Simulating operant conditioning during free behavior using TD(λ) actor-critic reinforcement learning

Ryan Eaton
Final project report
CSE/NeuBeh 528
June 12, 2009

Summary of conditioning experiment (Eaton et. al. 2008): Operant conditioning has been used to investigate neural mechanisms mediating performance of motor tasks that are typically constrained compared to free behavior. To extend this approach to reward activity under unconstrained behavior in the monkey we sought intracranial sites whose stimulation supported operant behavior. A nemestrina monkey was trained to perform an alternating wrist step-tracking task rewarded by applesauce reinforcement. We then tested whether stimulation of intracranial sites in ventral tegmental area and nucleus accumbens (NAc) could maintain the task performance. We found positive sites in NAc and quantified the rate of wrist responses as a function of stimulus intensity, pulse rate and train duration. Parametric curves revealed a systematic variation of response rates as a function of these parameters. Trains of 1 mA, 0.2 ms biphasic current pulses delivered at 50 Hz for 500 ms sustained robust responding. Using the Neurochip, an implanted autonomous recording and stimulating system, we triggered NAc stimuli from pulses generated by muscle activity as the animal moved freely about its cage. Each NAc stimulus pulse was triggered by a large biphasic EMG event, and NAc stimulation was rate-limited at 50 Hz. Activity rates in 100 ms bins were stored to memory for subsequent download. EMG-contingent NAc stimulation was available for intervals of 5 minutes (time-in periods) alternating with 5-min time-out periods during which the stimulator was turned off. Over sessions lasting 20 hours, the rates of EMG activity were dramatically elevated during time-in periods as compared to time-out periods. The monkey was observed to make limb movements activating the reinforced muscle. Immediately after the transition from time-out to time-in periods, pulse rates increased more rapidly for transitions at the end of the 20-hour session than the beginning, indicating a learned association between muscle contraction and NAc stimulation. Immediately following the termination of NAc stimulation EMG rates transiently increased early in the session, but not toward the end, characteristic of extinction behavior. Such activity-dependent stimulation of intracranial reinforcing sites with the Neurochip could be used to operantly condition a variety of neural and motor responses during free behavior.

Project Aims: 1) Simulate contraction rates driving activity-contingent, behaviorally-reinforcing brain stimulation using actor-critic temporal-difference learning that incorporates eligibility traces [TD(λ)]. 2) Examine simulation output for similar conditioning behaviors and learning emitted by the freely-behaving primate (i.e. forming the muscle activity-brain stimulation association, acquisition of the brain stimulation availability schedule).

The model. Given its simulated physical and behavioral state, the model employs a policy to inform decisions to either contract (a = 1), or relax (a = 0) a muscle that could potentially trigger behaviorally-reinforcing brain stimulation. The policy, an array of preference values—one for each state-action pair—guides a stochastic choice of action depending upon the model's current state. The learning algorithm uses information from state vectors to update a policy, or strategy, to maximize future reward. For this model, state vectors have three elements. The first indicates that the muscle is relaxed, ($s_t^{(1)} = 0$) or contracted ($s_t^{(1)} = 1$) at the current time step, t. The second element, $s_t^{(2)}$, is a binary

indicator of stimulus pulse delivery. A coarse appraisal of muscle fatigue level comprises the third element. It may take one of four values: none, low, high or severe.

The possible values of each state vector element implies the model has a total of $2\times2\times4 = 16$ states in its state space. Because brain stimulation can only be triggered by muscle contraction, four of the sixteen possible states are inaccessible (see state definitions table in appendix). Actor-critic TD(λ) learning is most easily implemented for a finite number of discrete states and actions. The above-described behavioral paradigm (Eaton et. al. 2008) lends itself to discrete formalism. The presence (or absence) of a stimulus pulse at each time step is a binary event. A continuous electromyography signal either passes through an appropriately calibrated discrimination window (or it doesn't)—again binary. An animal's appraisal of its own fatigue is more qualitative than quantitative, perhaps bracketed fatigue levels suffice for this purpose as a continuous measure would over-estimate the precision of the animal's perception (St Clair Gibson et. al. 2003).

