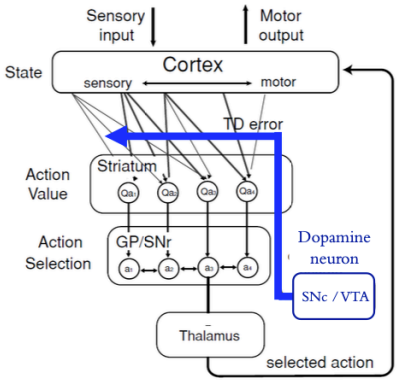
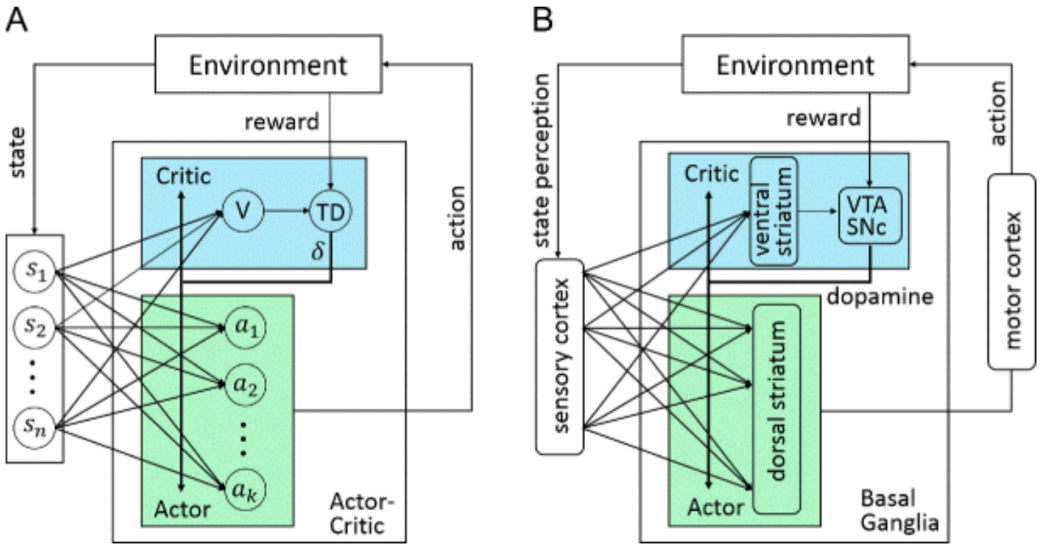
Addiction as a Computational Process Gone Awry

Striatum incorporates environmental state sensory/motor information from the cortex -> adjusts weights on the actions depending on the dopamine RPE from VTA -> striatum outputs selected action back to cortex -> influences movements, decision-making, and further reward processing.



These processing steps are exactly the actor critic model

- Dorsal striatum learns how to output.
- Ventral striatum learns the values of objects from reward error signals and helps the dorsal striatum to learn the outputs.



The structure of the Actor-Critic algorithm for decision making (a), and the corresponding brain regions and pathways (b). The critic function and actor function, which correspond to the ventral striatum and dorsal striatum in basal ganglia individually, are shown in blue area and green area respectively. TD error is carried by dopamine which is elicited from VTA and SNc. The Actor-Critic algorithm mainly focuses on the simulation of actor and critic function in basal ganglia. It is too slow to solve the complex tasks since lacking of contextual information from working memory

If the brain mechanism really works like TD updates structurally, we can use the same TD structure to model and understand compulsive behaviors better.

Addicted TDAgent Theory

The dopamine error signal is not equivalent to pleasure; instead, it is an ***internal signal indicative of the discrepancy between expectations and observations***.

- The agent selects actions proportional to the expected benefit that would be accrued from taking the action (from behavior matching law)

- Because δ transfers ***backward from reward states*** to anticipatory states with learning, ***actions can be chained together*** to learn sequences. This is the heart of the TDRL algorithm.
- Action selection then would be counted as happens in a semi-Markov state space.

$$\delta(t) = \gamma^d[R(S_l) + V(S_l)] - V(S_k) \quad (2)$$

$$V(S_k) \leftarrow V(S_k) + \eta_V \delta \quad (3)$$

Usually speaking, phasic increases in dopamine are seen after unexpected natural rewards; however, ***with learning, these phasic increases shift from the time of reward delivery to cueing stimuli***. Transient increases in dopamine are now thought to signal changes in the expected future reward (i.e., unexpected changes in value). These increases can occur either with unexpected reward or with unexpected cue stimuli known to signal reward and have been hypothesized to signal δ.

