AGH University of Krakow

FIELD OF SCIENCE: ENGINEERING AND TECHNOLOGY

SCIENTIFIC DISCIPLINE: AUTOMATION, ELECTRONICS, ELECTROTECHNICS AND SPACE TECHNOLOGIES

SUMMARY OF ACCOMPLISHMENTS

Prediction model of an autonomous vehicle's behavior, based on Artificial Intelligence methods

Author: Nikodem Pankiewicz

Supervisor: dr hab. inż. Piotr Bania

Completed in: AGH University of Krakow, Faculty of Electrical Engineering, Automatics, Computer Science and Biomedical Engineering

Kraków, 2024

Abstract

We are witnessing rapid automotive development and an increasing demand for newer and more versatile driver assistance systems (ADAS). With the accompanying intensive development of artificial intelligence based on deep neural networks, it is possible to replace classical solutions and extend the scope of their operation. One potential area of profitability for AI is in solving the problem of vehicle behavior planning.

The thesis focused on inventing a methodology for developing a vehicle behavior policy in adaptive cruise control mode. The policy was designed to plan a target acceleration value which was achieved in a designated time horizon by planning and executing a trajectory. The strategy was intended to operate on highways, and the planning took into account the road situation, which included the road topology and all vehicles detected by the vehicle's perception system. The policy aimed to be effective in a real environment and handled a broad range of road situations. Research was conducted to explore the potential of using reinforcement learning (RL) and imitation learning methods.

Preliminary analysis showed that imitation learning, which was based on collected real data, was limited by a restricted amount and poor data quality. On the other hand, the policy optimized by reinforcement learning methods in simulation often appeared to be suboptimal in real-world conditions. It was due to the deviation between simulation and reality (sim2real gap) and missing replicated real-world phenomena.

The dissertation presented methods to mitigate these problems. Firstly, it presented a method of improving the quality of real-world data using numerical optimization. Secondly, it presented a way to combine both learning methods to minimize the sim2real gap and increase the distribution of situations known to the agent.

The policies obtained using the presented solutions were assessed using the proposed evaluation methods. Additionally, the policy effectiveness was compared with the performance of a test driver and a baseline policy trained with the standard RL approach.

Furthermore, the thesis presented a methodology for evaluating the reliability of the behavior model using statistical analysis to examine the impact of input elements into the system on the selected action.

The results suggested that the proposed algorithms are promising from the perspective of developing vehicle behavior planning. The presented evaluation method allowed a better understanding of the predicted behavior policy and increased its reliability.

1. Motivation

Recently the automotive industry has been working towards minimizing car accidents, mainly fatal ones, caused by human error. They implement advanced assistance systems to prevent common errors, support decision-making, and monitor driver attention. The ultimate goal is to replace human drivers with fully automated vehicles. To achieve this, a holistic system of perception, behavior planning, and vehicle control modules must be developed and integrated to ensure safe and efficient driving.

Among others, behavior planning is a demanding field that requires more attention. In the past, solving decision-making problems involved coding responses into decision trees or procedural algorithms. Although this approach offered control and transparency, it was cumbersome and prone to errors without ensuring optimal solutions. Nowadays, data-based methods that derive behavior rules from analyzing a wide range of traffic situations to optimize efficiency are considered more promising. One such method is Reinforcement Learning (RL) which optimizes the behavior policy according to some reward function (control objective).

For driving applications, using Reinforcement Learning (RL) directly on real roads is impractical due to safety, speed, and cost concerns. These concerns make computer traffic simulations a preferred training environment. However, policy performance in real conditions depends on the reproduction of real-world phenomena in simulation. Transferring a policy to real-world conditions reveals its limitations, potentially encountering unseen states, variations in state perceptions, distinct state transitions, or outlier behavior of traffic participants. This discrepancy, known as the sim2real gap, poses a significant challenge in applying RL to real-world applications.

Given the above thesis focused on developing methods that alleviate the sim2real gap and distributional shift issues. The proposed methods integrated simulated data with real driving data in the learning process. The approach derived inspiration from offline and online Reinforcement Learning. Offline RL is used for learning only the static dataset of experience collected in advance, usually by some human expert. This type of data harvesting provides close to optimal real-world experience, facilitating learning and eliminating exploration problems. However, the limited scope of the dataset restricts policy to handle a wide range of situations. On the other hand, online RL allows unrestricted exploration of the simulated environment, therefore, learning all possible state transitions and action consequences. It leverages offline learning in terms of explored situations, however, simulated experience is affected by the aforementioned sim2real gap.

The advantages of these two approaches seem to mitigate each other's drawbacks. Therefore the proposed methods concentrated on the development of techniques that enabled applying data from various sources to train behavior-driving policy which would be efficient in a real environment. At the same time, the work studied the disadvantages of offline learning and proposed a method for increasing the quality of real data to support the learning process.

Besides the development process, the thesis presented a novel approach to understanding the decisions made by ANN-based policy. The method allowed comprehending the pattern of policy behavior and cross validate it with human intuition.

2. Project Foundations

2.1 Control Objectives

The thesis focused on enhancing the development process of the Adaptive Cruise Control (ACC) driving mode with reinforcement learning-based methods. The ACC system allows users to set a maximum desired velocity for the vehicle, which is only pursued when the road ahead is empty and free of obstructions. Otherwise, the system dynamically adjusts the vehicle's speed to maintain a safe distance from other vehicles, either preceding or intending to merge into the host lane. It should prioritize safety and preemptively respond to emerging situations.

To provide a comfortable ride for passengers, the ACC system follows specific acceleration and deceleration limits. This feature reduces sudden movements and contributes to a smoother driving experience. The system's capability to choose and modify the speed according to the actions of the target vehicle, whether it's directly ahead or planning to merge into the host lane, is crucial in order to maintain a consistent velocity and prevent abrupt speed changes.

The objectives of the ACC system are to maximize the configured speed while ensuring safety, enhance driving comfort by reducing speed oscillations caused by the movement of other vehicles, and minimize instances of heavy braking and unsafe lane changes. The system's target selection and predictive capabilities are key to achieving these goals, helping to maintain a steady flow of traffic and accommodate both high-speed travel and congested conditions on highways.

For the foundation of the project, the thesis formulated a system as a Partially Observed Markov Decision Problem (POMDP). The action in POMDP was provided by some stationary policy π parameterized by vector θ based on the observation $o_{s,t}$ of the current state in time t. The policy π was optimized concerning the given objective function by methodology presented in the thesis.

