PSY 2350R: Laboratory on Reinforcement Learning and Decision Making Semester Report

Aneesh Muppidi
aneeshmuppidi@college.harvard.edu

December 16, 2023

1 TD Learning and Value RNNs (First Half of the Semester)

Objective

The primary goal for the first half of the semester was to computationally replicate the ”Rapid Learning of Odor–Value Association in the Olfactory Striatum” experiment and to develop computational models to estimate the value of a cue post-onset observation.

Initial Setup and Experiment Class Creation

- **Experiment Class Creation:** I developed an experiment class, which allowed for customization of various parameters such as number of trials, observations, rewards, reward delays, intervals between cues, and total number of cues.

- **Visualization Functions:** I implemented functions to create visual representations of the experimental data, which were helpful for interpreting the experiment's dynamics.

- **Controlled Experiment Initiation:** I set up a controlled experiment where the first half included original cues, and in the second half, two original cues were replaced with new ones.

Model Development and Analysis

- **CSC Representation Model:** The first model I explored was based on the Complete Serial Compound (CSC) representation. In CSC, every time step following stimulus onset is represented as a separate feature. This approach allowed us to model the internal representation of value associated with every time step after observing a cue, with values varying depending on the time elapsed and our gamma value. This model was pivotal in establishing a baseline for our experiments, showing that the learning of novel cues was not faster than that of original cues.

- **Shared CSC Model:** Following the initial model, I proposed a model where all cues shared the same CSC representation. However, this model did not perform as expected and I didn't have enough time to fully develop it.

- **Exploration of ValueRNN:** Jay and I were particularly interested in whether the ValueRNN, would exhibit faster learning for new cues (Figures 1, 2). In this model, I varied the number of hidden units to observe their effect on learning. We found that with too few hidden units, such as 2, the network showed worse convergence of the mean squared RPE for predicting value compared to the baseline. However, beyond a certain threshold of hidden units, the model demonstrated faster learning for new cues, highlighting the potential of the ValueRNN in our experimental setup.

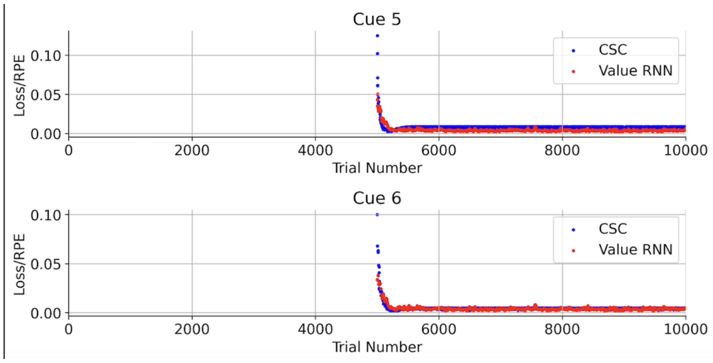


Figure 1: A direct comparison of the RPEs for cues 5 and 6 (the new cues introduced during the second half of the experiment) during learning by the CSC representation model and the ValueRNN. We see that the ValueRNN starts off with a much lower loss than the CSC representation model.