Normally speaking once the value function correctly predicts the reward, learning stops. The value function can be said to compensate for the reward: The change in value in taking action would counter-balances the reward achieved on entering the next state. When this happens, δ = 0. In another word, ***taking transient dopamine as the d signal correctly predicted rewards produces no dopamine signal***.

However, Cocaine produce a transient increase in dopamine through ***neuropharmacological mechanisms***, producing dopamine surge that can be modeled by assuming that these drugs induce an increase in δ that cannot be compensated by changes in the value. The effect of addictive drugs is to ***produce a positive*** δ ***independent of the change in value function*** (idea inspired by neuropharmacological mechanisms), making it impossible for the agent to learn a value function that will cancel out the drug-induced increase in δ and the constant increases approaches values to infinity.

$$\delta = \max\{\gamma^d[R(S_l) + V(S_l)] - V(S_k) + D(S_l),\ D(S_l)\} \quad (4)$$

- Values of states leading to natural rewards asymptotically approach a finite value (the discounted, total expected future rewards, approximated by TD update), leading to asymptotic balances.
- Values of states leading to drug receipt increase without bound.
- The more the agent traverses the action sequence leading to drug receipt, the larger the value of the states leading to that sequence and the more likely the agent is to select an action leading to those states (***early use of drugs occurs because they are highly rewarding, but this use transitions to a compulsive use with time***).
- In reality, it is also unlikely for the value of cocaine to go to infinity. Biological compensation mechanisms are likely to limit the maximal effect of cocaine on neural systems, including the value representation.

Rational to Irrational Decision Perspective

The TDRL theory proposed in this paper differs from that of ***rational addiction*** where the drug just stands for a higher value because TDRL proposes that addiction is inherently irrational: It uses the same mechanisms as natural rewards, but the system behaves in a nonoptimal way because of neuropharmacological effects on dopamine and the value function cannot compensate for the D(s) component, the D(s) component eventually overwhelms the R(s) reward terms. Eventually, the agent behaves irrationally and rejects the larger rewards in favor of the (less rewarding) addictive stimulus.

Dopamine and Delta both as “Wanting”

The value function is a means of guiding decisions and thus is more similar to ***wanting than to liking in the terminology of Robinson and Berridge***. In TDRL, dopamine does not directly encode wanting, but because learning an appropriate value function depends on an accurate d signal, dopamine will be necessary for acquisition of wanting.

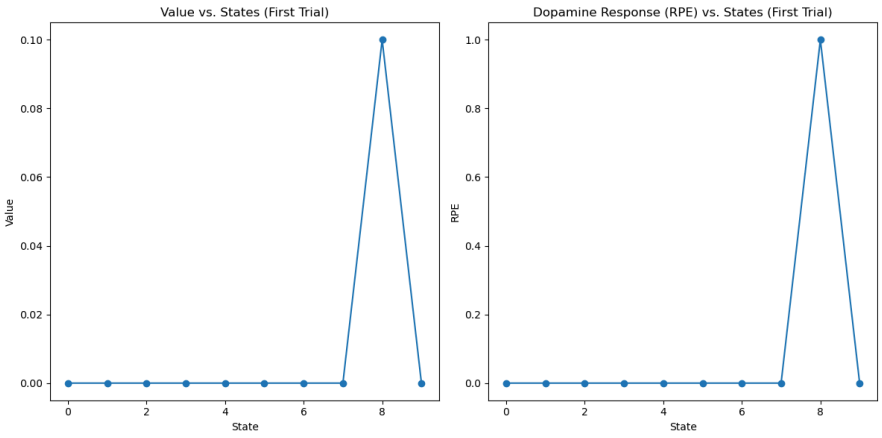
Harder Extinction Biologically

In normal TDRL, the value of states leading to reward decay back to zero when that reward is not delivered. This follows from the existence of a strongly negative d signal in the absence of expected reward. Although firing of dopamine neurons is inhibited in the absence of expected reward, the inhibition is dramatically less than the corresponding excitation. In general, the simple decay of value seen in TDRL does not model extinction very well, particularly in terms of reinstantiation after extinction.

