Reinforcement Learning Project - Task 2

Abdelaziz Guelfane

abdelaziz.guelfane@student-cs.fr

GitHub Repository

Other Members:
Idriss Mortadi, Abdellah Oumida & Aymane Lotfi

April 2025

Abstract

This work extends the `highway-env` environment to train a reinforcement learning (RL) agent with continuous throttle and steering controls. We implement Proximal Policy Optimization (PPO) within an actor-critic framework to stabilize training and optimize for high-speed, right-lane driving while minimizing collisions. The agent is rewarded for maintaining lane discipline, high-speed driving, and penalized for off-road behavior. Through the use of continuous control, the agent learns smoother, more precise driving dynamics compared to a discrete-action DQN baseline. We evaluate the performance of the PPO agent, observing significant improvements in driving stability, reward accumulation, and reduced collision rates. However, the continuous control setting introduces greater training complexity and sample inefficiency compared to discrete approaches.

Contents

1 Task 2 : Continuous Environment

1.1 Continuous Actions for Highway Driving

We extend *highway-env* to train an RL agent with continuous throttle and steering over 60s episodes. Rewards encourage right-lane driving (+0.5 pts/s), high speed (+0.1 pts/s at max), and penalize collisions (-1 pt).

- **Action space**: continuous 2D (`throttle`, `steering`)
- **Timing**: simulation at 5 Hz; policy updates at 1 Hz (each action held for 5 steps)
- **Constraints**: steering changes limited to once per second; negative reward for off-road driving

These settings promote steady, lane-centered driving behavior while preventing oscillations and off-road exploitation.

1.2 Actor-Critic PPO Implementation

1.2.1 Algorithm Overview

Proximal Policy Optimization (PPO) is an on-policy, policy-gradient method that updates policies in a constrained way to prevent unstable parameter shifts. Our implementation:

- Uses an actor-critic framework: actor selects actions, critic evaluates expected rewards
- Collects full-episode trajectories, then performs multiple epochs of PPO updates
- Employs clipped surrogate objective to limit policy changes

The PPO loss is given by:

$$L^{\text{CLIP}}(\theta) \;=\; \mathbb{E}\big[\min\big(r_t(\theta)\,A_t,\;\text{clip}(r_t(\theta),1-\epsilon,1+\epsilon)\,A_t\big)\big],$$

where

$$r_t(\theta)=\frac{\pi_\theta(a_t \mid s_t)}{\pi_{\theta_{\text{old}}}(a_t \mid s_t)},\quad A_t=\text{advantage estimate},\quad \epsilon=0.2.$$

The clipping prevents probability ratio changes beyond ±20% in a single update, providing stability.

1.2.2 Network Architecture

While we initially experimented with a shared-backbone architecture, we ultimately implemented separate networks for actor and critic to avoid gradient interference issues in PyTorch:

- **Input:** flattened occupancy grid (vehicle positions & velocities & cosine, sine of heading)
- **Actor network:**
  - Two fully-connected layers (512, 128 units) with ReLU
  - Outputs mean $\mu(s)$ and standard deviation $\sigma(s)$.
- **Critic network:**
  - Two fully-connected layers (512, 128 units) with ReLU
  - Outputs state value $V(s)$ for advantage computation
- **Action sampling:** $a_t \sim \mathcal{N}(\mu(s_t),\sigma(s_t))$, clipped to environment bounds

The critic estimates $V_\phi(s)$ by minimizing:

$$L_{\text{critic}}=\mathbb{E}[(V_\phi(s_t)-R_t)^2]$$

where $R_t=\sum_{k=t}^{T}\gamma^{k-t}r_k$ represents cumulative discounted rewards. The advantage $A_t = R_t - V_\phi(s_t)$ reduces gradient variance during policy updates.

