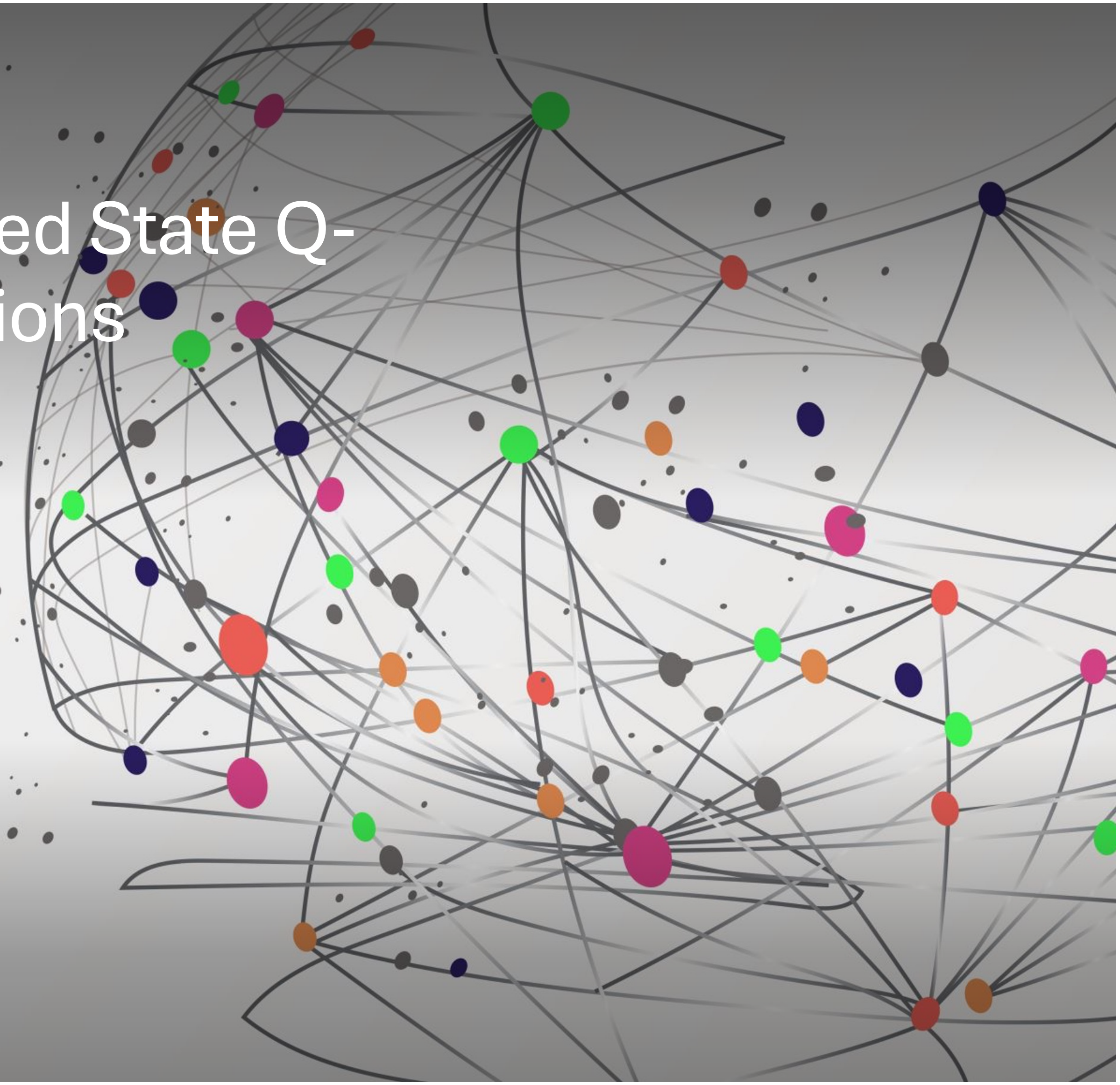
Discrete Muti-addicted State Q-agents Making Decisions