2.2 Objective function

The control objective was to maximize the sum of cumulative rewards:

$$J = \left(\sum_{t=0}^{T-1} \gamma^t R(s_t, u_t, s_{t+1}) \right)$$

The rewards function was multifactorial and its general form was constituted such as:

$$R(s_t, u_t, s_{t+1}) = -(v_{s,\max} - v_{t+1}^{\text{vcs}})^2 - (a_{s,t+1}^{\text{vcs}})^2 - c_0(\Delta s) - c_1(\Delta s)$$

where v_{t+1}^{vcs} is the absolute velocity of the controlled object in the Vehicle Coordinate System in state s_{t+1}; $v_{s,\max}$ was targetted preset velocity; a_{t+1}^{vcs} was absolute acceleration in VCS achieved by executing u_t in s_t.

$$c_0(\Delta s) = \begin{cases} 1, & \text{if } \Delta s \leq d_{\text{lon\_min}} \\ 0, & \text{if } \Delta s \geq d_{\text{lon\_min}} \end{cases}$$

where $d_{\text{lon\_min}}$ was the minimal longitudinal distance defined according to the Responsible Sensitive Framework. Δs was the distance alongside the lane centerline between the front bumper of the controlled vehicle and the rear bumper of the vehicle in front. The term c_1 determined the collision event of the host and front vehicle.

2.3 Motion Stack

The policy was part of the holistic motion planning stack (Fig. 2.1) which consists of consecutive modules :

- **Route planning** determines the optimal path from the current location to the destination, relying on high-definition maps that detail the road network, traffic regulations, and environmental features. This module guides the vehicle on which lanes to follow to reach its destination.
- **Behavior Planning Module** decides on the driving strategy by selecting maneuvers or setting parameters like speed and acceleration, based on data from the perception module, routing instructions, and user preferences. This includes adapting to dynamic traffic conditions to maintain safety and efficiency.
- **Trajectory Generation Module** plans a detailed path that the vehicle will follow, considering the vehicle's dynamics and the surrounding traffic. This trajectory is crafted to ensure safety, efficiency, and comfort, taking into account the desired maneuvers from the behavior planning module.
- **Control Module** executes the trajectory by adjusting the vehicle's actuators—throttle, steering, and brakes—according to the current state of the vehicle and the planned path, ensuring that the vehicle adheres to the intended trajectory accurately.

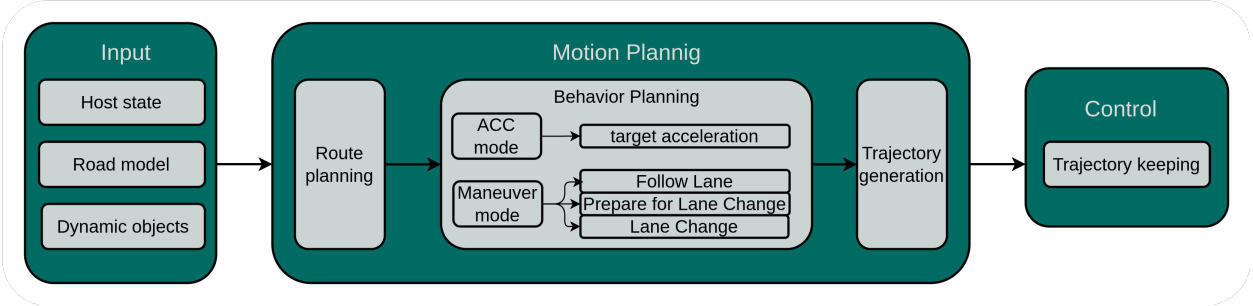

Figure 2.1: Motion planning architecture. The optimized policy was part of the behavior planning module.

3. Experimental Setup

3.1 Policy definition

The policy was part of the behavior planning module and was responsible for planning the continuous value of acceleration that should be reached within 0.5s by trajectory generation and control modules. The policy was defined as a normal probability density function in which mean and standard deviation parameters were returned by Artificial Neural Network (ANN). The ANN weights were subject to the optimization process. The policy returned values in the range of $< -3.5, 1.5m/s^2 >$.

3.2 Neural Network Architecture

The thesis proposed modular ANN architecture (Fig. 3.1) that consists of three major modules:

- **Perception Module** processed separate data for the host, targets, and roads, using straightforward feed-forward layers for hosts and targets. A more complex Graph Neural Network was used for road observation, which accounts for the intricate structure and logic of road networks. The encoded features and one additional trainable input were passed as tokens into the Transformer Encoder layer. The feature output token associated with the trainable input token should represent all relevant details about the traffic scene in the latent state.

- **Brain Module** had a form of LSTM layer. It was designed to remember key past features for control tasks. It filtered and combined current data with memory to optimize future decisions.

- **Control Module** included one fully connected layer with tanh activation function and a single trainable parameter. The first generated the mean value of Normal Distribution μ and the latter specifies its natural logarithm of standard deviation $ln(\sigma)$.