1.2.3 Training Challenges & Solutions

Key issues included spinning, off-road driving, and unstable control. We addressed these by:

- Capping steering changes to $\pm 0.1\,\text{rad/s}$
- Penalizing off-road behavior
- Adding entropy bonus (0.01) for stable exploration

Progress was monitored via TensorBoard, confirming reward growth from near 0 to over 120 and improved driving stability.

1.3 Performance Analysis

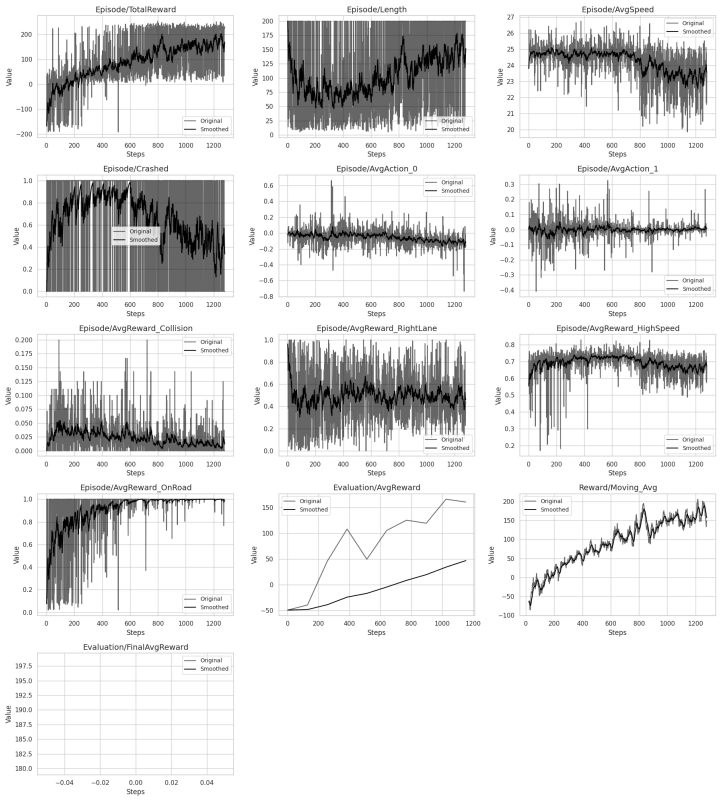


Figure 1: Training and evaluation metrics of the PPO agent on `highway-fast-v0` with continuous actions. The plots show the evolution of rewards, episode lengths, collision rates, and control signals over 1200 training steps. Smoothed curves highlight trends in performance, including increased total reward, improved driving stability, reduced crashes, and consistent high-speed lane adherence.

The PPO agent demonstrated significant performance gains over the course of training, as evidenced by smoothed metrics across key behavioral indicators. Here is an example of an episode (Google Drive link).

- **Reward progression:** The total episodic reward (Fig. 1) shows a clear upward trend, stabilizing above 120, aligning with consistent high-speed and right-lane adherence. The moving average further confirms long-term improvement.
- **Driving duration:** Episode lengths (Fig. 1) increased steadily, suggesting improved survivability and fewer collisions. This is corroborated by the decline in the crash rate (Fig. 1).
- **Control behavior:** The average values of actions (steering and throttle) show reduced variance, indicating more consistent and purposeful driving. The average speed (Fig. 1) remains near 25, suggesting the agent maintains a high but controlled pace.

- **Reward decomposition:** The agent optimizes for task-specific goals:
  - **Right-lane adherence** (`AvgReward_RightLane`): stabilized around 0.5, indicating successful lane preference.
  - **High-speed driving** (`AvgReward_HighSpeed`): values converge to 0.7, close to maximum possible.
  - **Collision avoidance** (`AvgReward_Collision`): low and decreasing, showing avoidance behavior is learned.
  - **On-road consistency** (`AvgReward_OnRoad`): converges to 0.9, suggesting the agent largely remains within road boundaries.
- **Evaluation performance:** Periodic evaluation rewards (Fig. 1) confirm policy generalization and transferability across episodes. The sharp rise with reduced variance reflects robust policy convergence.

