
VRAIL: Vectorized Reward-based Attribution for Interpretable Learning

Jina Kim
School of Computing
jinakim@kaist.ac.kr

Youjin Jang
School of Computing
jyj22@kaist.ac.kr

Jeongjin Han
School of Computing
hjj22@kaist.ac.kr

Abstract

We propose **VRAIL** (Vectorized Reward-based Attribution for Interpretable Learning), a bi-level framework for value-based reinforcement learning (RL) that learns interpretable weight representations from state features. VRAIL alternates between value estimation and RL training with shaped rewards. Experiments in the Taxi-v3 domain show that VRAIL improves convergence and stability over standard DQN without environment modifications, while revealing subgoals such as passenger possession. These results suggest that VRAIL enhances both learning performance and interpretability in a model-agnostic manner within value-based RL.

1 Introduction

In reinforcement learning (RL), designing an effective reward function remains a critical yet inherently challenging task. While the principle that “reward is enough” suggests that reward maximization alone can explain intelligent behavior [4], real-world environments often involve sparse, delayed, or ambiguous reward signals due to their complex and nuanced dynamics [1]. Humans make decisions by subjectively evaluating multiple factors, weighing changes in possession or outcomes against internal value judgments. For instance, consider the act of donating money: although it results in a loss of monetary possession, it may simultaneously produce a positive emotional outcome. Motivated by this insight, we introduce VRAIL (Vectorized Reward-based Attribution for Interpretable Learning), a novel method that approximates reward shaping through learnable feature weighting. VRAIL aims to accelerate agent convergence while making interpretability of learned values, by modeling how state transitions affect both possession and perceived value.

A toy experiment shown in fig. 1 further illustrates the necessity of our proposed method. In the standard Taxi-v3 environment from the OpenAI Gymnasium library [5], the agent’s observations do not include explicit wall information, which plays a critical role in successful navigation. As shown in fig. 1, without this structural information, a vanilla DQN baseline fails to converge in 2 out of 10 random seeds, highlighting the instability that can arise from incomplete state representations. Although we found that manually adding wall information resolved the issue and led to consistent convergence, such interventions are impractical in real-world settings. Our expectation is that VRAIL achieves stable performance even in the original, unmodified environment by learning to attribute value to latent features within state transitions. Therefore, VRAIL will offer a practical and potentially generalizable approach for RL tasks where full state observability cannot be guaranteed.

2 Method

We propose **VRAIL** (Vectorized Reward-based Attribution for Interpretable Learning), a framework for learning interpretable weight via bi-level optimization. As illustrated in fig. 2, the process consists of two interacting stages: a reinforcement learning (RL) stage and a deep learning (DL) stage. The overall objective is to help the agent recognize meaningful feature transitions as implicit subgoals, improving sample efficiency and interpretability.

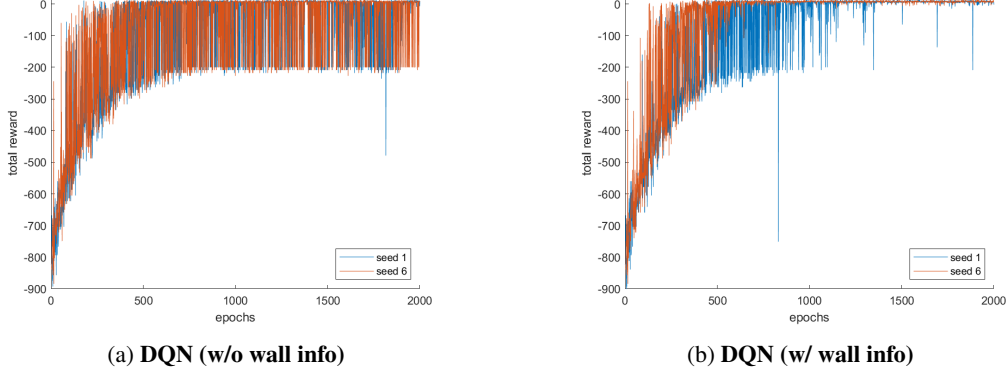


Figure 1: **Effect of wall information on DQN convergence.** (a) Without wall information, DQN fails to converge in 2 out of 10 seeds. (b) Including wall information leads to consistent convergence across all seeds. For clarity, only the 2 failure cases are shown.