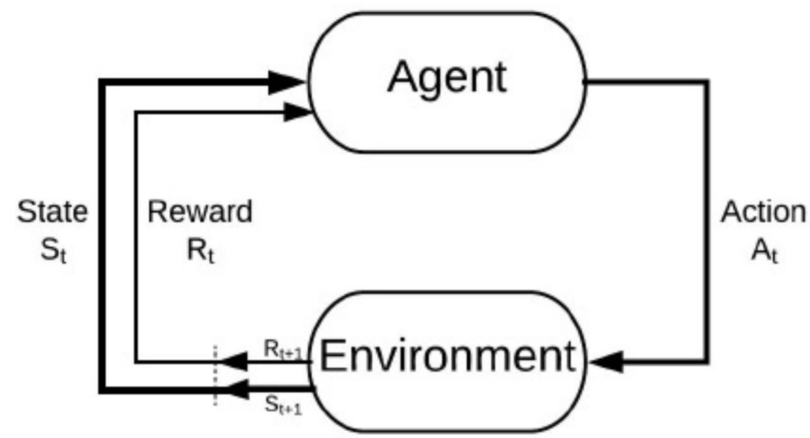


By Kaiwen Bian

What & Why?

To what extend does a monotonic decreasing dopamine surge function effect a Q-agent’s ability to make decision under a discrete chain multi-addiction states setting?

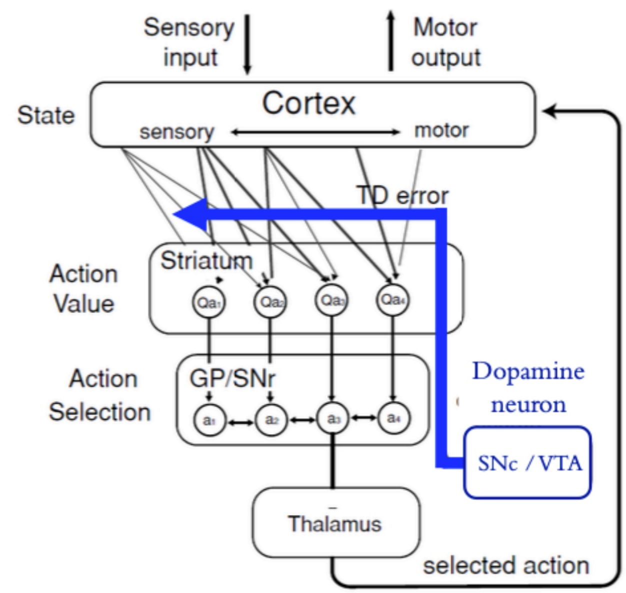
1. Action matters, different stage dictates decision.
2. Search strategies matters, more rewards presented, current best may be deceptive.
3. Discrete & simple -> mathematical insights to the algorithm



Background on TD/Dopamine

Striatum incorporates environmental state from the cortex + dopamine reward prediction error signal from the VTA -> adjusts the weights on action selection back to the cortex, influencing movements.