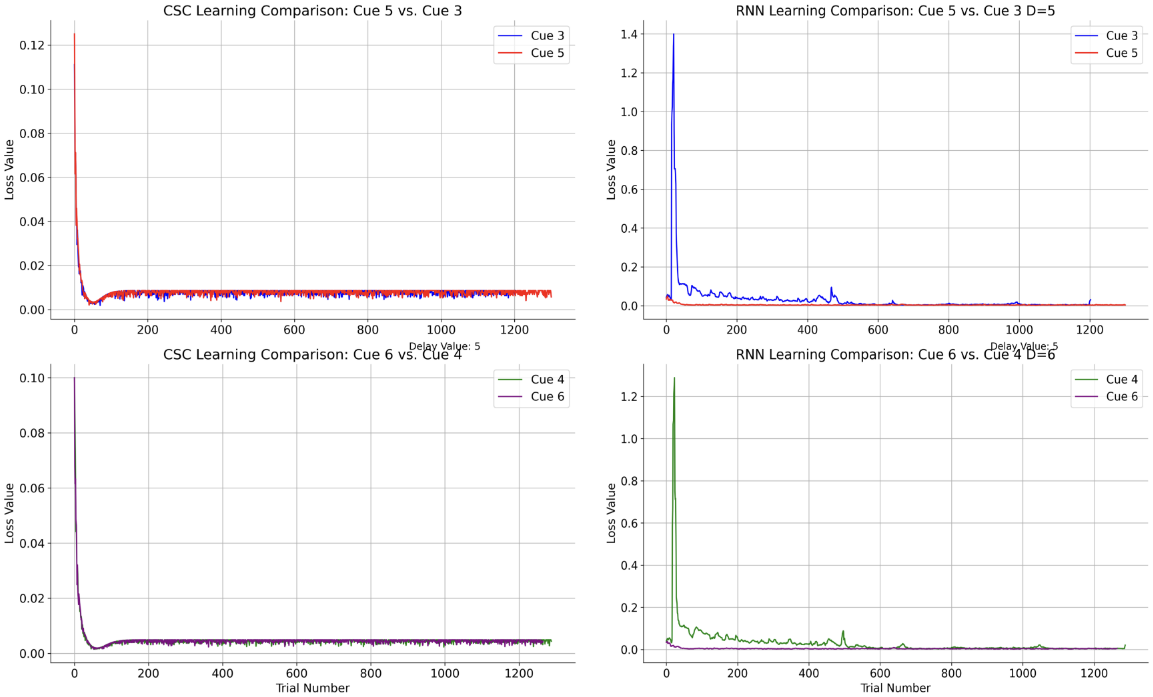
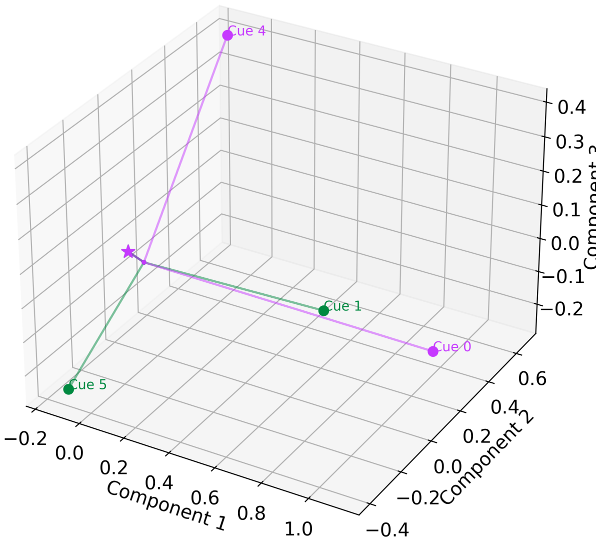


Figure 2: A direct comparison for both the CSC representation model and the ValueRNN model in the Loss/RPE learning for the original cues 3 and 4 in the first half the experiment compared to learning the new cues 5 and 6. We see that the ValueRNN learns the new cues much faster than the original cues, but this is not true for the CSC model.

Advanced Visualizations

Building on Jay's initial work on visualizing hidden unit activity trajectories, I advanced this by visualizing the hidden unit trajectories corresponding to each cue using PCA, as seen below. I did this for before and after the RNN learned the novel cues. This provided a deeper understanding of the model's learning dynamics.



Learnings and Contributions

Throughout this semester's first half, my key learnings included:

- Implementing TD learning from scratch.
- Computationally representing experiments.
- Developing and applying an RNN in Pytorch for TD learning.
- Understanding the relationship between the ValueRNN's loss function and the TD error.
- Visualizing experiments and RNN activity.
- Gaining insights into CSC representations.

