Accelerated VDBE: Q-Learning with Differentiable Value-Difference Based Exploration Rate Updates

Daniel Huber

December 2025

Problem Formulation

- Frequently seen constraints on modern autonomous systems:
  - Limited onboard computers.
  - Cannot afford to train a deep neural network with GPU-dependent algorithms.
- How to conserve compute power?
  - Robots can learn simple tasks with a simpler method than a deep neural net.
  - Hence, Q-learning

Q-Learning Overview

Q-Value Update Rule:

$$Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \alpha[R_{t+1} + \gamma \max_a Q(S_{t+1}, a) - Q(S_t, A_t)] \quad [2] \tag{1}$$

ε-Greedy Policy:

$$\pi(s) = \begin{cases} \text{argmax}_{a \in \mathcal{A}(s)} Q(s, a) & \text{if } \zeta > \varepsilon,\ \zeta \in [0, 1] \text{ is randomly selected each time step} \\ \text{Random action} & \text{else} \end{cases} \tag{2}$$

The Exploration-Exploitation Dilemma

Exploration versus Exploitation [2]: how to balance?

- Exploration rate ε is a constant between 0 and 1.
- Common Approaches:
  - Set at a constant value
  - Decreases gradually over time.
  - Selected with some other heuristic strategy.
- Goal: Make ε updates smarter.
- Solution: Improve VDBE [1] using a differentiable parameter.

Previous Work: Standard Value-Difference Based Exploration (VDBE) [1]

Update Rule:

$$\varepsilon_{t+1} = \delta \tanh \frac{2\alpha|T|}{\sigma} + (1-\delta)\varepsilon_t \tag{3}$$

- $|T|$: Absolute value of temporal difference: $R_{t+1} + \gamma \max_a Q(S_{t+1}, a) - Q(S_t, A_t)$. This is basically the difference between current Q-value estimate and previous Q-value estimate.
- $\tanh \frac{2\alpha|T|}{\sigma}$ = "surprise function" (normalized temporal difference)
- σ: Positive scaling parameter called inverse sensitivity [1]
- δ: Parameter determining influence of surprise

Contributions: Accelerated VDBE

Hypothesis: Standard VDBE makes updates to ε highly volatile.

Solution: Smooth out updates to ε by introducing differentiability.

The Derivative of Surprise:

$$\begin{gathered} \text{surprise } U(T) = \tanh \tfrac{|T|}{\sigma} \\ \delta_* = |\tfrac{d}{dT} U(T)| = \tfrac{1 - U^2(T)}{\sigma} \end{gathered} \tag{4}$$

New Update Rule:

$$\varepsilon_{t+1} = \frac{U - U^3}{\sigma} + (\frac{\sigma + U^2 - 1}{\sigma})\varepsilon_t \tag{5}$$

Behavior Comparison

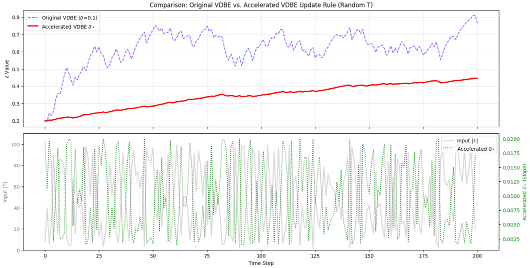


figure 1: $|T|$ varies randomly

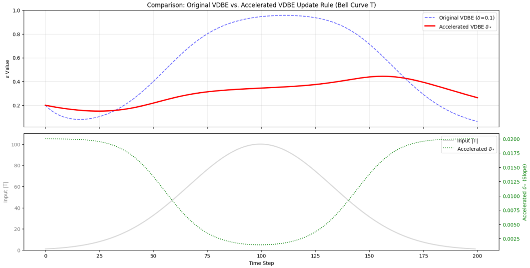
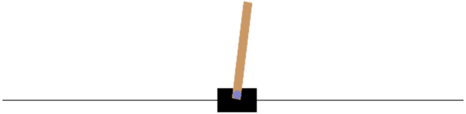


figure 2: $|T|$ follows a smooth curve

Testing in Cartpole Environment



Screenshot of cartpole environment [3]

- **Goal:** "Balance the pole by applying forces in the left and right direction[s]" [3].
- **Failure:** Pole falls below $\pm 12^\circ$ of vertical.
- **Success Criteria:** Keep the pole upright for an average of 225 time steps for 100 episodes in a row.
- Credit to Farama Foundation [3] for Q-learning and Environment script templates, which I modified to build this project.

Results

AGENT	SOLVED (Avg Eps)	TIME (Avg)	BEST RUN
Normal Q-Learning	5000.0 ± 0.0	54.12s	5000
Standard VDBE (σ=1)	5000.0 ± 0.0	77.05s	5000
Standard VDBE (σ=5)	2126.4 ± 443.0	53.21s	1518
Standard VDBE (σ=20)	3904.8 ± 1002.0	91.99s	2328
Standard VDBE (σ=100)	5000.0 ± 0.0	63.57s	5000
Accel VDBE (σ=1)	5000.0 ± 0.0	117.48s	5000
Accel VDBE (σ=5)	1882.1 ± 448.3	52.93s	1079
Accel VDBE (σ=20)	2846.4 ± 1465.7	84.06s	1068
Accel VDBE (σ=100)	5000.0 ± 0.0	105.52s	5000

Accelerated VDBE presents a 13% improvement over normal VDBE in the best case scenario of $\sigma = 5$

Other Attempts to Improve Q-Learning

- Initial Attempt: Differentiable learning rate proportional to rate of change of reward.
- Result: Did not work well for Q-learning.
  1. Skews Q-values, rendering previous exploration useless.
  2. Difficulty in taking the derivative of reward with respect to actions or states in a discrete environment.

Conclusions and Future Work

Conclusion:

- Accelerated VDBE updates exploration rate proportional to the derivative of the agent's level of "surprise."
- Performs better than regular VDBE and normal Q-learning in testing.

Future Work:

- Turn this project into a publication.
- Implement an exploration-action selection mechanism that is more intelligent than a purely random choice.
- Determine whether it might be possible to combine Accelerated VDBE with n-step methods.

References

1. Tokic, Michel, and Günther Palm. Value-Difference Based Exploration: Adaptive Control between Epsilon-Greedy and Softmax. KI'11: Proceedings of the 34th Annual German conference on Advances in artificial intelligence, 4 Oct. 2011. https://www.tokic.com/www/tokicm/publikationen/papers/KI2011.pdf
2. Sutton, R. S., & Barto, A. G. (2021). Reinforcement Learning an Introduction. MTM.
3. Farama Foundation. (2025). Gymnasium documentation. Cart Pole - Gymnasium Documentation. https://gymnasium.farama.org/environments/classic_control/cart_pole