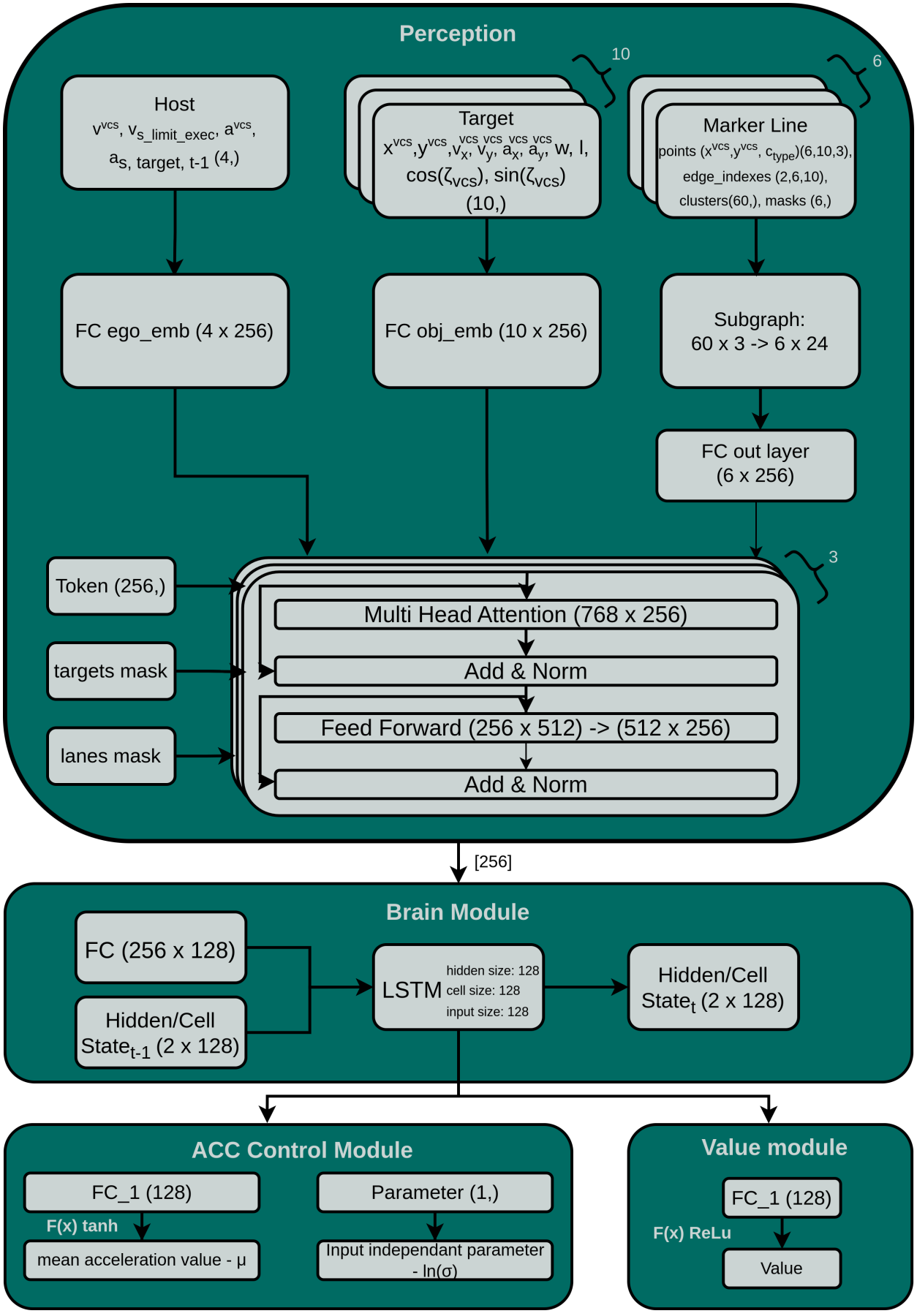

Figure 3.1: ANN consists of 3 major modules: perception, brain, and control. The subgraph is shown in detail in Fig. 3.2

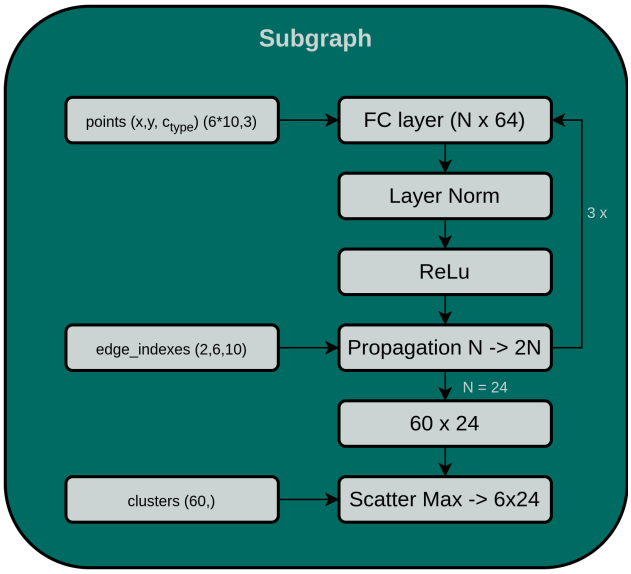

Figure 3.2: Subgraph is a part of ANN 3.1 which processes the lanes geometry which is represented as a directed graph.

3.3 Dataset

The training dataset used for experiments consisted of 672211m driving experience. The data set covered various traffic situations, road geometries, weather conditions, and broad traffic flow distribution. The data set comprised around 60000 state-expert action pairs sampled every 0.5s from raw data. Data was collected on the highways and roads outside cities, mainly in Germany. Raw data was collected from vehicle sensors such as the front-facing camera, all-around radars, and internal car sensors such as IMUs and wheel encoders. The raw data was processed to meet the requirements of the reinforcement learning training framework (SARS' tuple). The recorded acceleration values were recalculated based on the achieved velocity, as defined by the action in the POMDP. This transformation was necessary because the driver did not steer the vehicle according to the system interface, and due to the presence of noisy sensor readings.

4. Methodology

4.1 Baseline PPO Training

The thesis delivered the baseline optimization of policy in the form of Reinforcement Learning training. The training utilized on-policy RL algorithm Proximal Policy Optimization (PPO) and simulated highway environment with multiple lanes and various traffic densities. It was executed with 80 distributed CPU workers for experience gathering and a GPU-based trainer for policy optimization. This comprehensive process spanned around 12 million steps, mimicking 74 days of continuous driving and covering a distance of approximately 144038 km. The optimization process was structured around collecting a training batch, subdividing it into mini-batches for sequential optimization, and iterating this process a specified number of times, with batch sizes and the iteration count being critical to the efficiency of the training regime.

As the training progressed, the control module parameter $ln(\sigma)$ was decreasing, leading to a more constrained distribution of action selection. The gradual reduction limited the agent's exploration indicating growing confidence in action choices. This training phase sets a foundational PPO policy, serving as a benchmark for evaluating the effectiveness of subsequent, more advanced methodologies against this baseline performance.

4.2 Improving Experts' Experience

Offline Reinforcement Learning (RL) algorithms are often affected by ground truth data imperfections. To solve this issue, the thesis proposed a novel method called Optimization-based Imitation Learning (ObIL). The main idea behind it is to optimize the expert's actions before using it in training. Experts' actions are often suboptimal due to human factors and the inability to predict future traffic situations perfectly, which is crucial for selecting appropriate control signals.

The ObIL method suggested a two-step algorithm. First, the experts' actions are improved using gradient optimization methods. In the second step, the improved actions are applied in policy optimization, utilizing the Imitation Learning process.