2.1 RL Stage: Reward Shaping with Learned Potentials

In the RL stage, we train a value-based reinforcement learning agent using rewards shaped by the output of the DL stage. By the potential-based reward shaping theorem [3], we can guarantee policy invariance under transformations of the form $R'(s, a, s') = R(s, a, s') + \gamma\Phi(s') - \Phi(s)$. For the shaping potential $\Phi(s)$, we use an estimated value function $\hat{V}(x_s; \theta)$, which approximates the expected cumulative return from $x_s \subseteq s$, where x_s is the extracted feature state vector s and θ represents the model parameter of the estimator. This choice is motivated by the observation that $V(s)$ inherently encodes how close the agent is to achieving the main goal, thus implicitly capturing useful subgoal structures. By shaping the reward using $V(s)$, we aim to encourage behaviors that guide the agent through intermediate progress toward the final objective.

2.2 DL Stage: Learning a Interpretable Value Function

In the DL stage, we make an estimator $\hat{V}(x_s; \theta)$ to approximate the state-value function $V(s)$. The function $\hat{V}(x_s; \theta)$ is parameterized using either a linear form (Linear VRAIL) or a quadratic form (Quadratic VRAIL) as

$$\hat{V}(x_s; w) = w^\top x_s \quad \text{or} \quad \hat{V}(x_s; W) = x_s^\top W x_s,$$

where $w \in \mathbb{R}^d$ and $W \in \mathbb{R}^{d \times d}$ are learnable parameters.

2.3 Bi-level Optimization Loop

After each RL stage, the updated value function $V(s)$ is collected and used to refine the reward parameters in the DL stage. This iterative procedure forms a closed bi-level optimization loop, where interpretable reward functions are continuously improved to support more effective learning.

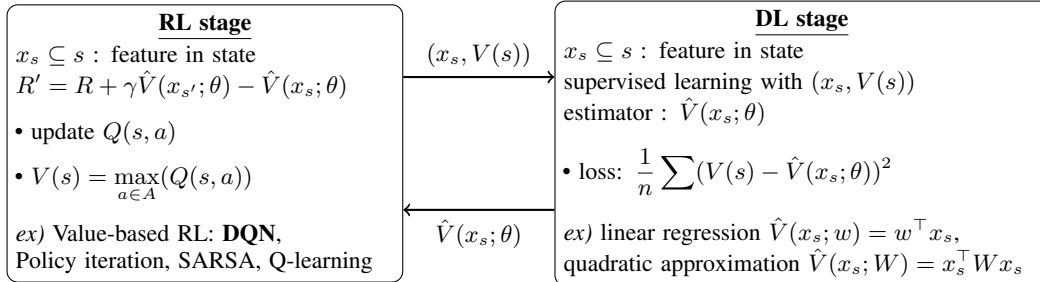


Figure 2: **Bi-level optimization framework of VRAIL.** The DL stage learns reward parameters from state-value estimates, and the RL stage uses them to shape rewards and train the agent’s policy.

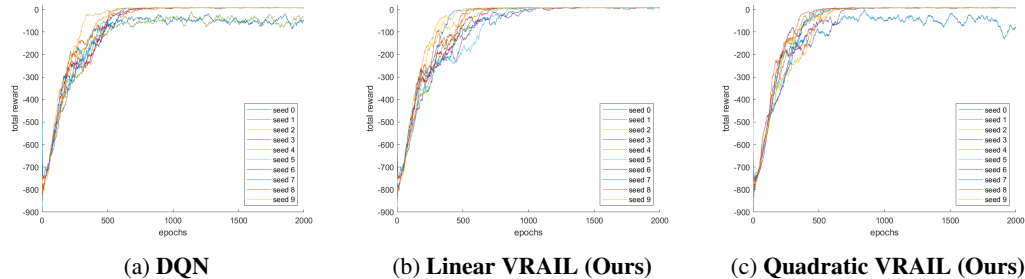


Figure 3: **Training stability comparison across models.** Total reward per episode over 2,000 training epochs. VRAIL variants (b, c) show improved convergence stability over the DQN baseline.