In the actor-critic TD(λ) learning algorithm, step rate parameters α and β scale the amount by which the expected reward function V(s) and preference array P(s,a) are modified in proportion to δ, the error in predicted reward. The temporal discount-rate parameter γ determines the present value of future rewards: a reward received k time steps in the future is worth only γ^{k-1} times what it would be worth if it were received immediately (Sutton and Barto 1998). Eligibility traces $e_C(s)$ and $e_A(s,a)$.can be viewed as a temporary record of the occurrence of an event—they are a basic mechanism for temporal credit assignment. Parameter λ determines the rate by which these temporary records decay. Parameters $\alpha = 0.2$, were set for the simulation described below. This set was chosen to make TD(λ) learning gradual but with memory of past experience via slowly decaying eligibility traces. Actions were selected at 200 Hz over each episode during which contraction contingent stimulation was available during the first 5 minutes, and not available for the remaining half.

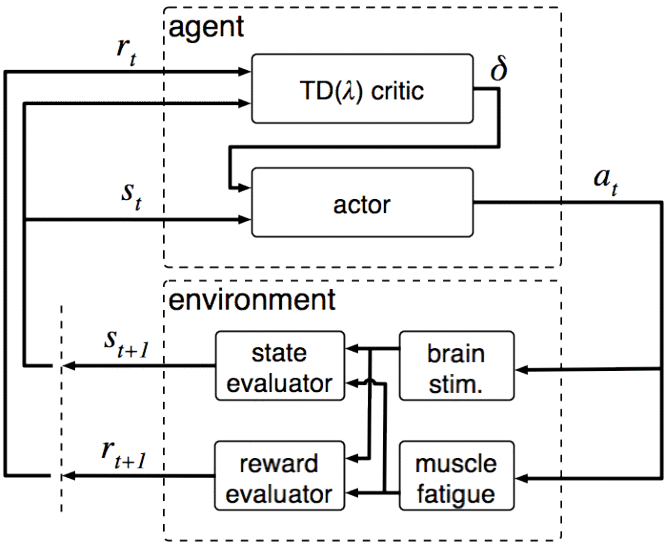


Figure 1. Schematic of agent-environment closed-loop interaction during simulated muscle conditioning. The particular state and reward experienced at time step t are denoted s_t and r_t respectively. The error in expected future reward (δ) after choosing action a_t leading to state s_{t+1} and reward r_{t+1}, scales the update of expected reward assigned to state s_t as well as the preference for choosing action a_t in state s_t. The vertical dashed line illustrates the transition to next iteration for which variables s_{t+1} and r_{t+1} of the previous run are assigned to s_t and r_t of the next.

Selecting the reward function. Previous findings indicate rate of responding increases to an asymptote with repetitive stimulus trains of increasing frequencies of re, behaviorally-reinforcing, brain stimulation (Mora et al. 1979, Eaton et. al. 2008). Responding extinguishes following stimulus train cessation. To generate a continuous reward signal that accumulates with each stimulus pulse, a kernel of appropriate shape can be convolved with the pulse train. A suitable kernel would generate a rise to saturation when convolved with a pulse train of sufficient frequency and then gradually decay down to baseline after the train terminates. Although any function with a rising phase followed by a falling phase would satisfy these criteria, the alpha function (in the domain $t > 0$) is familiar and can be fully characterized by just two parameters: time to function peak and a scaling constant.

Heuristic knowledge suggests muscle fatigue severity depends upon the intensity and duration of the contracting muscle. Also, that recovery from fatigue (fatigue-decay) occurs during rest periods following contractions. Similar to the brain stimulation reward function, a continuous fatigue penalty function meeting the above requirements can be attained by selecting an alpha function as the kernel.

Behaviorally-reinforcing brain stimulation increases the likelihood that the animal re-emits the operant with which it is paired—in this case muscle activity. In contrast, muscle fatigue is generally unpleasant and likely negatively reinforces its associated operant. The net effect of the two (their difference) might be a generalized reward signal

to inform a policy governing muscle contraction as to maximize expected rewards in this paradigm.

The relative widths and amplitudes of the brain stimulation and muscle fatigue reward kernels were tailored to reward low-frequency muscle activity and negatively reinforce high-frequency bursts. The t_{peak} of the muscle fatigue alpha function was 2.5 times the value of the brain stimulation kernel but only one-quarter the amplitude. Fatigue accumulated and decayed slower than reward from brain stimulation. Successive brain stimulation pulses were rate-limited at 50 Hz when triggering contractions surpassed this frequency. Fatigue surpassed brain stimulation for contraction rates greater than 50 Hz, resulting in negative net reward.