This period was instrumental in demonstrating that the ValueRNN could effectively showcase rapid learning. My contributions primarily involved implementation and exploration of these models, while Jay provided guidance on broader research questions. I also engaged in formulating my own models such as the shared CSC representation, and other specific questions on how we can measure ”how fast” a model learns.

2 Q-learning and Go/No-Go Task: Striatum vs Cortex (Second Half of the Semester)

Transition to Q-learning

Following our discussions with Sam, we decided to explore rapid learning mechanisms in Q-learning and investigate potential neural correlates in the experimental data. Our model projected input to an internal representation through a W matrix, and then from this internal representation, it projected values for two actions through a Q matrix. We aimed to test two hypotheses with this model, focusing on the dimensions of the internal representation and the relative importance of learning in the W and Q matrices.

Hypotheses

- **Pre-Learned Q Matrix Hypothesis:** We hypothesized that a Q matrix pre-adapted to cues 1-4 would facilitate faster learning of new cues 5 and 6.
- **Role of W Matrix Hypothesis:** We posited that the W matrix, which maps cues to their representations, could efficiently leverage the existing structure in the Q matrix for new mappings of cues 5 and 6.

Implications of the Hypotheses

1. **Transfer of Learning:** Faster learning of cues 5 and 6 compared to 3 and 4 would indicate the agent's capability for transfer learning.
2. **Efficiency of Q Learning Modification:** Successful learning of new cues with a frozen or slowed-down Q matrix would suggest an efficient and adaptable learning process.
3. **Mapping Efficiency of W:** Quick mapping of new cues to an established Q matrix by the W matrix would demonstrate the agent's adaptability with minimal alteration to its learned strategy.

Neural Correlates and Model Implementation

The cortex's role in transforming sensory inputs into abstract representations for decision-making is mirrored in the model by 'z = Wx'. This transformation represents the mapping of external cues to internal representations. The striatum is represented in the model by 'v = Qz'. This step reflects the use of abstract representations from the cortex to evaluate actions and outcomes.

OpenAI Gym Implementation

To accommodate the Go/No-Go task, I set up the "Rapid Learning of Odor–Value Association in the Olfactory Striatum" experiment in an OpenAI Gym environment. This allowed for more precise control over the experiments. Jay suggested the need for such an environment, and I expanded on this idea by customizing the gym environment with a single seed value. This ensured that for any given seed, all cues (rewarded or not), the order of cues, each trial, and the exploration and exploitation parameters remained consistent, enabling reproducible results.

Exploration into Hierarchical Bayesian Meta-learning

As our project progressed, Jay and I delved into the potential connections between our work and hierarchical Bayesian meta-learning, considering the cues in our experiments as different tasks. To deepen our understanding, I conducted a literature review, focusing on key papers like "Learning Overhypotheses with Hierarchical Bayesian Models" and suggested a joint review with Jay on "Recasting Gradient-based Meta-learning As Hierarchical Bayes" for further insights.

Reflections and Insights

During this exploration, we faced a realization regarding the nature of our task and its alignment with hierarchical Bayesian principles. We noted a fundamental distinction between probabilistic estimation and policy learning. In our context, whether the Q-value for the correct action was high (e.g., 1,0) or moderately high (e.g., 0.6,0.4), the chosen action would remain the same. This observation led to the understanding that in such a simple task framework, the variations in Q-values did not significantly impact the speed of learning new cues or alter the policy substantially when a new cue was introduced.

Future Directions

Given these insights, we concluded that directly relating our work to the concept of learning overhypotheses in the realm of hierarchical Bayesian meta-learning would be more challenging than initially anticipated. This realization prompted us to decide that this line of inquiry, while intriguing, would need to be revisited in the future.