Simulation of Addictive TDAgent (Value & Dopamine Error)

First Trial Value & Dopamine Error Signal

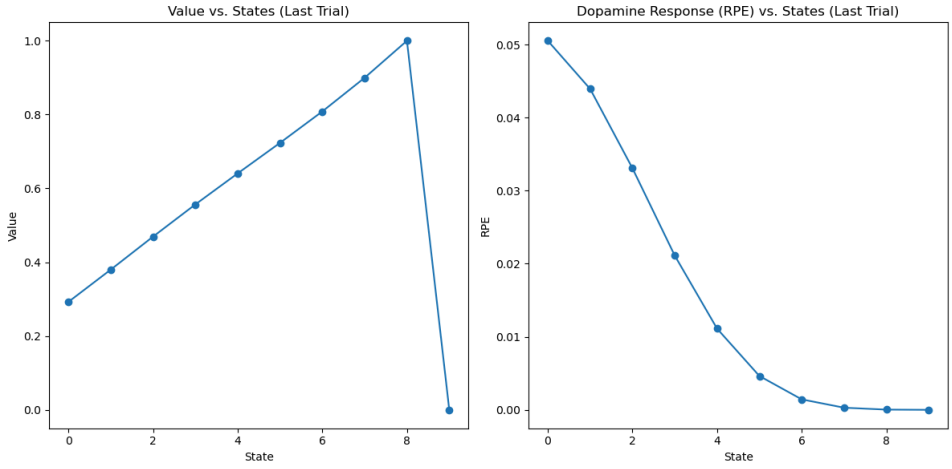
The value of a given state and the dopamine error response both peak at state 8, which is the state leading into the reward state, indicating a big error in prediction due to the unexpected reward stage. As for the value graph, the agent learns the value of the state leading to the reward state as a high value state since the TD update equation of a current state is updated with the next state’s value.



Last Trial Value & Dopamine Error Signal

This represents a more stable understanding of the environment where the value forms a linear rise with the increases in state and has a sharp drop exactly at the reward state (because the next state is zero value). On the other hand, the dopamine prediction error is a much smoother curve where there is a higher positive prediction error in the starts of the state and then lowers the errors in the later states.

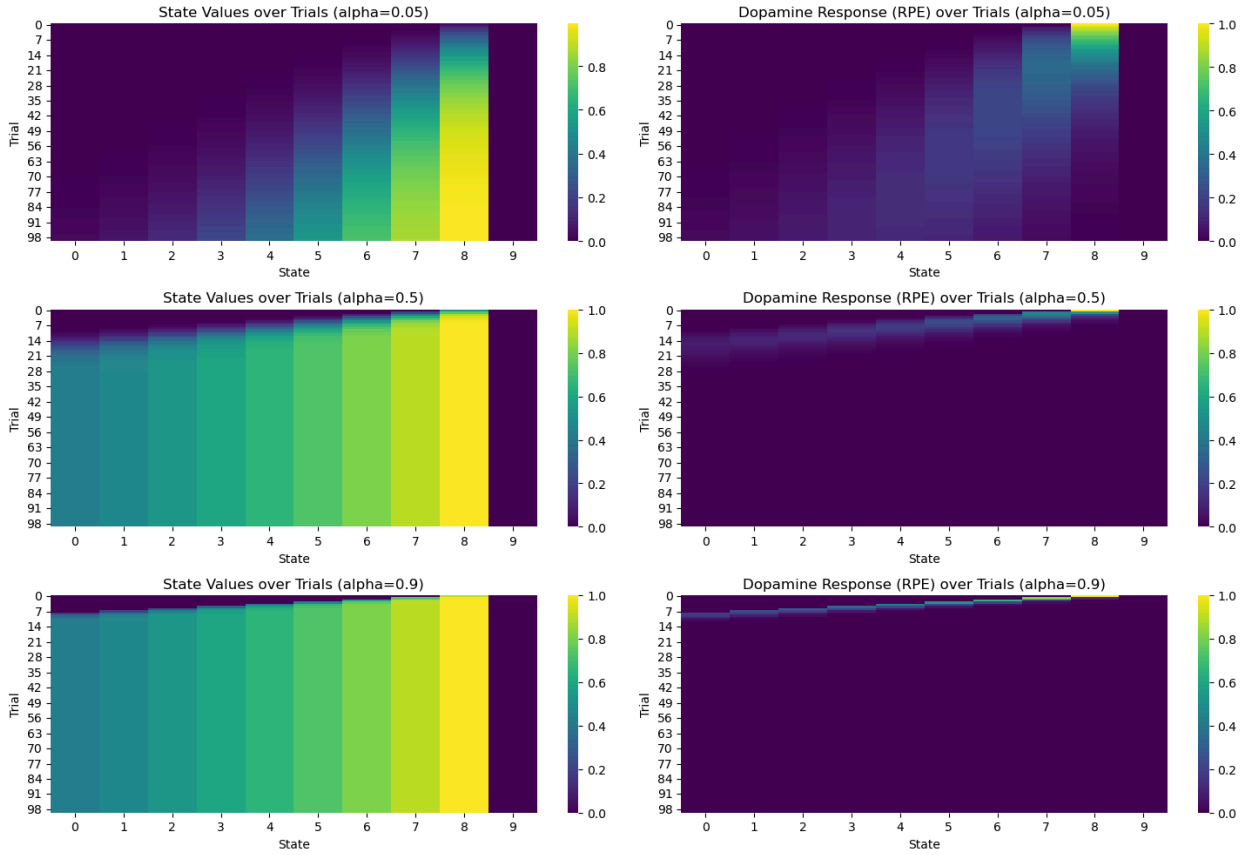
1. This might be caused by the fact that it is closer to make more correct predictions as it is closer to the reward state and it gets harder to understand the values when the states gets more distance from the actual rewarding state, the “signal” is a lot less strong and it becomes harder to predict the actual correct value of the state (temporal nature).
2. Moreover, because later states’ value is essentially carried away by the reward directly, it is easier to estimate the values. However, it becomes harder in the beginning because the estimates would depend on the prediction value of the future state, which is not as stable of a signal to use as the reward itself.