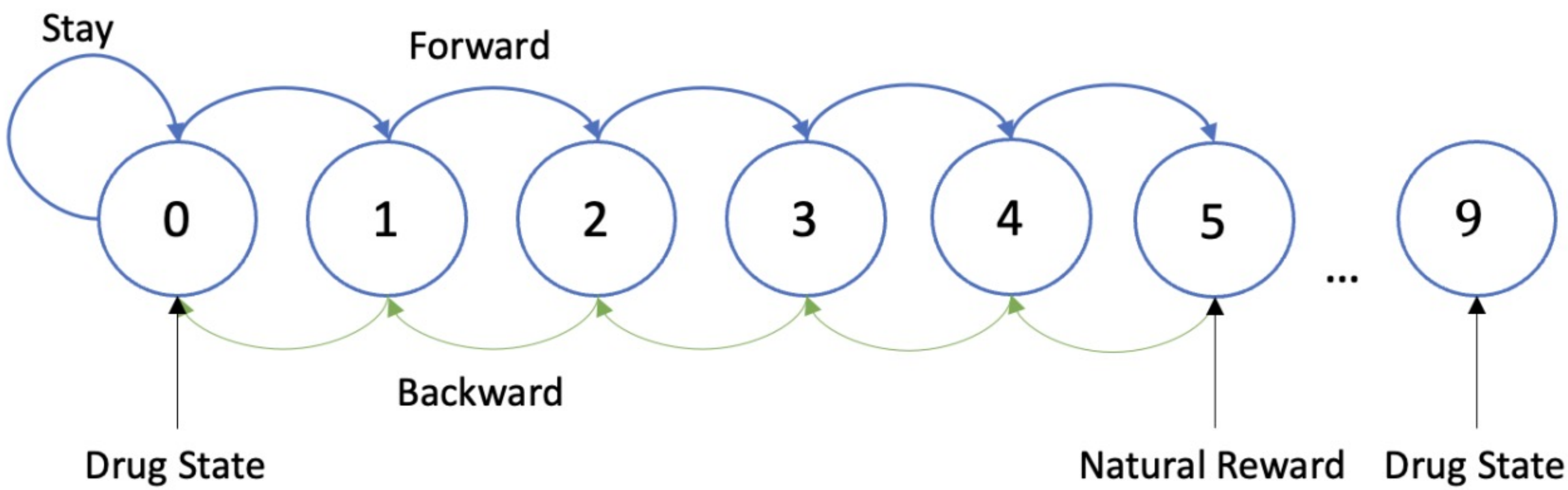
- TD error have been thought to do such job.
- Dopamine does not encode pleasure, rather expectancy.
- Dopamine surge = TD error surge



Methods & Setup

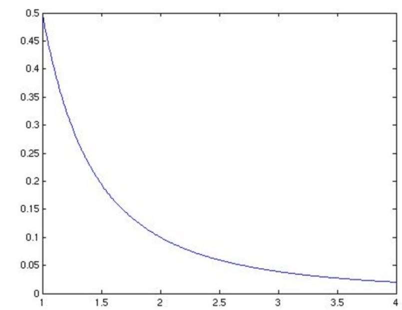
Environment Chain Setup

- Greedy, epsilon greedy, and Boltzmann exploration strategies
- Two addicted states + natural reward state setting
- Addicted or not (dopamine surge flag) setting
- Stochastic processes (random walk) modeling & comparison