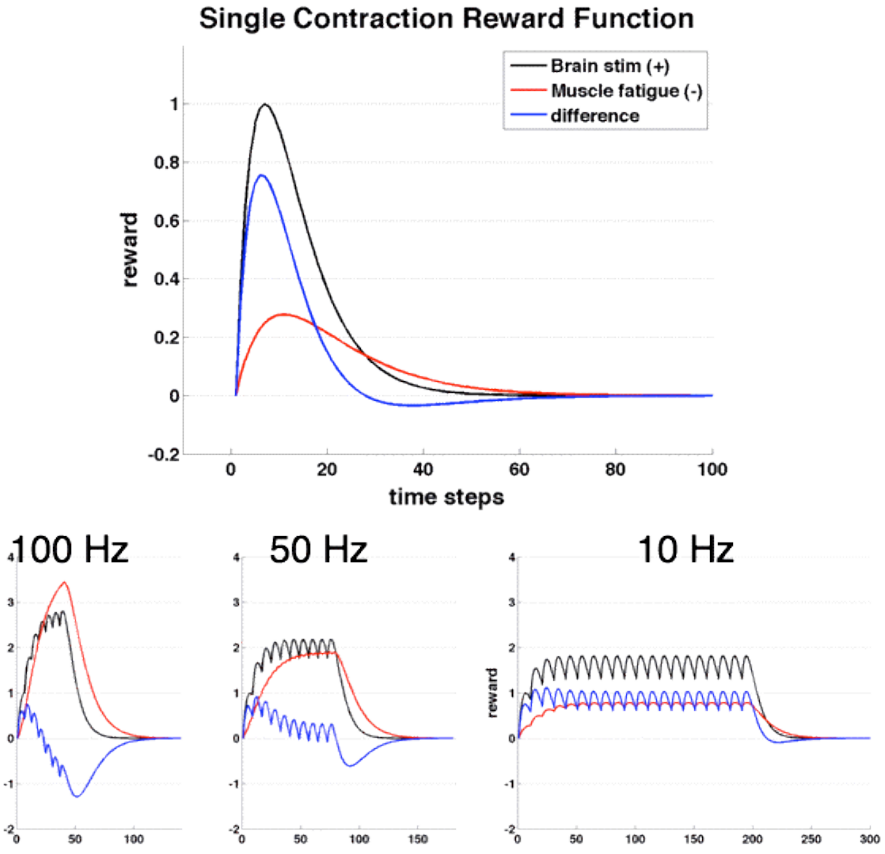


Figure 2: Reward function chosen to positively reinforce low-frequency contraction rates and negatively reinforce high-frequency bursts. Brain stimulation reward profile accumulates with each stimulation pulse up to, but not above, 50 Hz. Muscle fatigue reward profile accumulates with each and every contraction. Subtracting the fatigue profile from the stimulation profile generates a net reward level received by the agent. Muscle fatigue level—the third element of the state vector--is assigned to one of four ascending discrete domains: "none", "low", "high" or "severe". For contraction rates greater than 50 Hz, not every contraction event triggers a pulse of brain stimulation—this permits the reward magnitude from muscle fatigue to surpass that from brain stimulation. The net reward level is negative (e.g. train of 20 contractions at 100 Hz).

Results/Discussion:

Five-minute mean contraction rates, and mean simulated rates during time-in periods had similar ranges (20 to 60 Hz) over the course of the 20-hour conditioning session (figure, 3). However, physiological mean contraction rates during time-out periods were much more variable than mean rates of simulated contractions that did not trigger stimulation. Monkeys exhibit a variety of active behaviors overnight as they move about their cages. Variability in time-out muscle activity is likely owed to other unmonitored behaviors. The model did not account for these, and are likely culprits for long periods of relative inactivity (e.g. sleep) observable in the in vivo record but not the simulation.

The lowest rate of rise in activity following off-to-on transitions occurred during first third of the sessions for both the monkey and the simulation (figure 4). Activity accelerates greatest for the last third of the sessions for both animal and model as well. Animal and simulation records differ in that across session slopes in pooled physiological muscle activity span a much broader range, and begin to rise much earlier relative to off-to-on transitions than for simulated contraction rates.