Testing Projection Models with Different Presets

Initially, some experimental runs indicated faster learning of a novel cue compared to an old cue when we slowed the learning rate of the Q matrix or froze it. Jay and I were initially enthusiastic about these findings. However, I harbored skepticism regarding their generalizability. To validate these initial observations, I conducted a more extensive analysis, running the experiment 100 times with 100 different seeds. The comprehensive analysis of these runs revealed that the initial findings were outliers. In response to these insights, Jay suggested trying different presets, such as freezing the model and comparing the impact of altering the Q and W matrices. I sought a more systematic approach to analyze the results and, therefore, developed six different presets along with a rigorous testing pipeline.

Presets Description

- **OnlyWPlastic:** The W matrix is plastic (adaptive), with the Q matrix's learning rate set to zero (alpha q = 0), allowing no changes in the Q matrix after new cues are introduced.
- **OnlyQPlastic:** The Q matrix is plastic, with the W matrix being an identity matrix (alpha w = 0), allowing no changes in the W matrix.
- **QWPlastic:** Both W and Q matrices are plastic, with no freezing of the Q matrix after new cues are introduced.
- **OnlyWPlasticFreeze:** The W matrix is plastic (alpha q = 0), with freezing of the Q matrix after new cues are introduced.
- **OnlyQPlasticFreeze:** The Q matrix is plastic (alpha w = 0), with freezing of the Q matrix after new cues are introduced.
- **QWPlasticFreeze:** Both W and Q matrices are plastic, with freezing of the Q matrix after new cues are introduced.

Testing Pipelines and Observations

The testing involved two primary pipelines:

1. A comparison of each preset for a single cue at various learning rates.
2. A comparison of learning for two cues (an old cue and a novel cue) across every preset at different learning rates, focusing on the RPE over trials.

Findings from the Second Pipeline: We observed that models heavily dependent on changing W and resistant to changes in Q (either through freezing after the first half or a low learning rate) did not demonstrate rapid learning of novel cues. This was evident from the RPE trends over trials for novel rewarded and non-rewarded cues introduced in the second half of the experiments, as compared to an old rewarded or non-rewarded cue's RPE (Fig 3)

Observations from Pipeline 1: When testing presets against each other for a specific novel cue's RPE, several trends emerged at various learning rate combinations. These trends suggested the differing importance of the Q and W matrices under different conditions. Notably, learning the Q matrix appeared more crucial than learning the W matrix or both in most scenarios.

Conclusion: The results from our extensive testing suggest that a malleable Q matrix is more important than a malleable W matrix for rapid learning of novel cues, offering little support for our original hypothesis.

Impact of Reduced Latent Internal Representation Dimensionality

An integral aspect of our research involved exploring how varying the dimensionality of the internal representation (z) affects learning. Typically, our baseline involved matching the dimensionality of z with the number of input cues. However, we experimented with reducing this dimensionality to understand its impact on the learning process.