The feasibility of this method was assumed in the case where the actions of the agent only affect the agent's state and have a negligible effect on the rest of the environment. The effectiveness of this method was based, in the first place, on leveraging the assumption that all states in the trajectory are already known, therefore, the motion prediction of adjacent vehicles is no longer necessary and prediction imperfection does not degrade action selection. The idea behind this approach to offline learning is to provide the trajectories with enhanced actions that would benefit the agent training. It is believed that the agent will be able to

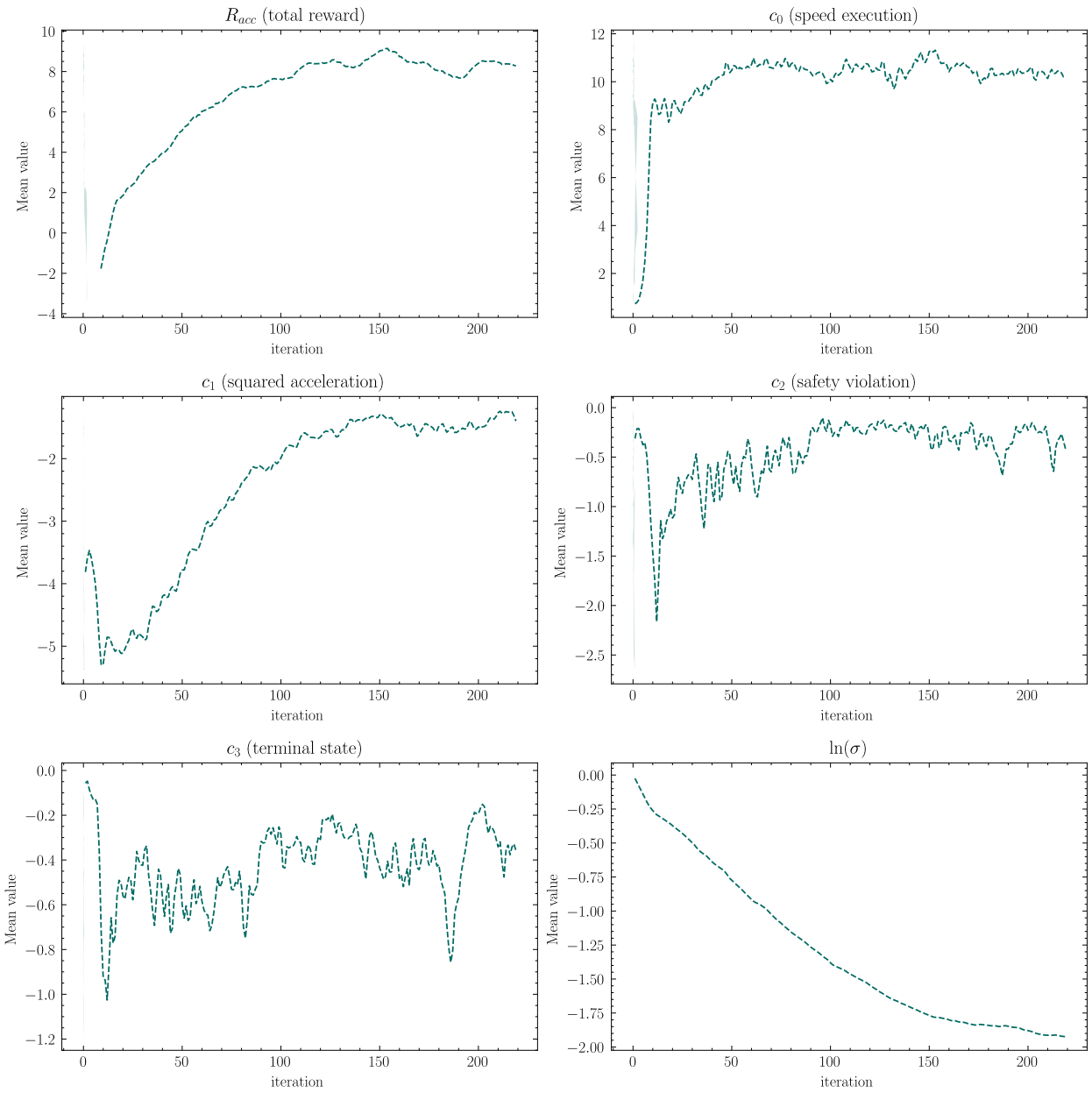

Figure 4.1: The course of training baseline PPO agent in terms of collected rewards and value of $ln(\sigma$ parameter.

recognize patterns in the data and utilize them as a basis for learning, ultimately exceeding the performance of the human expert.

The method assumed that the expert trajectory contains a set of actions that could be parameterized by the number of parameters, which are subject to optimization. The optimization process aims to minimize the cost function which is coherent with the ACC objective function and defined constraints such as maximal and minimal acceleration and distance to the leading vehicle. The thesis presented an example in which acceleration was expressed as a continuous BSpline function. BSpline function consisted of the number of coefficient knots that were subject to optimization. The SLSQP optimization method was selected for the gradient optimization process.

The thesis presented proof of concept which empirically verified that policy trained on optimized trajectories outperformed policy trained on original samples. The OBiL was then applied in optimization of ACC policy which is the subject of the thesis. First of all optimization process was conducted on all trajectories that were included in the training dataset. The example of the original and optimized trajectories are shown in Figure 4.2, 4.3.

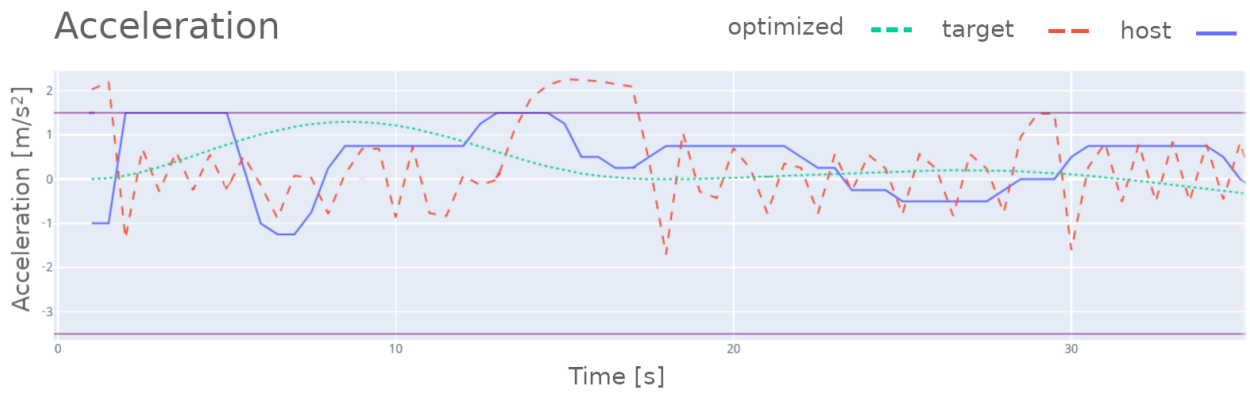

Figure 4.2: The acceleration of the leading target (red), human driver (blue), and optimized one (green). The horizontal lines represent the limit values of acceleration (-3.5, 1.5 m/s^2).