Addicted Q-agent

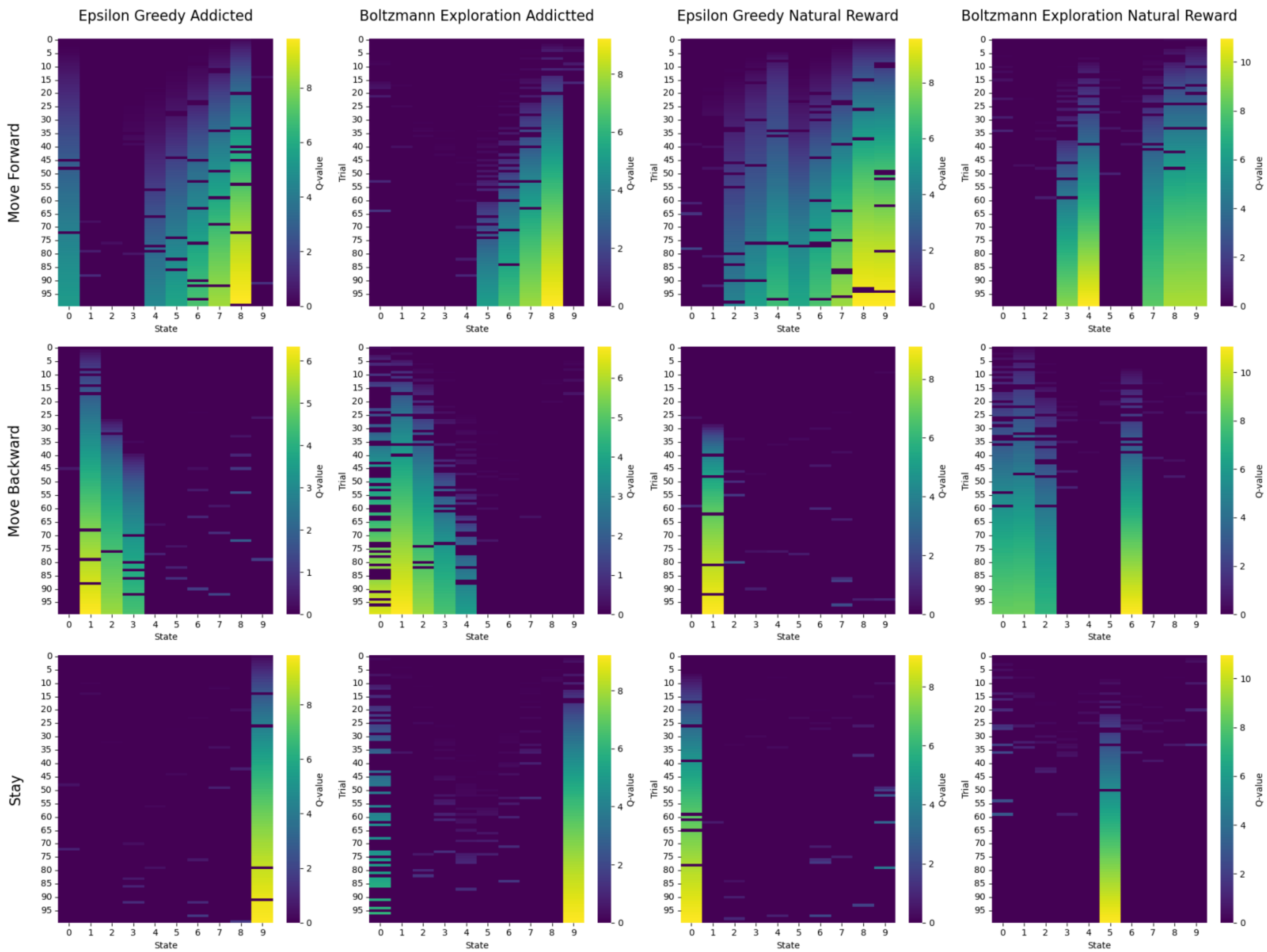
$$Q(s_i, a_i) \leftarrow Q(s_i, a_i) + \alpha \left(\max_2 \left(\left(R(s_i) + \gamma \max_{a \in A_{i+1}} Q(s_{i+1}, a_{i+1}) - Q(s_i, a_i) \right) + D(t_0)d^t, D(t_0)d^t \right) \right)$$



The key terms are described in the following list:

- $R(s_i)$: True reward of state s_i
- $Q(s_i, a_i)$: Action-value function for taking action a_i in state s_i.
- α: Learning rate, determining how much new information overrides old information.
- γ: Discounting factor, determines how important later rewards are
- $R(s_i) + \gamma \max_{a \in A_{i+1}} Q(s_{i+1}, a_{i+1}) - Q(s_i, a_i)$: Temporal difference error term, where $R(s_i)$ is the reward.
- $\max_{a \in A_{i+1}} Q(s_{i+1}, a_{i+1})$: Maximum future value given the current understanding at the next state s_{i+1}.
- $D(t_0)d^t$: Dopamine surge monotonic decreasing function.

