants of human causality judgments and response rates. *Journal of Experimental Psychology: Animal Behavior Processes*, *27*(3), 187.

Reed, P. (2003). Human causality judgments and response rates on drl and drh schedules of reinforcement. *Animal Learning & Behavior*, *31*(2), 205–211.

Reed, P. (2011). An experimental analysis of steady-state response rate components on variable ratio and variable interval schedules of reinforcement. *Journal of Experimental Psychology: Animal Behavior Processes*, *37*(1), 1.

Remijnse, P. L., Nielen, M. M., Uylings, H. B., & Veltman, D. J. (2005). Neural correlates of a reversal learning task with an affectively neutral baseline: an event-related fmri study. *Neuroimage*, *26*(2), 609–618.

Reverdy, P. B., Srivastava, V., & Leonard, N. E. (2014). Modeling human decision making in generalized gaussian multiarmed bandits. *Proceedings of the IEEE*, *102*(4), 544–571.

Reynolds, G., & McLeod, A. (1970). On the theory of interresponse-time reinforcement. In *Psychology of learning and motivation* (Vol. 4, pp. 85–107). Elsevier.

Schlinger, H. D., Derenne, A., & Baron, A. (2008). What 50 years of research tell us about pausing under ratio schedules of reinforcement. *The Behavior Analyst*, *31*(1), 39–60.

Schneider, B. A. (1969). A two-state analysis of fixed-interval responding in the pigeon 1. *Journal of the Experimental Analysis of Behavior*, *12*(5), 677–687.

Silberberg, A., Goto, K., Hachiga, Y., & Tanno, T. (2008). Schedule discrimination in a mixed schedule: Implications for models of the variable-ratio, variable-interval rate difference. *Behavioural processes*, *78*(1), 10–16.

Sims, C. R., Neth, H., Jacobs, R. A., & Gray, W. D. (2013). Melioration as rational choice: Sequential decision making in uncertain environments. *Psychological Review*, *120*(1), 139.

Singh, T., McDannald, M., Haney, R., Cerri, D., & Schoenbaum, G. (2010). Nucleus accumbens core and shell are necessary for reinforcer devaluation effects on pavlovian conditioned responding. *Frontiers in integrative neuroscience*, *4*, 126.

Speekenbrink, M., & Konstantinidis, E. (2015). Uncertainty and exploration in a restless bandit problem. *Topics in cognitive science*, *7*(2), 351–367.

Steyvers, M., Lee, M. D., & Wagenmakers, E.-J. (2009). A bayesian analysis of human decision-making on bandit problems. *Journal of Mathematical Psychology*, *53*(3), 168–179.

Sutton, R. S., & Barto, A. G. (1998). *Reinforcement learning: An introduction*. MIT press.

Tanaka, S. C., Balleine, B. W., & O'Doherty, J. P. (2008). Calculating consequences: brain systems that encode the causal effects of actions. *Journal of Neuroscience*, *28*(26), 6750–6755.

Tanno, T., Silberberg, A., & Sakagami, T. (2009). Single-sample discrimination of different schedules'reinforced interresponse times. *Journal of the experimental analysis of behavior*, *91*(2), 157–167.

Tanno, T., Silberberg, A., & Sakagami, T. (2012). Discrimination of variable schedules is controlled by interresponse times proximal to reinforcement. *Journal of the experimental analysis of behavior*, *98*(3), 341–354.

Van den Broek, M., Bradshaw, C., & Szabadi, E. (1987). Behaviour of impulsive and non-impulsive humans in a temporal differentiation schedule of reinforcement. *Personality and Individual Differences*, *8*(2), 233–239.

Watkins, C. J. C. H. (1989). *Learning from delayed rewards* (Unpublished doctoral dissertation). King's College, Cambridge.

Weiner, H. (1969). Controlling human fixed-interval performance 1. *Journal of the Experimental Analysis of Behavior*, *12*(3), 349–373.

Wilson, M. P., & Keller, F. S. (1953). On the selective reinforcement of spaced responses. *Journal of Comparative and Physiological Psychology*, *46*(3), 190.

Yi, M. S., Steyvers, M., & Lee, M. (2009). Modeling human performance in restless bandits with particle filters. *The Journal of Problem Solving*, *2*(2), 5.