3 Experiments

3.1 Experimental Setup

We conduct experiments on the classic Taxi-v3 environment [5], where the agent navigates a gridworld to pick up and drop off a passenger. The environment provides discrete observations of the taxi’s position, passenger location, and destination. To evaluate VRAIL without modifying the original state space, we use the default environment without injecting features like wall information. While VRAIL supports various value-based RL algorithms [2], we focus on DQN for empirical evaluation. All models are run with 10 random seeds to assess convergence stability and robustness.

3.2 Training Stability and Convergence Speed

We evaluate the convergence behavior of VRAIL in terms of both training stability and speed, using a standard DQN baseline for comparison. A run is considered converged if the agent consistently achieves a moving average reward above a fixed threshold during training. As shown in fig. 3, the baseline DQN converged in 8 out of 10 runs. In contrast, Linear VRAIL converged in all 10 runs and Quadratic VRAIL converged in 9 out of 10 runs, demonstrating greater robustness than the baseline.

Table 1: Average epochs to reach reward thresholds (10 seeds, excluding top/bottom 2 outliers).

Reward Threshold	−10	−5	0	+5
DQN	600.00	612.17	648.17	717.67
Linear VRAIL (Ours)	614.17	643.17	652.50	735.83
Quadratic VRAIL (Ours)	538.17	562.83	594.33	660.17

We evaluate convergence speed by measuring the number of training epochs required to reach moving average reward thresholds of -10 , -5 , 0 , and $+5$. As shown in table 1, Quadratic VRAIL consistently reaches all thresholds faster than both DQN and Linear VRAIL. While Linear VRAIL converges more reliably, it takes slightly longer than DQN to achieve the same rewards. These results highlight a trade-off: Linear VRAIL offers robustness, while Quadratic VRAIL combines speed with stability.

3.3 Effect of Shaped Reward Transfer

To further isolate the effect of learned reward shaping, we evaluate whether the reward function learned by VRAIL can benefit standard RL agents when applied independently. Specifically, we pretrain Linear VRAIL to learn a potential-based reward function, then fix this function and apply it to a vanilla DQN agent without further reward updates. As shown in fig. 4, the baseline DQN fails to converge in 1 out of 5 seeds. In contrast, DQN trained with the reward shaped by pretrained Linear VRAIL successfully converges in all 5 seeds. This suggests that VRAIL’s learned reward function serves as a robust shaping mechanism, improving convergence stability even when decoupled from the full bi-level optimization loop.

4 Analysis

To highlight the interpretability of VRAIL, we analyze the learned parameters in our models.