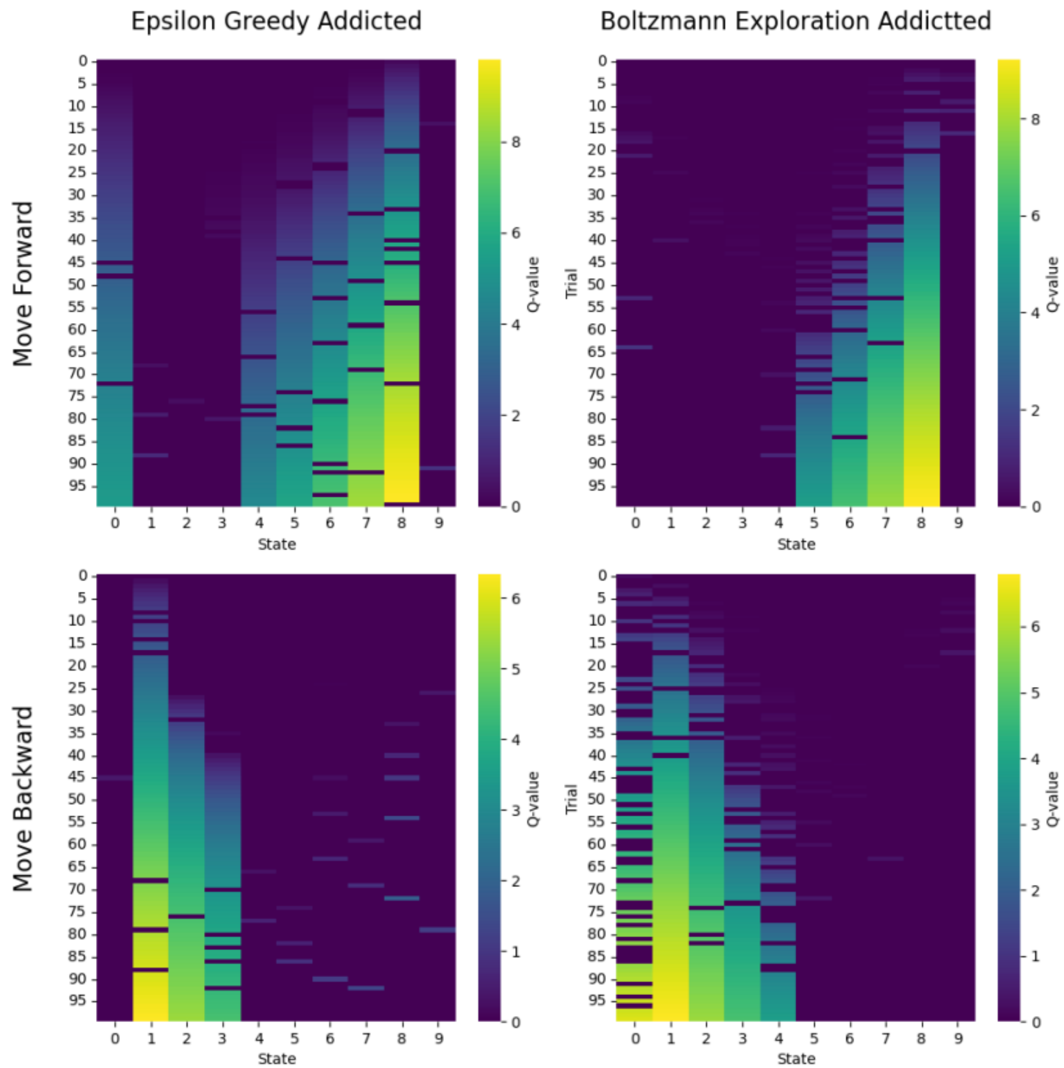
Preliminaries



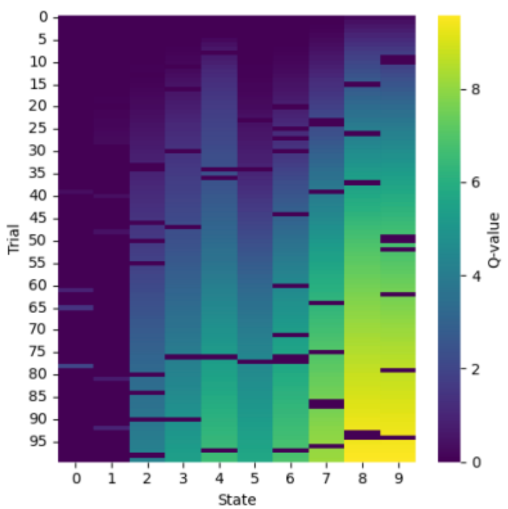
Heatmap shows the learning process:

1. Reward propagates backward/forward to other states.
2. A_e fails to find natural reward, but A_b does.
3. A_e representation not robust.

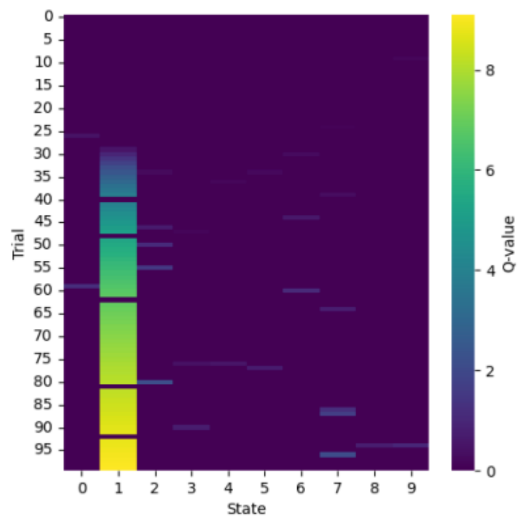
Propagate backward/forward



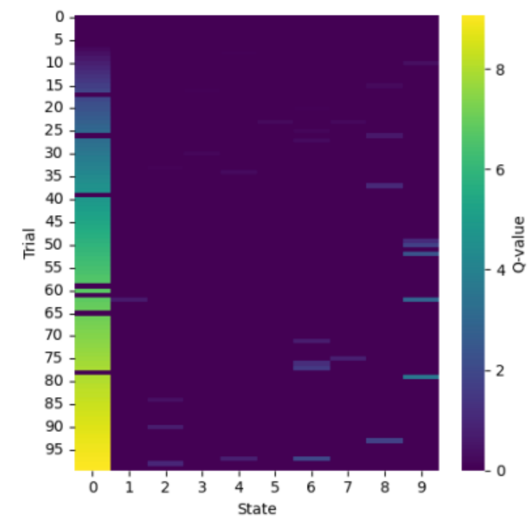
Epsilon Greedy With Natural Reward



Forward

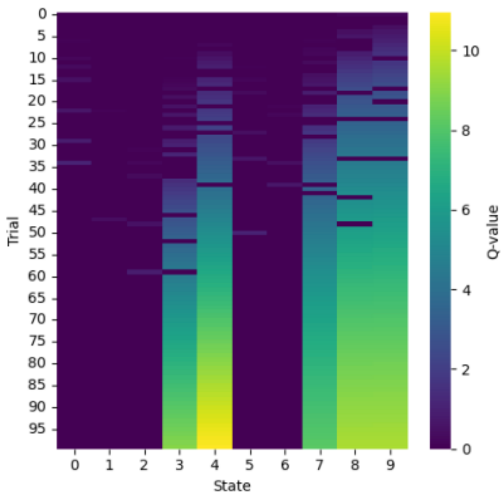


Backward

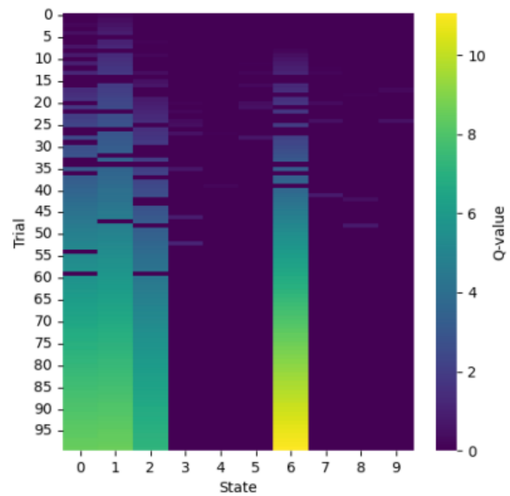


Stay

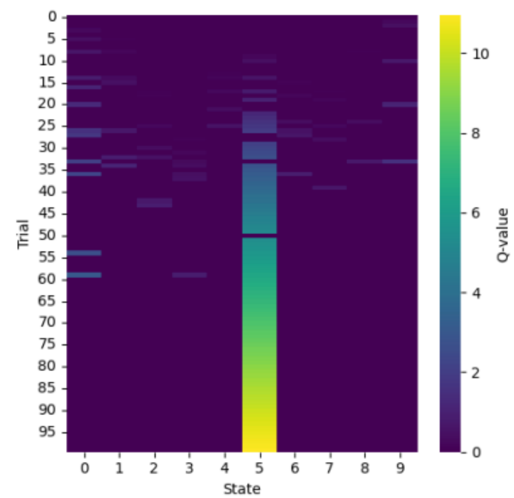
Boltzmann Exploration With Natural Reward