Immediately following on-to-off transitions, both physiological and simulation contraction rate averages show a sharp, transient peak in responding. In addition, the height of the peaks (as measured from pre-transition contraction rates) as well as their durations are quite comparable—on the order of 30 Hz and 3 seconds respectively. However, on-to-off post transition peaks in activity means decayed as muscle conditioning progressed while post-transition peaks actually increased, in both amplitude and width, over the course of the simulation.

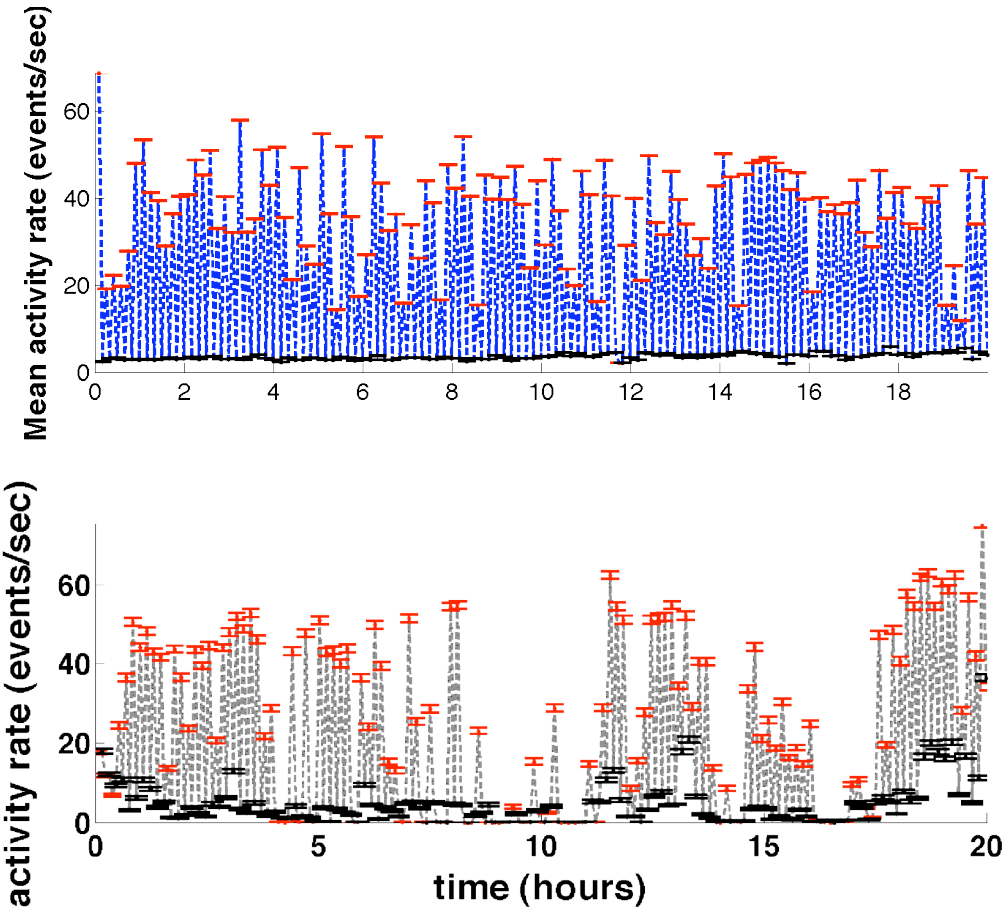


Figure 3: Simulated (top) and actual (bottom) mean activity rates reinforced by contraction-contingent brain stimulation in a freely behaving macaque. Each data-point indicates mean EMG activity over 5 minutes. Contraction-contingent current pulses delivered to nucleus accumbens during time-in periods (red, bottom). Pulses driving the simulated reward signal positive occurred during time-in periods in the top plot. During time-out periods (black), simulated and *in vivo* brain stimulation was not delivered.

