Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence (IJCAI-25)
Doctoral Consortium Track

Reward Adaptation via Q-Manipulation: Provably Beneficial Reward Function Transfer in Reinforcement Learning

Kevin Jatin Vora
Arizona State University, School of Computing and Augmented Intelligence (SCAI)
kvora1@asu.edu

Abstract

Reinforcement Learning has made great strides in game playing and robotics but faces challenges with sample complexity and generalization. Transfer learning, which allows agents to reuse knowledge from prior tasks, offers a promising solution. My current research focuses on Reward Adaptation, where agents adjust to new reward functions while leveraging knowledge from tasks with different reward functions. I propose Q-Manipulation (Q-M), a method that adapts Q-functions to new rewards by computing and iteratively tightening bounds, akin to value iteration. This allows for action pruning before learning begins, enhancing sample efficiency without compromising policy optimality. Through empirical comparisons I demonstrate its effectiveness, generalizability, and practicality. Future work will handle changes in transition dynamics and continuous MDPs.

1 Introduction

High sample complexity of Reinforcement Learning (RL) remains a significant challenge, as it often requires vast amounts of data and interactions to learn optimal policies. Transfer learning is one solution that addresses this challenge, enabling agents to leverage knowledge from related tasks to accelerate learning in new ones, similar to humans. One critical but underexplored aspect of transfer learning is **Reward Adaptation (RA)**, where an agent adapts to a target reward function by utilizing learned behaviors from related reward functions in the same environment dynamics. For example, in autonomous driving, an agent must combine driving behaviors optimized for speed (e.g., goods delivery) and comfort (e.g., passenger transport), adapting seamlessly to a new task of transporting goods with passengers.

While learning from scratch is feasible, it is inefficient when source behaviors are available. Existing methods, such as Successor Feature Q-Learning (SFQL) [Barreto *et al.*, 2020] attempt to combine source behaviors to initialize learning in the target domain. However, they may struggle when target behaviors significantly differ from source behaviors, limiting their generalizability. To illustrate the RA challenge, consider the Dollar-Euro domain , shown in Figure 1. The agent starts at a yellow location, with terminal locations offering rewards for collecting dollars (pink) or euros (green). The task is to adapt behaviors optimized for dollars and euros into a new behavior that balances both. In such cases we need a more general transfer learning algorithm. My research contributes in two key areas: 1) the development of a general knowledge transfer method for RA, and 2) the establishment of a theoretical framework for transfer in RL.

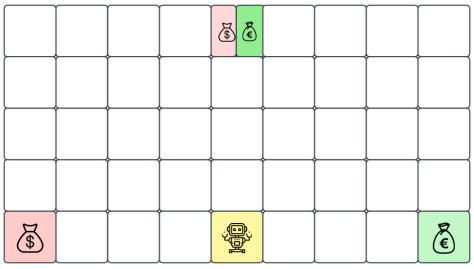

Figure 1: Dollar-Euro domain.

2 Proposed Approach

My research introduces "*Q-Manipulation*" (Q-M), a method that leverages the Q-functions of source behaviors to prune suboptimal actions, thus improving sample efficiency by reducing the exploration space. We assume the existence of a function, referred to as the *combination function*, that relates the source reward functions to the target reward function. Based on such a relationship, Q-M computes an upper and lower bound of Q-function in the target domain to identify actions that cannot contribute to the optimal behavior via an iterative process similar to value iteration. It enables us to prune those actions before learning the target behavior without affecting its optimality. Such a method is particularly advantageous in dynamic, real-world settings.

2.1 Problem Statement

Given an RA problem where variants of the Q functions are accessible for the source domains (e.g., Q^*'s (best returns policy) and Q^μ's (worst returns policy) under the source reward functions), determine the optimal policy under a target reward function $\mathcal{R}$ that is a known function of the source re-

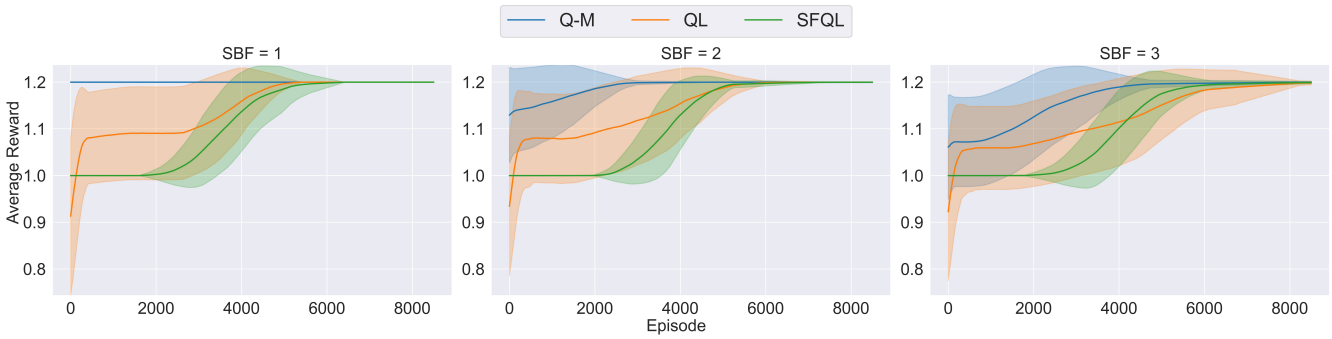

Figure 2: Convergence plots for Dollar Euro Domain comparing Q-M with Q-learning and successor feature Q learning

ward functions: $\mathcal{R} = f(R_1, R_2, \ldots R_n)$, To solve this, I propose Q-M, an action-pruning strategy that ensures only unnecessary actions are pruned.

