Term Project Proposal: Creating a marine environment in gymnasium and path finding optimization using MDPs

Benjamin Knöbel del Olmo , Florian Schechner

October 6, 2025

Abstract

We aim to design and implement a reinforcement learning environment in Gymnasium that simulates a boat navigating a continuous two-dimensional aquatic space. The agent controls two independent rudders to learn an optimal trajectory from a start to a target location while avoiding obstacles. The task is formulated as an infinite-horizon Markov Decision Process with a continuous state space and a finite discrete action set. This setup enables systematic evaluation of policy optimization algorithms under continuous-state dynamics defined by basic force and momentum balances, and it can be extended to include nonlinear effects such as drag or currents. The environment further provides a foundation for exploring advanced reinforcement learning methods, including off-policy control and policy-gradient approaches.

1 Motivation and Problem Definition

We consider a rowing-boat agent equipped with two independently controlled rudders, one on the left and one on the right side of the boat. Each rudder can perform three actions—row forward, row backward, or remain idle—resulting in a total of $3^2 = 9$ discrete action combinations. These actions define the agent's finite action space. The environment represents a river or lake modeled as a continuous two-dimensional plane in which the boat can translate and rotate. The agent's state is defined by its (x, y) coordinates and orientation, forming a continuous three-dimensional state space. While it is assumed, that the agent does not know the underlying motion dynamics, the simulator defines them explicitly through force and momentum balances. Each rudder is assigned a force magnitude and lever arm relative to the boat's center of mass, producing either forward or backward translation when both rudders act in the same direction, or rotational motion when they differ. Given an assigned mass and inertia, this setup allows realistic physical motion and feedback. The problem is formulated as an infinite-horizon Markov Decision Process (MDP). The reward function remains negative at each timestep until the goal location is reached, at which point the episode terminates. This encourages shortest-path behavior. Future extensions may include environmental disturbances such as water currents or obstacles, and a visualization module for analysis and demonstration.

2 Related Work

Gymnasium is a widely used platform for developing and benchmarking reinforcement learning (RL) algorithms. Many classical control and navigation environments—such as CartPole, MountainCar, and LunarLander—serve as canonical testbeds for evaluating policy optimization methods. These examples, documented in the official Gymnasium repository,[1] illustrate common environment designs and reward structures used throughout the RL literature. Beyond discrete control, continuous-control benchmarks such as Pendulum-v1 and BipedalWalker-v3 demonstrate how agents can learn smooth, physically consistent motion under continuous state and action spaces. Similarly,

[1] https://gymnasium.farama.org

navigation-oriented environments like grid worlds or CarRacing-v2 establish frameworks for goal-reaching and path-planning behaviors.

The novelty of our project lies in applying reinforcement learning to a new aquatic environment with distinct physical conditions. The boat is controlled through two independent rudders, resulting in nine discrete control combinations and continuous motion in position and orientation. This introduces a fundamentally different form of path finding compared to grid-world or terrestrial tasks. The combination of continuous dynamics, coupled control inputs, and the physical complexity of aquatic motion makes this environment significantly more challenging and distinct from existing Gymnasium benchmarks.

3 Solution Method

We focus on two main algorithmic families for solving the proposed control problem:

1. *Generalized Policy Iteration (GPI)* methods with function approximation, such as Semi-Gradient SARSA.

2. *Value-based* methods derived from Generalized Value Iteration (GVI), such as the Deep Q-Network (DQN).

These approaches are well suited for environments with unknown dynamics, continuous state spaces, and finite discrete action sets, as they can learn approximate value functions directly from sampled experience. Implementing both allows a systematic comparison between on-policy and off-policy learning behaviors.

For evaluation, we will benchmark each algorithm on the proposed boat environment using the following metrics

- **Success rate:** proportion of episodes in which the agent reaches the target.
- **Convergence speed:** number of learning iterations required to reach a predefined success threshold.
- **Solution optimality:** average path length or cumulative reward of the learned trajectory.

Depending on the available time and computational resources, additional methods (e.g., Actor–Critic or policy-gradient algorithms) may be implemented to validate theoretical expectations, such as differences in stability or convergence performance under function approximation.

4 Conclusion

This project develops a novel reinforcement learning environment within Gymnasium that models a rowing boat controlled through two independent rudders. The problem is formulated as an infinite-horizon Markov Decision Process (MDP) with a continuous state space and a discrete action space of nine possible control combinations. The simulator is based on basic force and momentum balances, enabling realistic motion and structured reward feedback for goal-directed navigation. The environment serves as a benchmark for evaluating the performance of classical reinforcement learning algorithms in continuous-state control tasks. In particular, we compare representatives of the Generalized Policy Iteration family (e.g., Semi-Gradient SARSA) and value-based methods derived from Generalized Value Iteration (e.g., DQN). Evaluation focuses on success rate, convergence speed, and solution optimality, providing empirical insight into algorithmic efficiency and stability. Beyond its immediate use for method comparison, the environment contributes a reproducible testbed for physics-based RL research in aquatic domains. Future extensions may incorporate additional complexities such as water currents, obstacles, or visualization tools. The expected outcome is both a functional simulation platform and a clearer understanding of how different learning architectures behave in continuous-state, physically grounded control problems.