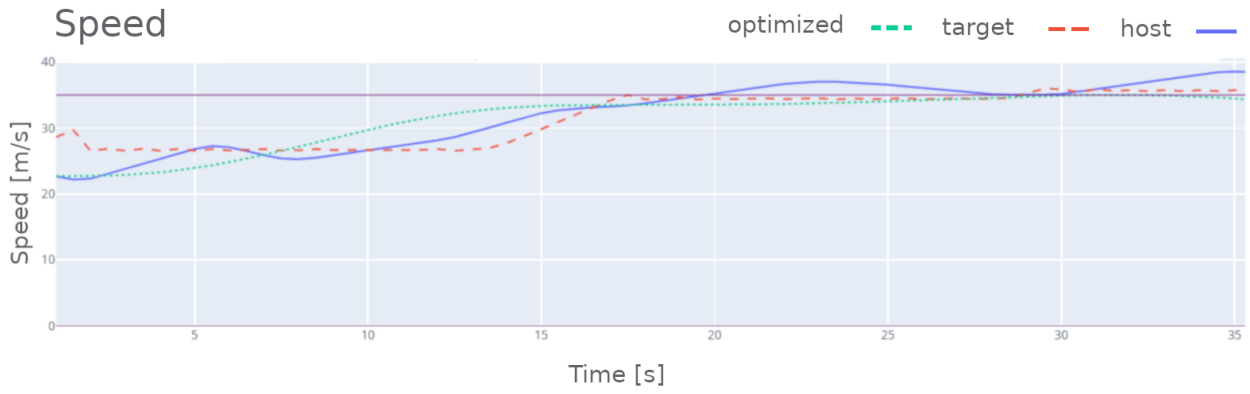

Figure 4.3: The speed of the leading target (red), the human driver (blue), and the speed calculated from the optimized spline function which represents the acceleration function (green). The horizontal lines represent the limit values of velocity (0, 35 m/s)

4.3 Offline Learning on Dataset

To achieve policy optimized against the real and simulated data, the imitation learning trainings were conducted on original and optimized datasets through Behavioral Cloning and MARWIL algorithms. The conduct of learning, from the perspective of training and testing losses, is depicted in Figures **??**. The thesis proposed several improvements to tackle the emerging problems and enhance the training process.

During initial attempts, the training and testing losses revealed significant variance in optimization steps, suggesting the initial batch size was too small. To address this, the accumulated gradients algorithm was

implemented, allowing for an effective increase in batch size without exceeding VRAM limitations by accumulating gradients across mini-batches before applying optimization.

Further enhancements included addressing the issue of favoring extreme acceleration values by policy. It resembled the problem of an unbalanced dataset in supervised learning. To address this, the thesis proposed to sample equally from clustered datasets. The clusters were created based on the range of action values. The best result was achieved with two clusters. One consisted of actions that correspond to decelerations and the other with accelerations.

The final improvement was made by updating the LSTM hidden states in samples after each optimization step. It was necessary to incorporate modifications in the LSTM weights as a result of optimization. The update significantly decreased computational requirements and made the training process simpler by eliminating the need to specify the number of samples to select for burning hidden states. This approach ensured that the policy remained adaptable and effective, reflecting the dynamic nature of continuous episode predictions.

The offline training generated four distinct policies evaluated based on the proposed criteria. The comparison identified that the best policy, from the perspective of the ACC objective, was the one trained with the MARWIL algorithm on an optimized dataset. This policy was used for further training in the final step of the proposed algorithm.

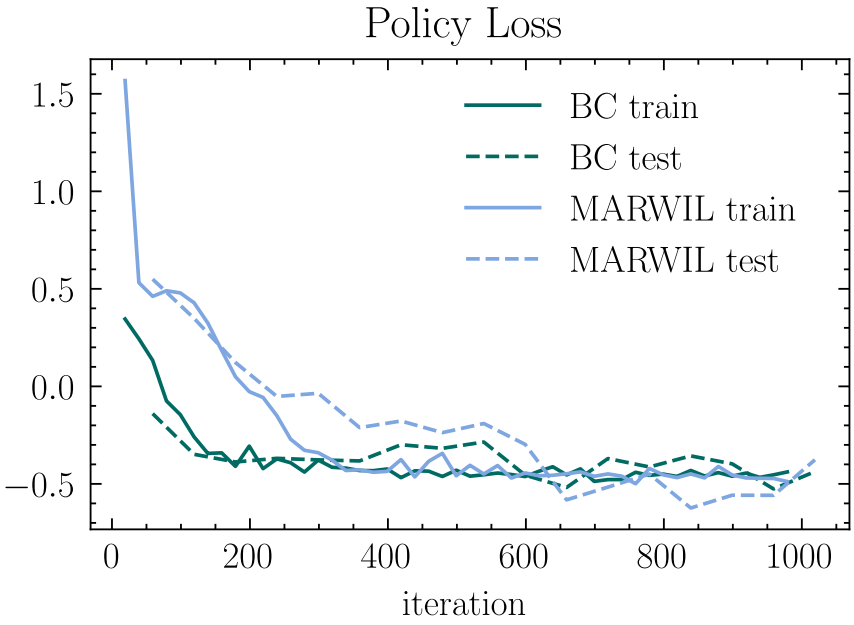

Figure 4.4: Training and testing loss of BC and MARWIL agent on original dataset D_{train}.