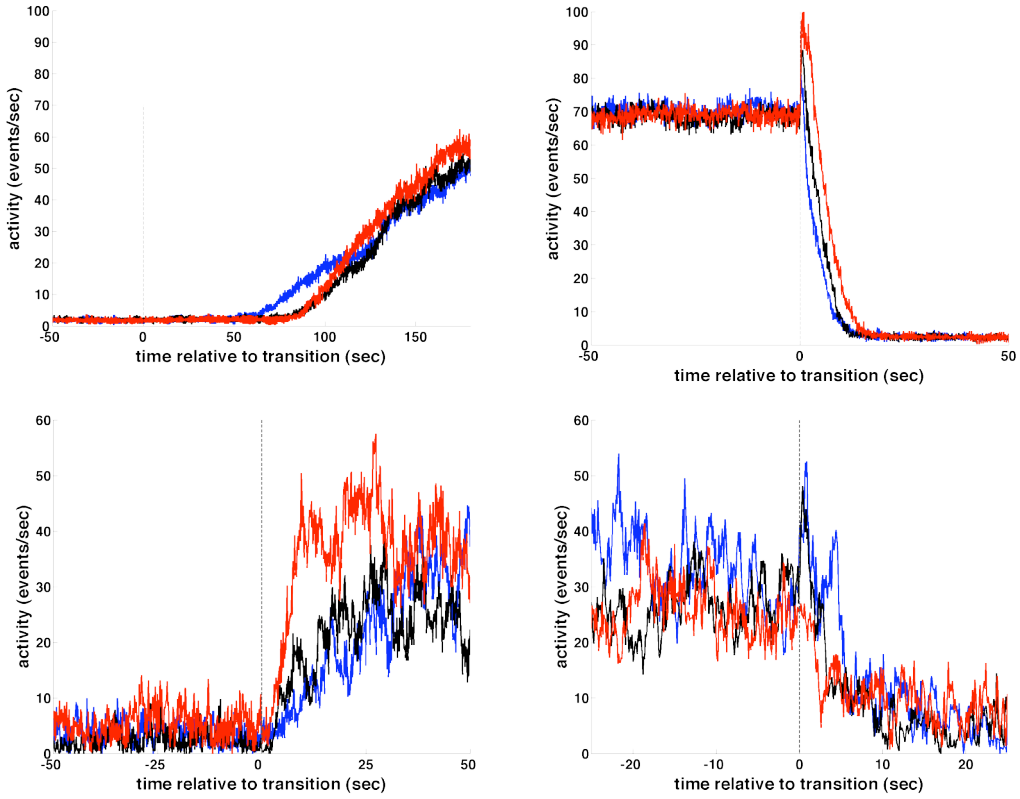
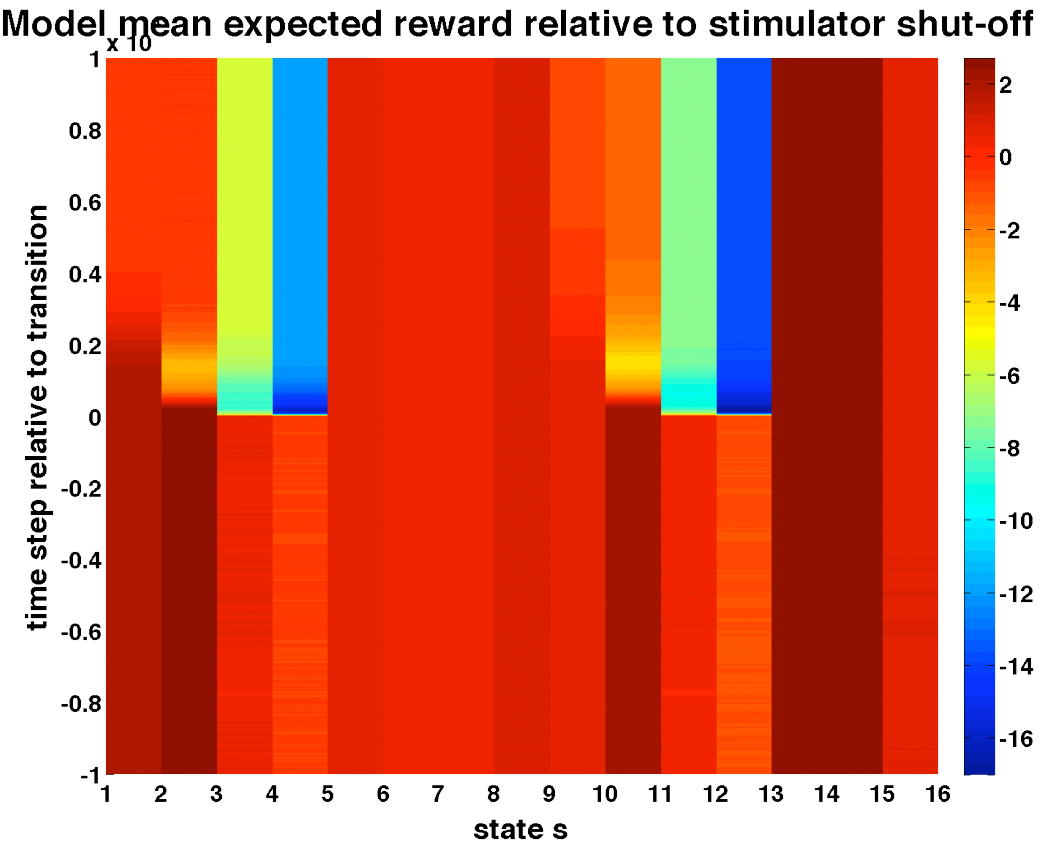