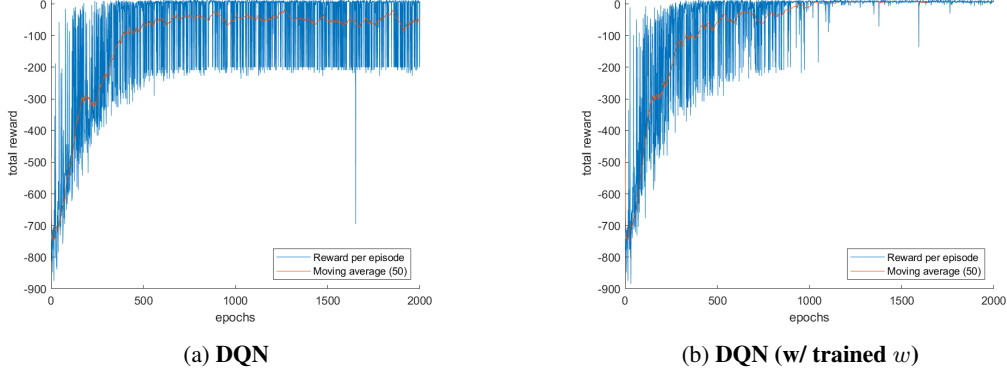


Figure 4: **Effect of reward shaping using a pretrained Linear VRAIL model on DQN training.** Standard DQN fails to converge in 1 out of 5 seeds, while DQN trained with the reward function learned by Linear VRAIL converges in all 5 runs.

4.1 Linear VRAIL

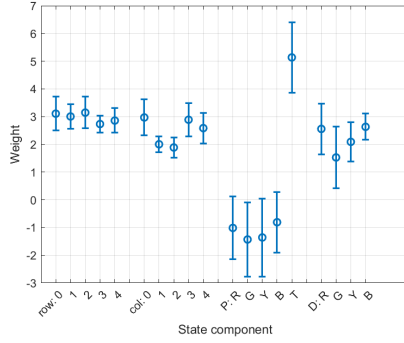
In Linear VRAIL, the value function is represented as $\hat{V}(x_s; w) = w^\top x_s$, where $w \in \mathbb{R}^{19}$ is a weight vector, with each element corresponding to a state feature. To capture a range of learned behaviors across different runs, we average the learned weight vectors over 10 random seeds and visualize the resulting feature importance. As shown in fig. 5a, the feature assigned the highest weight corresponds to whether the passenger is currently in the taxi, represented as "P: T". This suggests that the model implicitly identifies *passenger possession* as a key subgoal. Notably, this aligns with the structure of the Taxi environment, where incorrect pickup or drop-off actions incur a penalty of -10 , thereby reinforcing proper pickup and drop-off.

4.2 Quadratic VRAIL

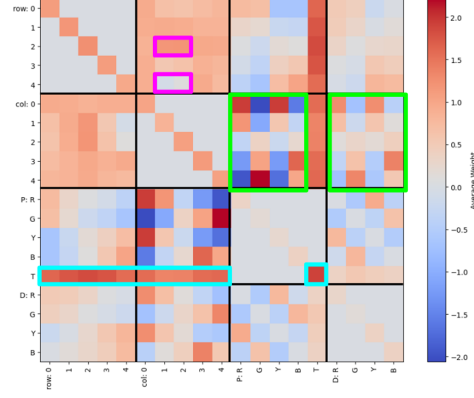
In Quadratic VRAIL, the value function is defined as $\hat{V}(x_s; W) = x_s^\top W x_s$, where $W \in \mathbb{R}^{19 \times 19}$ captures pairwise interactions between state features. We averaged the learned W over 9 converged seeds and visualized it as a heatmap. In fig. 5b, each axis represents a state feature, and cell intensity reflects the strength of interaction between feature pairs. In the **magenta** box, we observe interactions between spatial (row, column) features. Higher weights at accessible positions (e.g., (2,1)) indicate more frequent movement, while lower weights at blocked positions (e.g., (4,1)) reflect restricted mobility. In the **green** box, we observe strong interactions between column features and both passenger and destination locations. Although both show similar patterns, the passenger-related interactions are consistently stronger, suggesting that the agent prioritizes *pickup* before heading to the destination. Finally, the **mint** box highlights interactions involving the "P: T" feature (passenger is in the taxi). This region exhibits consistently high weights across multiple feature combinations, indicating that *passenger possession* plays a central role in shaping the agent’s policy.