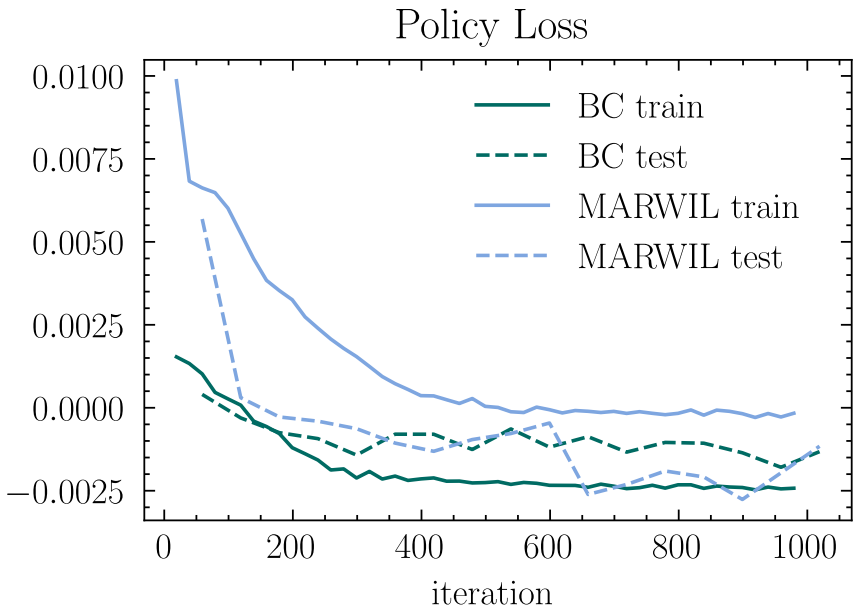

Figure 4.5: Training and testing loss of BC and MARWIL agent on optimized dataset D_{opt}.

4.4 PPO Training with Real Data

The policy trained with MARWIL on the optimized dataset was further enhanced through the online reinforcement learning method PPO within a simulation environment. Such an enhancement was essential for optimizing the policy in scenarios not covered by the dataset. It leveraged the simulation to generate a limitless variety of traffic situations through the Domain Randomization method incorporated in scenario creation. The online approach also introduced stochastic behavior into policy, which led to intensified environment exploration by interactions with simulated traffic participants.

A significant focus was to prevent catastrophic forgetting of previously learned real-world experiences. The thesis proposed to include resimulation of logged scenarios in the portfolio of randomized scenarios. The process of log replay relied on reproducing detected objects and road lanes in consecutive steps of simulation while executing the host trajectory as required by policy.

The method integrated real-world scenarios as a unique type among those generated by the scenario generator, allowing the agent to optimize its policy on new situations, revisit known ones, and learn transitions close to dataset states. These were vital for adjusting the agent's behavior in situations that deviated from known trajectories.

It was likely observed that resimulation may not accurately reflect the actual environment and could include issues known as the sim2real gap. However, using such samples could provide observations that are affected by real sensors and real traffic situations, which can decrease the gap between the simulated and actual world. Combining real trajectories with simulated ones provides another approach to the sim2real gap problem, in addition to domain randomization or adaptation methods.

The superior goal was to create a comprehensive distribution of scenarios that mirror expected real-world conditions in terms of road parameters, vehicle dynamics, and driver behaviors. This aimed at an

approximation of the scenario distribution to real-world conditions, ideally broader to ensure coverage of most potential situations without compromising policy quality by overgeneralizing to implausible scenarios.

The combination of generated scenarios and those based on driving logs was recognized as not perfectly matching the natural distribution but was seen as a step towards achieving a more accurate approximation of real-world conditions.

The training parameters were the same as used in the baseline PPO training. The training lasted for 400 epochs, and the course of training in the context of rewards optimization is visible in Figure 4.6.