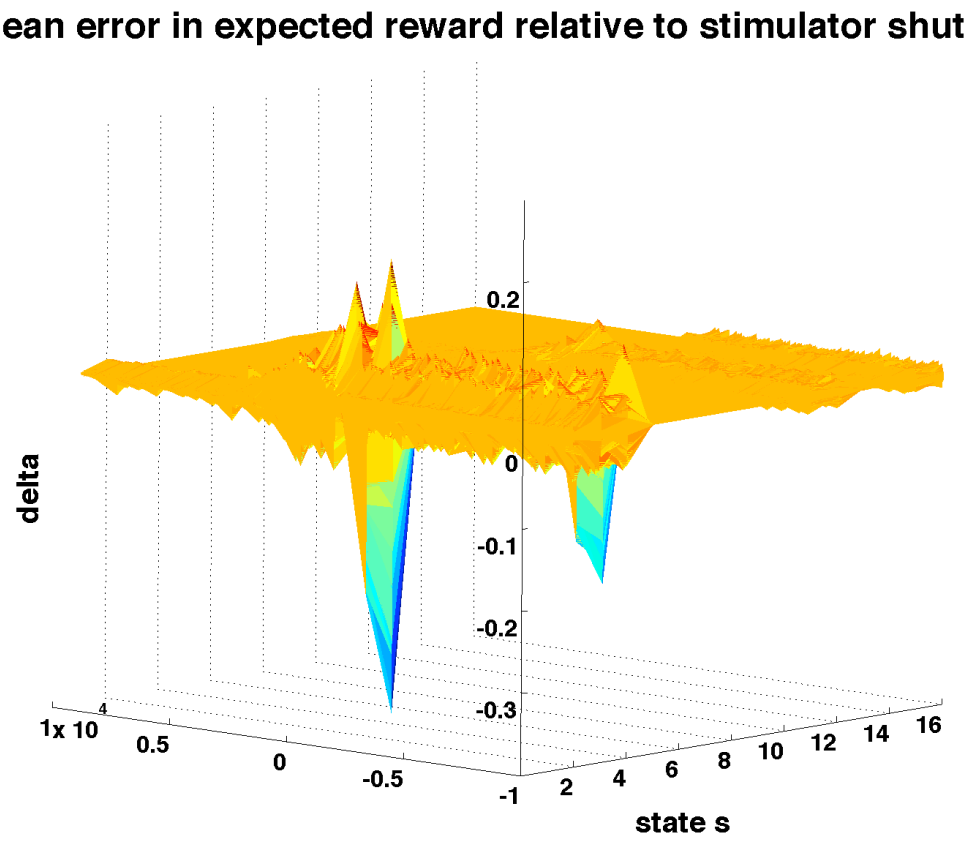