5 Conclusion

We introduced VRAIL, a vectorized reward attribution method aimed at enhancing both learning effectiveness and interpretability in reinforcement learning. Built on a bi-level optimization framework, VRAIL shapes rewards using feature-based value function estimators, enabling agents to identify meaningful subgoals without modifying the environment. Experiments in the Taxi-v3 domain demonstrate that VRAIL improves both convergence stability and learning speed. Analysis of the learned parameters further reveals interpretable patterns in agent behavior, such as prioritizing passenger possession. These results highlight VRAIL’s potential for real-world RL scenarios with sparse or delayed rewards. As future work, we plan to extend our evaluation to more complex and partially observable environments such as MicroRTS and Crafter, to further assess VRAIL’s generalizability and effectiveness in dynamic, high-dimensional settings.



(a) Linear VRAIL (w)



(b) Quadratic VRAIL (W)

Figure 5: **Visualization of learned parameters of DL stage.** (a) Linear VRAIL contains the importance of each feature, (b) while Quadratic VRAIL contain the pairwise feature importance.

A Team member Contribution

All team members actively contributed to every stage of the project, including project concretization, trials for method improvement, slides, and report writing. Further individual contributions are detailed below.

- Jina Kim : Backbone model implementation (Linear VRAIL), DQN toy experiment (Fig 1), self-attention based model trials, final refinement of the report.
- Youjin Jang : Idea proposal (vectorized rewards), presentation, quiz creation, visualization (graphs, diagrams).
- Jeongjin Han : Interpretation of learned parameters of DL stage., Effect of shaped reward experiment (Fig 4), α -scheduling trials, environment variants trials.

B Hyperparameter Settings

We set the number of outer optimization cycles to 20 and use a discount factor of $\gamma = 0.99$. The exploration rate ϵ decays as $\epsilon(t) = \max(0.995^t, 0.01)$. In the RL stage, we use DQN with experience replay. The agent is trained for 100 epochs with a learning rate of $1e-3$, batch size 64, and replay buffer size 50,000. The target network is updated every 10 epochs. In the DL stage, we train for 50 epochs with a learning rate of $1e-2$.

C Trial and Error

We conducted a series of trial-and-error experiments to explore the effects of various design choices in reward shaping and value estimation in VRAIL. Our study considered the following four dimensions:

- **Environment variants:** Sparse reward version (Only +20 when arrival, -10 for illegal actions) was tested, but didn't work well at all.
- **State value estimators:** We explored a self-attention-based model design, but due to the absence of an additional dimension over which to apply softmax, we switched to a quadratic model.
- **Reward shaping bonuses:**
 - Global shaping using $\alpha \times \tanh(w^\top x)$
 - Element-wise shaping: $\alpha \times \sum_i \tanh(w_i x_i)$

These methods worked well for some seeds, but showed instability across others, likely because this form of shaping alters the agent's ultimate objective.

- **Alpha scheduling:** Constant, linear, sigmoid, exponential, and logarithmic schedules were tested to control the mixing coefficient α . These schedules showed meaningful improvements, but our latest model no longer requires α .

References

- [1] Gabriel Dulac-Arnold, Daniel Mankowitz, and Todd Hester. “Challenges of real-world reinforcement learning”. In: *arXiv preprint arXiv:1904.12901* (2019).
- [2] Volodymyr Mnih et al. *Playing Atari with Deep Reinforcement Learning*. 2013. arXiv: 1312.5602 [cs.LG]. URL: <https://arxiv.org/abs/1312.5602>.
- [3] Andrew Y. Ng, Daishi Harada, and Stuart J. Russell. “Policy Invariance Under Reward Transformations: Theory and Application to Reward Shaping”. In: *Proceedings of the Sixteenth International Conference on Machine Learning*. ICML ’99. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1999, pp. 278–287. ISBN: 1558606122.
- [4] David Silver et al. “Reward is enough”. In: *Artificial Intelligence* 299 (2021), p. 103535. ISSN: 0004-3702. DOI: <https://doi.org/10.1016/j.artint.2021.103535>. URL: <https://www.sciencedirect.com/science/article/pii/S0004370221000862>.
- [5] Mark Towers et al. “Gymnasium: A Standard Interface for Reinforcement Learning Environments”. In: *arXiv preprint arXiv:2407.17032* (2024).