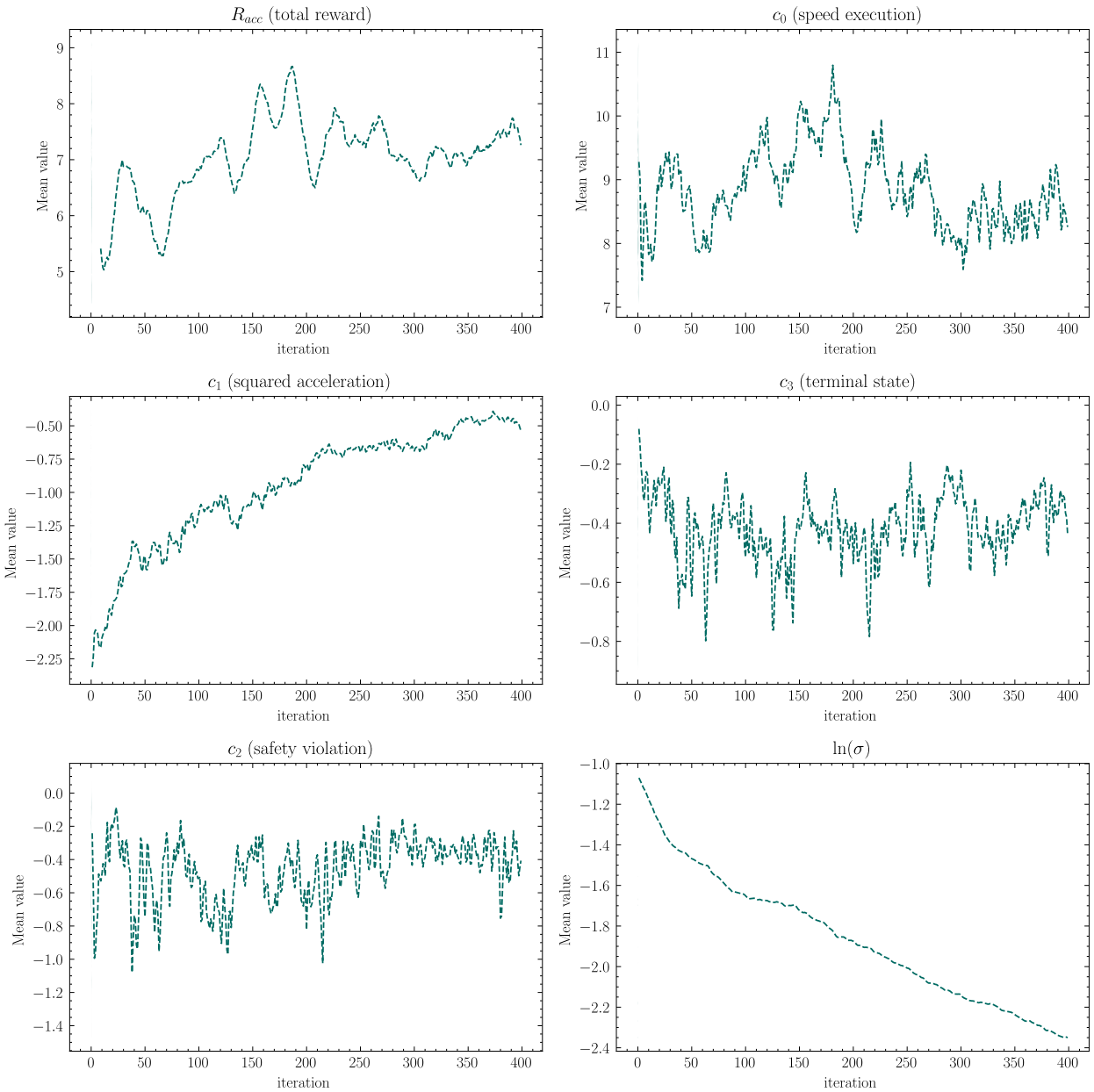

Figure 4.6: The plots of collected rewards during fine-tuning of MARWIL policy and ANN parameter $\ln(\sigma)$. The training was conducted on the ACC environment enhanced with scenarios created based on replaying of driving logs.

4.5 Evaluation

To assess the performance of all obtained policies, the mean sum of rewards from the reward function was used as the primary performance metric. Additionally, the evaluation involved tracking the mean value of each reward component to understand their impact on the total reward.

A set of Key Performance Indicators (KPIs) was defined to measure various driving performance aspects not captured by the reward function. It aided the reward function's design and adjustment by monitoring learning progress and optimizing multicriterial reward components. The KPIs were calculated during the training and evaluation phases. They included metrics related to speed, adherence to speed limits, acceleration, jerk, safety distance violations, heavy braking events, and others, with each KPI normed for standardized assessment. This approach allowed for a comprehensive comparison of different strategies and the exclusion of agents exploiting the reward system, ensuring a holistic evaluation of driving performance.

The evaluation phase was conducted after each training. In that phase the agents operated without exploration, directly passing to the trajectory generation module the mean acceleration value returned by the ANN. This phase was structured into three steps:

- **Closed Loop Testing In Simulation**: Agents were tested across 100 randomized episodes within the same training environment, with scenario distributions uniformly applied, assessing agent behavior in varied, randomized conditions.
- **Predefined Scenarios Testing**: Agents underwent testing in scripted scenarios to assess their responses to typical road situations, ensuring consistency across tests. This stage identified whether agents act as expected or deviate significantly, prompting further training or adjustments. This also facilitated a detailed analysis of agent behavior, aiding in reward function tuning and training scenario adjustments.
- **Testing on Logs**: In an open-loop setup, the agent's actions were executed in resimulated scenarios while maintaining the original trajectories of other vehicles. It was crucial for evaluating agent performance under real-life conditions and for comparing policy with expert performance.

The thesis methodology output a final policy which was trained initially on the optimized dataset using MARWIL and then finetuning with PPO in simulation. This policy achieved the highest mean sum of rewards among tested agents in resimulated scenarios. The policy excelled in balancing comfort and safety while optimizing vehicle speed, outperforming the human driver and the baseline PPO agent in terms of lower average absolute acceleration and fewer safety violations. Although the policy's driving behavior mirrored human drivers, it exhibited greater caution, maintaining larger distances from leading vehicles without significantly increasing safety distance violations. It adeptly adjusted velocity in response to the fluctuating speeds of leading targets. However, it was noticed that it required further refinement to address action flickering and other unwanted behaviors.

The evaluation highlighted a discrepancy in learning efficiency, with agents mastering complex tasks more readily than simple ones like maintaining speed in an empty lane. It suggested that a scenario portfolio was biased towards complicated ones over fundamental driving skills.

The course of the target, agent, driver, and optimized trajectories in the example log are shown in Figures 4.7 and 4.8.

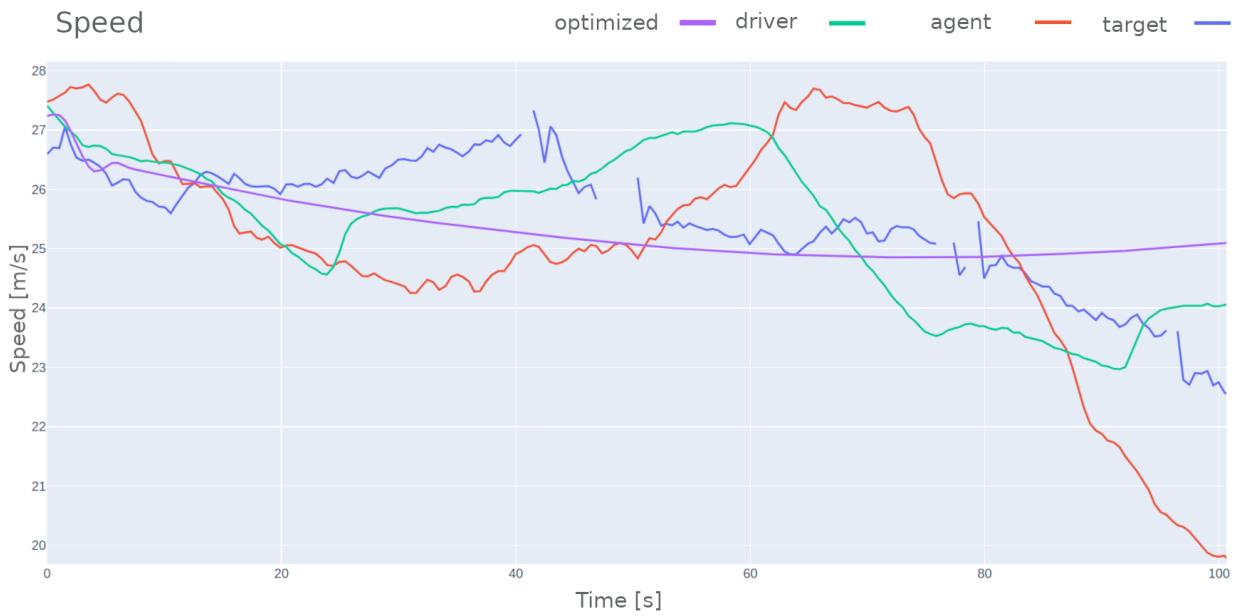

Figure 4.7: Comparison of executed velocity by different policies evaluated on resimulated log.