Learning Rate Heat Map Across Trials

Since the environment in this condition is not super complex, as the learning rate increases, the agent is able to achieve a good understanding of the environment faster. Same is reflected on the heat map:

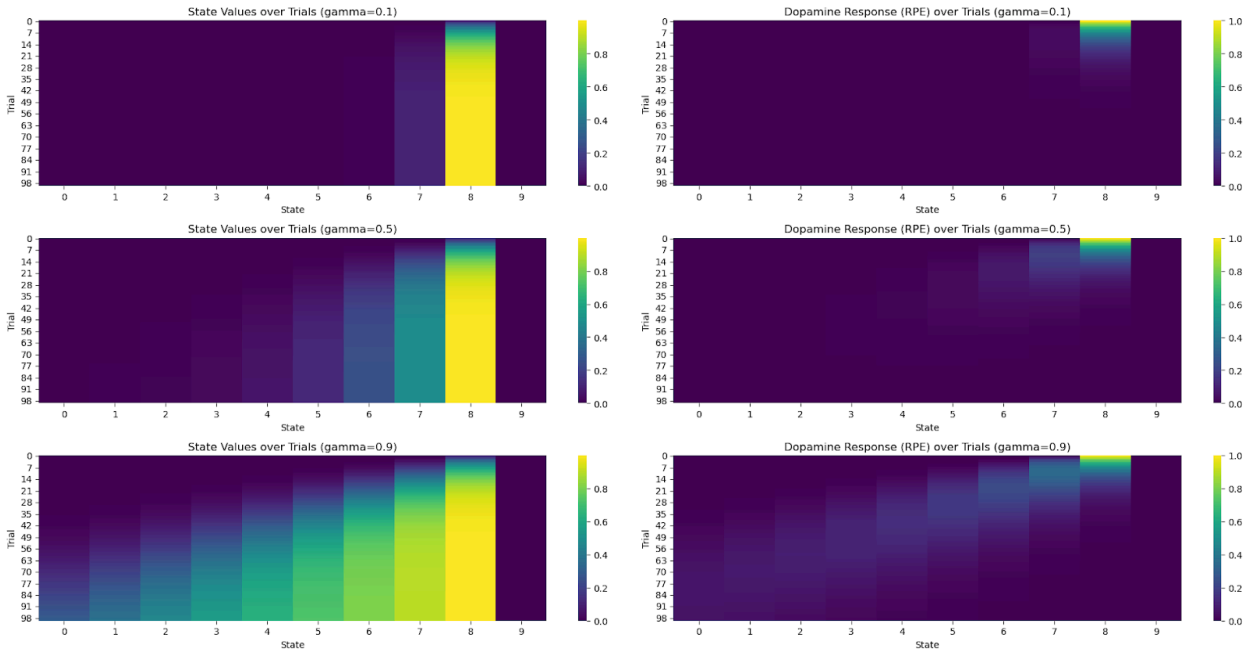
1. As getting closer to the reward state, the reward signal becomes stronger (more yellow), this is the same with previous line plots.
2. As the learning rate increases, the agent is able to build a robust understanding of the environment faster (0.05 agent is having a hard time understanding earlier states even at later trials while the 0.9 agent is able to grasp the understanding earlier in trials)
    a. Earlier states (further from reward states) are also harder to learn the values of as demonstrated in the previous question.
3. The highest error prediction is around the time of the reward in earlier trials, just as demonstrated previously. Gradually the prediction error, even for the low learning rate ones, drops to zero around the reward.
4. The increase in learning rate makes the drop of the prediction error around the reward state really quickly.