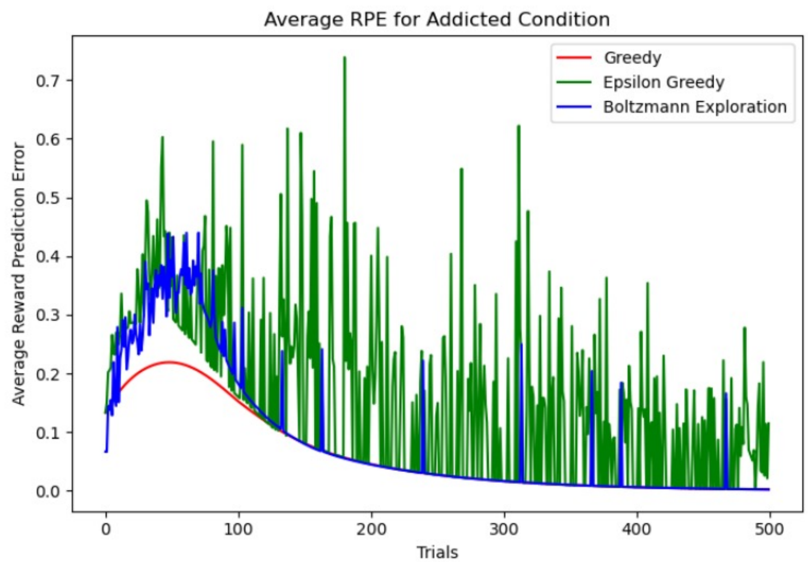
Forward

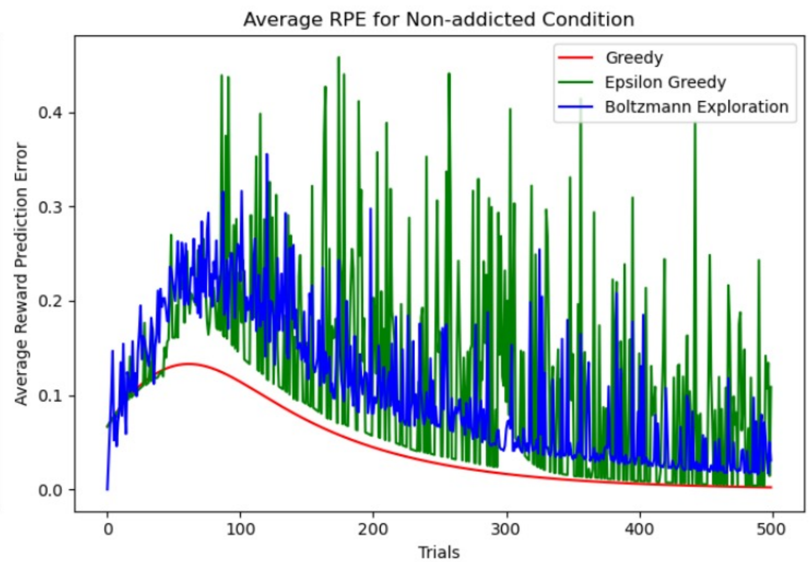


Backward



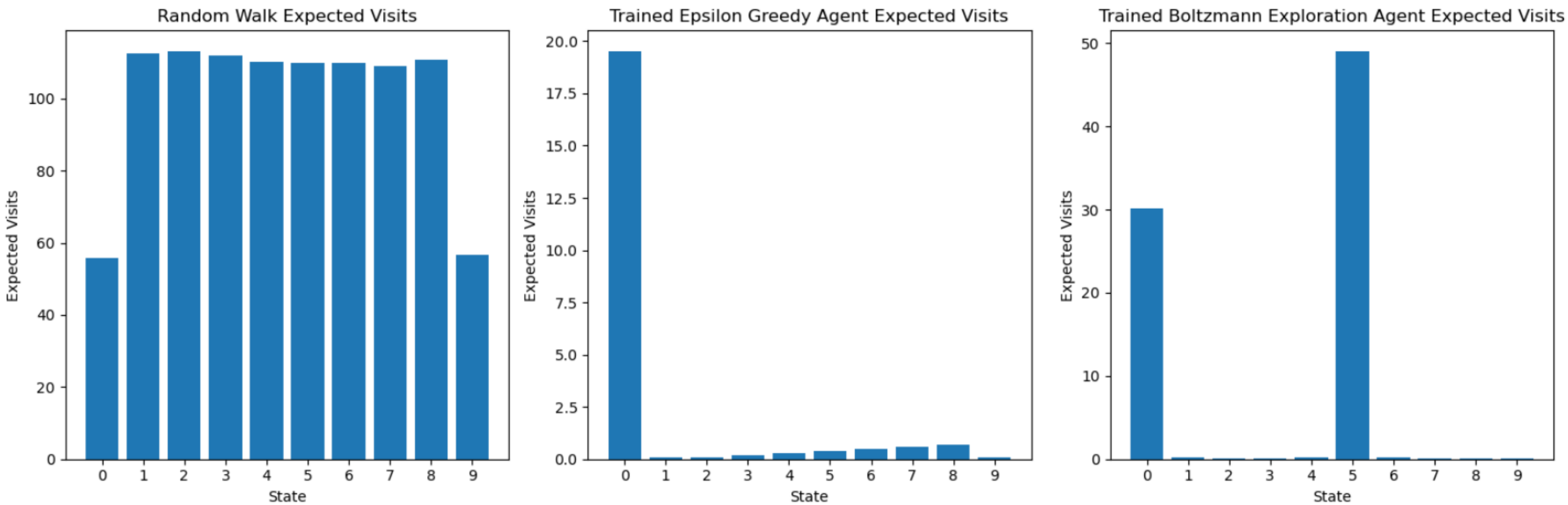
Stay





RPE shows what is unexpected and where to go:

1. Greedy build most robust (may be wrong), A_b second, A_e not robust.
2. Concavity exist, Suprises -> learning
3. Non-addicted agent more surprised -> drug grabs all “attention”



Compare with pure randomness?

1. Pure stochastic movement through the space stays more in the middle, showing the powerful effect of drugs.
2. Similar effect to heat map, A_b finds the higher reward, A_e stuck at drug.

What Now?

What's lacking & where to go next?

Propose novel perspective of introducing actions. Only points out the computation and theoretical aspect.

- More studies on the biological experimentation side to proof the validity (behavioral + neuronal counterparts)

Only touches on the most fundamental search strategies in discrete setting.

- Explore and incorporate more algorithms (UCB)
- Moving to continuous policy optimization (Actor-critic)