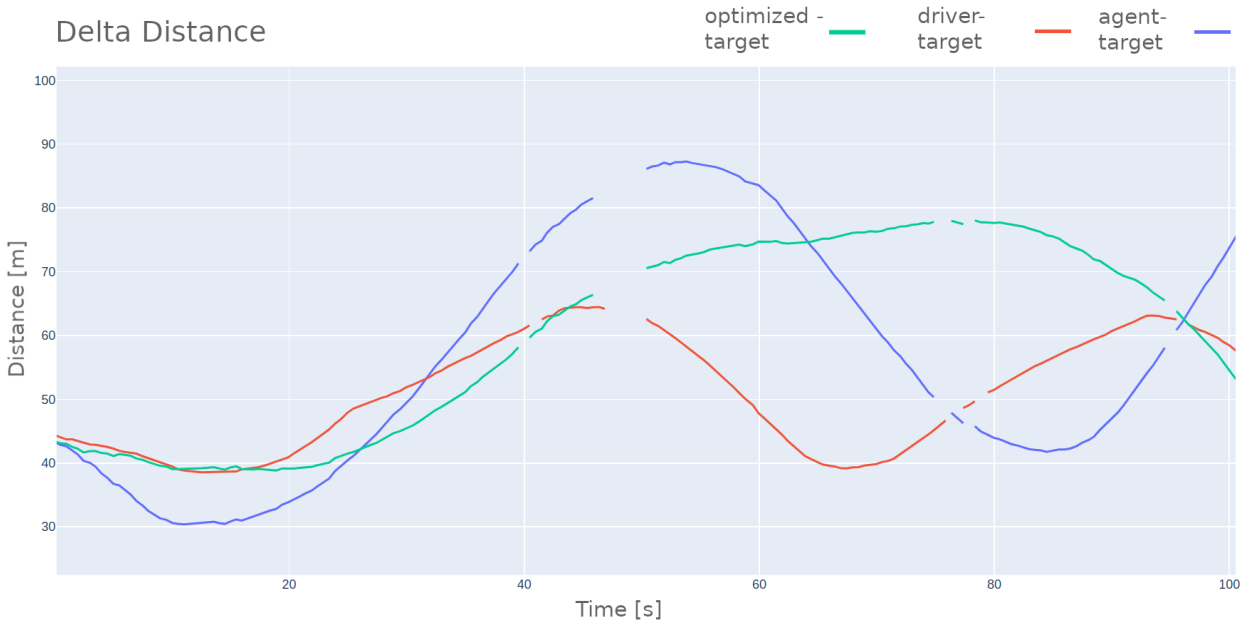

Figure 4.8: Comparison of distances to target during the evaluation of human driver (red), final policy (blue), and optimized solution on resimulated log (green).

4.6 Agent Explainability

As an addition to the evaluation process, the thesis proposed a new approach that could provide more transparency on the reasoning behind agent's decisions. The method used trajectory samples to determine the extent of an input feature's impact on the agent's action. The determination was calculated based on the Integrated Gradient method. It directly measured the magnitude of influencing of input to the output in ANN. The approach helped understand the policy behavior patterns and verify whether its decision-making process was coherent with human intuition.

The algorithm consists of several consecutive steps. Firstly the policy should be used for gathering statistically significant dataset which consists of input to and output from ANN. Then for all input features in all samples, the attributation was calculated and statistical analysis was performed. Firstly, it identified parameters with statistically significant attribution for the selected action and analysis of variance (ANOVA). Then it studied the linear and monotonic relationship between parameter allocation, input values, and given action with Pearson and Spearman's rank correlation coefficient Rho. Such analyses were then subject to interpretation. It was checked whether the detected strong correlation between the input value and the behavior matches the typical behavior pattern. The final step was to confirm that there was a strong correlation in places where it was expected.

The proposed method enhances the understanding of RL agents' decision-making, influenced by ANN-generated distributions. It identified key input features impacting decisions and their value correlations, while also pinpointing model or data errors, such as the vanishing gradient problem or incorrect data normalization. Addressing these issues improves system safety and predictability, particularly in automated vehicle motion planning, thereby increasing machine learning reliability for OEMs and consumers. The findings could be used for the refinement of ANN architecture or training pipeline, for instance, by adjusting the reward function to align with controller objectives or modifying ANN modules to consider neglected input features.

5. Summary and contributions

The thesis proposed an approach for developing a behavior policy for an Adaptive Cruise Control driving mode that could work in a natural environment. The policy was integrated into the motion planning architecture of automated vehicles, and its modularity supported the transparency requirements of vehicle OEMs.

The behavior module received input from the perception module and user preferences. Based on that the behavior policy denoted the target acceleration that the vehicle should achieve in 0.5 seconds. This acceleration value was used by the trajectory module to generate a continuous reference trajectory that was executed by the control module.

The presented approach combines both Offline and Online Reinforcement Learning algorithms to address their respective disadvantages. Online RL methods collect experiences and evaluations of actions in an online simulated environment to optimize the behavior policy, resulting in a policy that can handle a broad range of situations. However, this approach may not be feasible in real conditions due to the sim2real gap. In contrast, the Offline RL method uses data collected in natural conditions to optimize the behavior policy, resulting in a policy that can handle real-world situations, but covers a narrower range of cases.

The proposed method involved training an ANN-based policy on a real-world dataset using Offline RL algorithms with MARWIL and BC algorithms. The dataset was optimized with gradient optimization in order to alleviate the issue of suboptimality of experts' actions. The best-performing policy was then fine-tuned in a simulated environment with the PPO RL algorithm. Due to the fact that such fine-tuning can lead to catastrophic forgetting of knowledge inferred during offline training, the resimulation of real driving logs was incorporated as a part of the learning curriculum. The resimulation relied on replaying the scenarios included in the dataset but the agent might control its trajectory by selecting actions. In that setup, the policy could still optimize its behavior in a vast range of simulated scenarios as well as real situations.

The evaluation report based on real data showed that the agent trained according to the proposed method outperformed the human driver in some aspects. However, the human driver was faster and drove closer to targets than the trained agents. The proposed solution was promising, but extensive training with a carefully selected distribution of scenarios may be required to surpass the performance of human drivers in all scenarios.

Additionally to the evaluation phase, the study proposed a novel approach for explaining the actions of a Reinforcement Learning agent guided by ANN inference. The approach involved collecting observation-action pairs, calculating the significance rate of input features, and conducting a statistical analysis to identify meaningful correlations between actions and input components. The method allows for the interpretation of policy behavior patterns by matching them with human intuition.