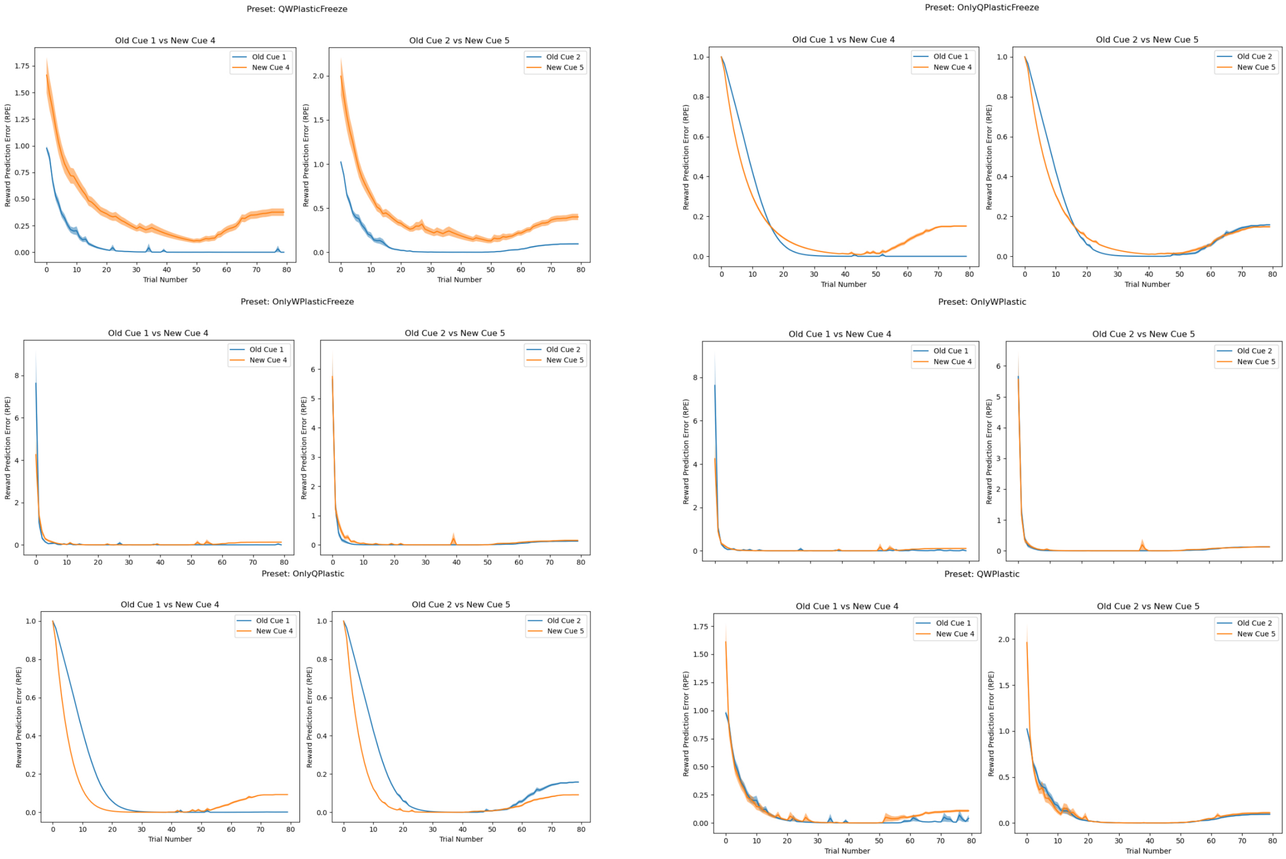


Figure 3: An example of plots generated for one learning rate configuration for comparing the learning of a novel cue vs. an old cue under different presets.

Experiment Setup and Observations

In one key experiment, we reduced the latent representation dimensionality to three, which is half the number of our input cues. We observed the effects of this reduction on the rate of learning new cues for every preset.

Findings: For both the OnlyWPlastic and QWFreeze presets, the rate of learning a new cue was either equivalent to or faster than our baseline scenario. This was clearly demonstrated in our results, as depicted in Figure 4. However, this was not the case for the other presets. We observe that this result is only seen when the dimensionality of our internal representation was atleast 3, and not any lower.

Implications of Reduced Dimensionality

A reduced dimensionality of the internal representation does not necessarily hinder the learning process. In some cases, it may even facilitate more efficient learning, possibly due to a more streamlined and focused representation of cues. A notable similarity between the QWFreeze and OnlyWPlastic presets is the inhibition of the Q matrix; in QWFreeze, this occurs halfway through the experiment, while in OnlyWPlastic, the Q matrix is consistently inhibited. This shared characteristic suggests that when limiting the adaptability of the Q matrix, either partially or throughout the experiment, a reduced dimensionality of internal representation can lead to faster learning outcomes.

Learnings and Insights

Throughout the course of this project, my understanding of RL, particularly Q-learning, has deepened significantly. The hands-on experience and the challenges I encountered provided me with invaluable insights and learnings.