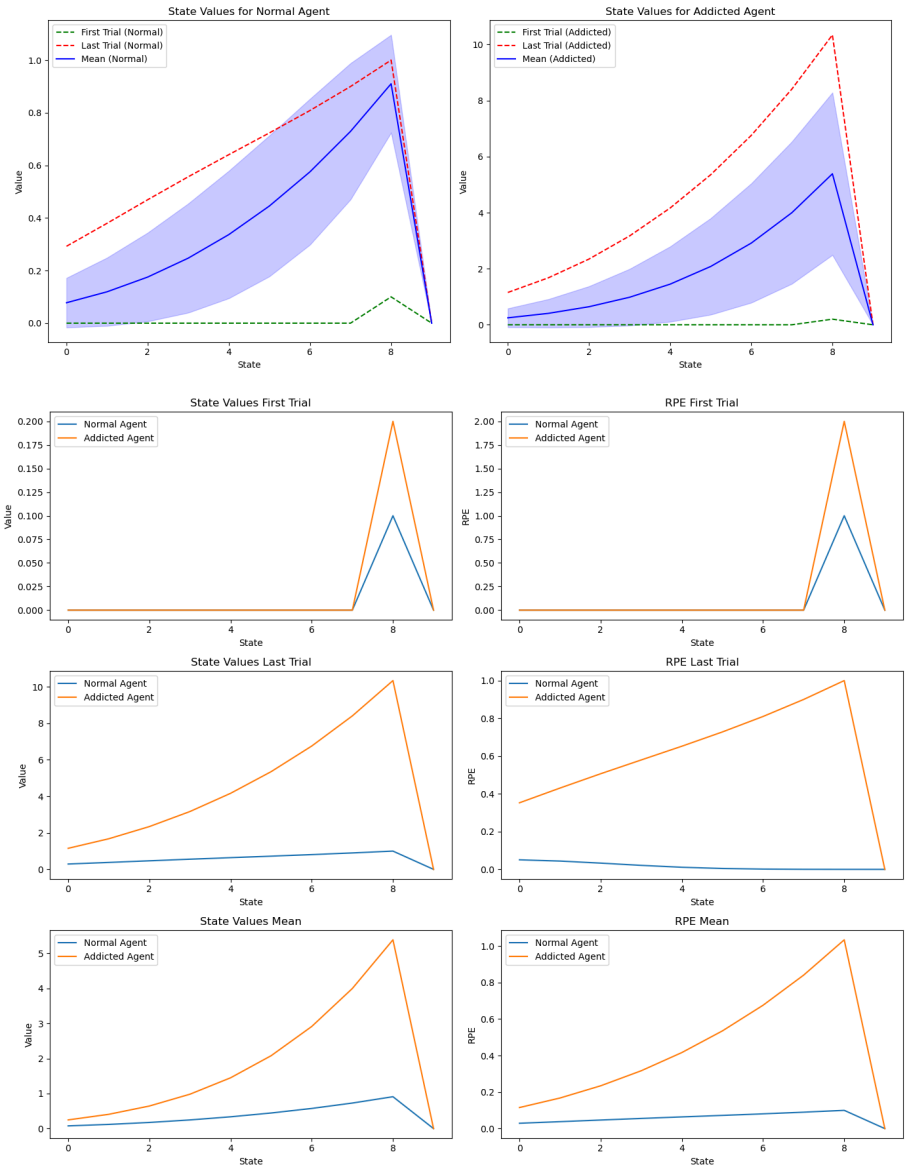
Discount Rate Heat Map Across Trials

The higher the discount rate, the more valuable the later state would impact the previous state's value. As can be seen from the state value over trials heat map, higher gamma value would lead to faster and sharper updates in the state values as there is a strong consideration for the future reward, so the whole sequences of states would be sequentially more activated getting closer to the reward. Under the same logic, RPE would initially be higher for the states leading up to the reward state because there is a surprising stronger correlation between the earlier states and the later reward.