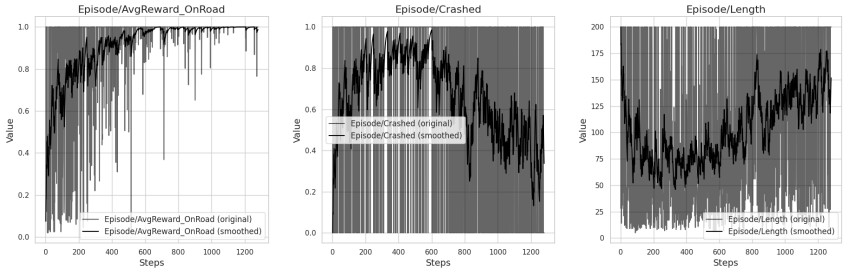
Despite early instability and suboptimal behaviors (e.g., spinning or erratic steering), the agent ultimately learned to balance speed, lane discipline, and safety. This supports PPO's capacity to handle continuous control in complex driving tasks when paired with careful reward shaping and training constraints.

1.4 Continuous vs. Discrete Actions

We compared our continuous PPO agent against a discrete-action DQN baseline:

1.4.1 Learning Dynamics

The DQN agent, operating in a discrete action space, exhibited faster initial learning due to simpler exploration dynamics. In contrast, the PPO agent with continuous controls required more time to acquire basic driving behaviors before improving its ability to avoid collisions.



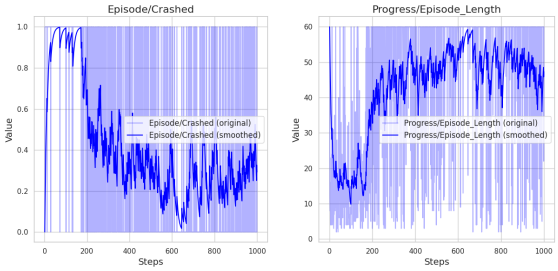
(a) On-road reward (b) Crashed episodes (c) Episode length

Figure 2: Training progress of the PPO agent: Initially, both the on-road reward and number of crashes increase as the agent learns to navigate. Over time, crash frequency decreases while episode lengths increase, indicating improved driving performance.

1.4.2 Implementation Tradeoffs

Continuous control advantages came with increased complexity:

- **Training complexity:** The actor-critic structure of PPO is more sensitive to hyperparameter settings compared to the simpler DQN setup. Achieving stable learning required tuning multiple components such as learning rate, clipping range, and advantage estimation parameters.
- **Sample efficiency:** PPO, being an on-policy algorithm, required significantly more interactions with the environment to learn effectively (1280 episodes). In contrast, DQN's off-policy nature allowed for more efficient reuse of past experiences, leading to faster convergence in terms of sample count (800 episodes was enough to reach reward plateau).



(a) Crashed episodes (b) Episode length

Figure 3: Training progress of the DQN agent: Contrary to PPO agent, DQN agent starts directly to minimize crashes which results in direct increase of episodes lengths.

1.5 Conclusion

We have demonstrated that an on-policy PPO agent with continuous throttle and steering can learn stable, high-speed, right-lane driving in `highway-env`, outperforming a discrete-action DQN baseline in terms of driving precision and safety. By enforcing action smoothing, steering caps, and tailored reward shaping, the PPO agent achieved consistent rewards above 120, low collision rates, and maintained on-road adherence.

- **Continuous control benefits:** Enables fine-grained speed and steering adjustments, yielding smoother trajectories and better lane discipline.
- **Training tradeoffs:** Requires more samples (1280 vs. 800 episodes) and sensitive hyperparameter tuning (learning rate, clipping range, entropy bonus) compared to DQN.
- **Stability mechanisms:** Steering rate limits, off-road penalties, and entropy regularization effectively prevent oscillations and unsafe maneuvers.