Implementing Q-Learning Agent from Scratch

The most profound learning experience came from manually implementing the Q-learning agent. This process, starting from the one-hot encoding of the input, projecting through the W matrix, and then to the Q projection, was intricate. Particularly, understanding how gradient descent interacts with these matrices and conducting the forward and backward passes by hand for debugging was pivotal. Implementing this projection model from scratch, without relying on existing libraries, was instrumental in developing an intuitive grasp of the mechanics behind the model and recognizing what was effective and what wasn't.

Utilizing Gym for RL Experiments

Another significant learning was the use of Gym for RL experiments. This platform proved to be invaluable for its versatility and precision in experiment control, making it an indispensable tool in my research. I am glad I discovered how to use this tool during this semeseter, and will continue to use it for future research.

Robust Testing of Models

I learned the importance of robustly testing models and that a single experimental run is insufficient to draw concrete conclusions. This understanding led to more comprehensive and reliable research methodologies, involving multiple runs and varied parameters to ensure the validity of our findings.

Literature Reviews and Meta Learning

Engaging in literature reviews expanded my knowledge in areas such as meta-learning, overhypotheses, and various perspectives on rapid learning. These reviews not only provided theoretical underpinnings but also inspired new ideas and approaches in our experiments.

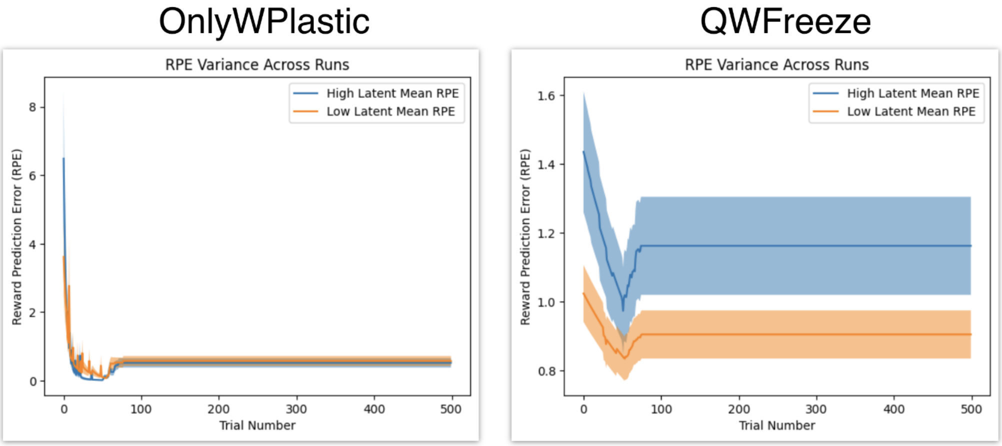


Figure 4: RPE for a learning a novel cue for a model with a 6-dim internal representation (blue) versus a model with a 3-dim internal representation (orange).

Development of New Presets and Neural Correlates

The development of new presets and the exploration of their neural correlates were also a part of my learning journey. This process enhanced my ability to conceptualize different models and understand their potential neural underpinnings.

Visualization and its Importance

Finally, I learned the importance of visualizing results. Effective visualization techniques were crucial in making complex data comprehensible and in communicating our findings clearly and effectively.

Acknowledgements

I would like to express my sincere gratitude to my mentor, Jay, for his invaluable guidance, support, and the generous allocation of his time throughout this semester. Jay's mentoring style, which encouraged me to ask pertinent questions and engage in critical thinking, was instrumental in my growth during the semester. He created a learning environment where I felt comfortable and never judged, no matter the nature of my questions. This atmosphere of openness and respect was pivotal in my development as a researcher.

Moreover, Jay's advice on effectively visualizing our findings significantly enhanced the presentation and impact of the work. His profound knowledge in TD learning and RL provided clear insights and guided me towards relevant resources and perspectives.

Furthermore, Jay's enthusiasm for exploration and intellectual discussion made our weekly meetings not only productive but also immensely enjoyable and enlightening.

Code and Research Logs

My code and research notebook/logs/notes can all be found at this repo