2.2 Q-Manipulation

Q-M initializes upper and lower bounds of $Q^*_{\mathcal{R}}$ and iteratively refines them. Bounds are updated as follows:

Upper Bound (UB): **Upper Bound (UB)**

$$Q_0^{UB}(s,a) > Q^* \qquad \text{[Initialization]} \tag{1}$$

$$Q_{k+1}^{UB}(s,a) = \min\Bigg(Q_k^{UB}(s,a), \max_{s' \in \hat{T}(\cdot|s,a)} \Big[\mathcal{R}(s,a,s') + \gamma \max_{a'} Q_k^{UB}(s',a')\Big]\Bigg) \tag{2}$$

Lower Bound (LB)

$$Q_0^{LB} < Q^* \qquad \text{[Initialization]} \tag{3}$$

$$Q_{k+1}^{LB}(s,a) = \max\Bigg(Q_k^{LB}(s,a), \min_{s' \in \hat{T}(\cdot|s,a)} \Big[\mathcal{R}(s,a,s') + \gamma \max_{a'} Q_k^{LB}(s',a')\Big]\Bigg) \tag{4}$$

Here, $\hat{T}(\cdot|s,a)$ denotes one step reachable states from s, a. Intuitively, if an action a's lower bound is higher than some other action $\hat{a}$'s upper bound under a state s, then $\hat{a}$ can be pruned for that state. This results in a more streamlined set of actions, ensuring that the process remains significantly more efficient when learning in the target domain. This theoretical foundation is verified to work with reward transfer.

2.3 Convergence and Optimality

I establish that the proposed update rule is a contraction mapping, ensuring convergence to a fixed point. Furthermore, I demonstrate that it is a non-strict contraction, implying that the fixed point may not be unique. Additionally, I prove that actions retained after the tightening of bounds preserve optimal solutions. For a comprehensive discussion on initialization, experiments, and formal proofs, please refer to my paper[Vora and Zhang, 2025].

Theorem 1 (Convergence)**.** *The iteration process introduced by the Bellman operator in Q-M satisfies* $\|\mathcal{T}Q_k - \mathcal{T}Q_{k+1}\|_\infty \leq \gamma \|Q_k - Q_{k+1}\|_\infty, \forall Q_k, Q_{k+1} \in \mathbb{R}^{|S\times A|}$.

This ensures that the Q function converges to a fixed point.

Theorem 2. *The Bellman operator in Q-M specifies only a non-strict contraction in general:* $\left\|\mathcal{T}Q - \mathcal{T}\hat{Q}\right\|_\infty \leq \left\|Q - \hat{Q}\right\|_\infty$

Corollary 1 (Non-uniqueness)**.** *The fixed point of the iteration process in Q-M may not be unique.*

Theorem 3. *(Optimality) For RA with Q variants, the optimal policies in the target domain remain invariant under Q-M when the upper and lower bounds are initialized correctly.*

3 Prelimnary Result

Computation of UB and LB is affected substantially by the stochastic branching factor (SBF) of a domain, as evident in Eqs. 2 and 4. SBF here is defined as the maximum number of next states reachable (or with a nonzero transition probability) from any state and action pair. To demonstrate the influence of SBF, for each evaluation domain, we gradually increase its SBF. As shown in fig. 2, we observe that Q-M converges substantially faster than the baselines. However, as expected, the performance of Q-M is negatively impacted as SBF increases. Extensive evaluations in simulated and synthetic domains demonstrate that even when MDP is not designed or the reward is not handcrafted Q-M outperforms others.

4 Future Work

The theoretical foundation of Q-M extends beyond RA, offering potential to improve transfer RL. Future work will focus on enabling the transfer of transition dynamics between source and target tasks. Additionally, I aim to develop approximate solutions using function approximations for transfer in continuous MDPs. Q-M represents a significant step toward practical, scalable RL with the potential to transform AI learning and adaptation in dynamic environments.

References

[Barreto *et al.*, 2020] André Barreto, Shaobo Hou, Diana Borsa, David Silver, and Doina Precup. Fast reinforcement learning with generalized policy updates. *Proceedings of the National Academy of Sciences*, 117(48):30079–30087, 2020.

[Vora and Zhang, 2025] Kevin Vora and Yu Zhang. Reward adaptation via q-manipulation, 2025. https://arxiv.org/abs/2503.13414.