On the other hand, a low discount rate would mean that future reward doesn't matter as much for earlier states and it is reflected by the state value graph where earlier states would not be activated on heat map at all as it is not considering later reward at all (initialized value at 0). Only the very near state next to the state prior to the reward state (state 7) would have value activation. Similarly, the RPE graph would only have unexpected surprising value errors only on the state prior to the reward state as that is the only one that has a difference from expectation (originally all the state values start from 0, so there would be no change in the expected value of earlier stages, hence no RPE signals).



Addictive/Normal Agent First, Last, and Mean Trial



No matter for the first trial, last trial, or the mean trial value, the addicted model is always getting higher values, which is an effect caused by the drug state reward propagating backward to the earlier state (can be seen from the last trial and mean trail graph for state values, the tail indicate that the reward is been propagated back to the earlier state and the closer a state gets to the drug state, the higher the reward it would be.

- The same can be noticed from the first two graphs, when the normal agent maintains a linear relationship of values taking up to the reward, the addicted agent forms a skewed relationship between values and states leading up to the reward states and the values are a lot higher in comparison.

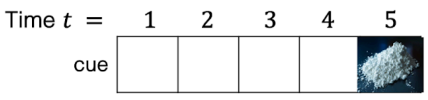
From the RPE graph, the addicted agent would maintain a continuously positive prediction error from the time of cue to the drug state because the the values of the values of any states would increase until infinity from the reward back propagation effect of the constantly increasing reward of drug state (always a plus to the expected values), neglecting the functional structure of the TD update and the termination of learning when reward is achieved.

- Learning goes on continuously until infinity in a wrong direction , anticipating more and more rewards, no stops when rewarded , never being satisfied from the reward.

Manual TDAgent Solving

The values of each stage since the cue are updated from the reward drug state back to the time of the cue -> propagation backward.

1. The values actually all initially start at zero for t=0 to t=4, but gradually there would be an increase that propagates backward due to cocaine, even when reward is not presented at state t=0 to t=4.
2. Values for state 5 (drug stage) gradually increase constantly, natural reward does not increase, no reinforcing, continuously increasing reward in the cocaine case.
3. Reward propagates backward and, the closer the state is to the drug state, the higher the state value would be.
4. Maintain a continuously positive prediction error from the time of cue to the drug state -> the values increase until infinity, neglecting the functional structure of the TD update.
5. "Learning" goes on continuously, anticipating more and more rewards, no stops when rewarded , never being satisfied from the reward.
6. Without external help, no value function can be learned to actually counteract the effect of the drug. Only when a large value exists for alternative choice, alternatives may be taken.



Trial 5

t	$r(t)$	$\delta(t)$	$\hat{V}(t)$
1	0	0.98	0.98
2	0	0.36	1.46
3	0	0.405	2.03
4	0	0.45	2.7
5	1	0.5	3.5

t=1 | $\delta(t) = 0 + 0.9(1.09) - 0 + 0 = \max(0.98,0) = 0.98$
$\hat{V}(t=1) \leftarrow 0 + 1(0.98) = 0.98$

t=2 | $\delta(t) = 0 + 0.9(1.62) - 1.09 + 0 = \max(0.368,0) = 0.368$
$\hat{V}(t=2) \leftarrow 1.09 + 1(0.368) = 1.46$

t=3 | $\delta(t) = 0 + 0.9(2.25) - 1.62 + 0 = \max(0.405,0) = 0.405$
$\hat{V}(t=3) \leftarrow 1.62 + 1(0.405) = 2.03$

t=4 | $\delta(t) = 0 + 0.9(3) - 2.25 + 0 = \max(0.45,0) = 0.45$
$\hat{V}(t=4) \leftarrow 2.25 + 1(0.45) = 2.7$

t=5 | $\delta(t) = 1 + 0.9(0) - 3 + 0.5 = \max(-1.5,0.5) = 0.5$
$\hat{V}(t=5) \leftarrow 3 + 1(0.5) = 3.5$