Figure 4: Averages of bicep EMG rate for transition from stimulator off-to-on (bottom, left) and on-to-off (bottom, right). Corresponding peri-transition averages of simulated contraction rates (top, left and right). The off-to-on stimulator transitions above illustrate that muscle activity, both simulated and physiological, rose more rapidly in the average of the last third of the transitions *(red)* compared to the average profiles from the first third *(blue)* and middle third *(black)*. This suggests the animal/simulation learned to associate biceps activity with reinforcing brain stimulation over the course of the 20-hour conditioning session. The initial third of on-to-off stimulator transitions show a transient increase in muscle activity immediately after end of reinforcement *(arrow)*, characteristic of extinction behavior. This post-transition peak disappeared in the last third of the session in the physiological records, but grew taller and wider in the simulation.



The above figure shows peri-transition expected reward averages across states and intra-episodic time. It is compiled using records from the 120 simulated time in-to-time out transitions. Notice the abrupt drop in expected reward for states 4 and 12—both severe-fatigue, pulse-absent states (see appendix). Immediately following effective stimulator shut-off, the sudden absence of reinforcing brain stimulation no longer balances the penalizing effect of muscle fatigue. The abrupt drop in preference for the state-action pair just preceding the transition affects a large change in reward expectation, and hence the policy, as the error in expected reward at the transition step is substantial (see plot below). The abrupt policy change that results likely causes the transient increase in post on-to-off transition responding observed in peri-transition averages of simulated contraction rates. The precise mechanism requires further investigation.

I tailored this TD(λ) actor-critic reinforcement learning model to mimic *mean* contraction rates paired with an alternating five-minute schedule of contraction-contingent behaviorally-reinforcing brain stimulation. Surprisingly, the model, while designed to *describe* animal reinforcement learning over the course of hours, roughly *predicts*, post on-to-off responding over periods on the order of 10 seconds or less. This suggests actor-critic TD(λ) learning may correlate with more fundamental physiological mechanisms underlying animal learning and behavior.



Lastly, choice of step and reward parameter sets generating the above simulation findings were by no means optimal; results shown here were generated using the fourth intelligently-chosen, but not refined, set of parameters that I investigated. In future work I will attempt to broaden the range of across session slopes following peri-off-to-on transitions and explore the precise mechanism generating the post on-to-off transition peak in simulated contraction rates.

References:

Eaton, R.W., Zanos, S.P., Fetz, E.E., (2008) Intracranial reinforcement of operants during controlled and free behavior in the primate. Society for Neuroscience Annual Meeting, Abstract/Poster 878.4

Mora F, Avrith DB, Phillips AG, Rolls ET. Effects of satiety on self-stimulation of the orbitofrontal cortex in the rhesus-monkey. Neuroscience Letters. 1979;13(2):141-145.

St. Clair Gibson A, Baden DA, Lambert MI, Lambert EV, Harley YXR, Hampson D, Russell VA, Noakes TD. The Conscious Perception of the Sensation of Fatigue. Sports Medicine 2003: 33(3) 167-176

Sutton, R.S., Barto, A.G., (1998) Reinforcement Learning: An Introduction. The MIT Press

Appendix:

State vector definitions:

	1	2	3	4	5*	6*	7*	8*	9	10	11	12	13	14	15	16
Muscle contracted?	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1
Pulse delivered?	0	0	0	0	1	1	1	1	0	0	0	0	1	1	1	1
Fatigue level	0	1	2	3	0	1	2	3	0	1	2	3	0	1	2	3

Fatigue level numbers defined: 0 = "none", 1 = "low", 2 = "high" and 3 = "severe"
*state vector not permitted by environment.

Intra-episodic actor-critic TD(λ) learning algorithm:

(adapted from On-line tabular TD(λ) and Watkins's Q(λ) algorithms found in Sutton and Barto, 1998)

Call from previous episode V(s), P(s,a), $e_C(s)$ and $e_A(s,a)$ for all $s \in \{s\}$ and all $a \in \{a\}$
where {s} is the set of all states and {a} the set of all actions.
Repeat (for each time step of current episode):
a ← action chosen by policy π = softmax(P(s,a)) for state s.
Take action a, call on environment(s,a) to generate reward, r, and report next state s'.
$\delta \leftarrow r + V(s') - V(s)$. [Error in predicted reward]
$e_C(s) \leftarrow e_C(s) + 1$. [Bump eligibility trace corresponding to current state s]
$e_A(s,a) \leftarrow e_A(s,a) + 1$ [Bump eligibility trace corresponding to current state-action pair].

For all states in {s}:
$V(s) \leftarrow V(s) + \alpha\delta e_C(s)$. [Update critic or reward evaluation]
$e_C(s) \leftarrow \gamma\lambda e_C(s)$. [Update critic eligibility traces]

For all actions in {a}:
$P(s,a) \leftarrow P(s,a) + \beta\delta e_A(s,a)$. [Update state-action pair preferences]
$e_A(s,a) \leftarrow \gamma\lambda e_A(s,a)$. [Update actor eligibility traces]

s ← s', store s' and a in memory.

End at next to final time step as the next state s' has already been chosen.

Make final state s accessible for next episode.