Partial end-to-end reinforcement learning for robustness towards model-mismatch in autonomous racing

by

Andrew Murdoch

Thesis presented in partial fulfilment of the requirements for the degree of Master of Engineering (Electronic) in the Faculty of Engineering at Stellenbosch University

Supervisor:	Dr. J.C. Schoeman
Co-supervisor:	Dr. H.W. Jordaan

July 2023

Declaration

By submitting this thesis electronically, I declare that the entirety of the work contained therein is my own, original work, that I am the sole author thereof (save to the extent explicitly otherwise stated), that reproduction and publication thereof by Stellenbosch University will not infringe any third party rights and that I have not previously in its entirety or in part submitted it for obtaining any qualification.

Date: 2023/07/01

Abstract

The increasing popularity of self-driving cars has given rise to the emerging field of autonomous racing. In this domain, algorithms are tasked with processing sensor data to generate control commands (e.g., steering and throttle) that move a vehicle around a track safely and in the shortest possible time.

This study addresses the significant issue of practical *model-mismatch* in learning-based solutions, particularly in reinforcement learning (RL), for autonomous racing. Model-mismatch occurs when the vehicle dynamics model used for simulation does not accurately represent the real dynamics of the vehicle, leading to a decrease in algorithm performance. This is a common issue encountered when considering real-world deployments.

To address this challenge, we propose a partial end-to-end algorithm which decouples the planning and control tasks. Within this framework, a reinforcement learning (RL) agent generates a trajectory comprising a path and velocity, which is subsequently tracked using a pure pursuit steering controller and a proportional velocity controller, respectively. In contrast, many learning-based algorithms utilise an end-to-end approach, whereby a deep neural network directly maps from sensor data to control commands.

We extensively evaluate the partial end-to-end algorithm in a custom F1tenth simulation, under conditions where model-mismatches in vehicle mass, cornering stiffness coefficient, and road surface friction coefficient are present. In each of these scenarios, the performance of the partial end-to-end agents remained similar under both nominal and model-mismatch conditions, demonstrating an ability to reliably navigate complex tracks without crashing. Thus, by leveraging the robustness of a classical controller, our partial end-to-end driving algorithm exhibits better robustness towards model-mismatches than an end-to-end baseline algorithm.

Acknowledgements

This thesis appears in its current form due to the assistance and guidance of several people. I would therefore like to offer my sincere thanks to all of them.

I am thankful to God for granting me this opportunity to study. I praise Him for His strength, sustenance, and unwavering faithfulness.

I would like to express my sincere gratitude to my parents, Ross and Jeanne Murdoch. You have been a source of inspiration, and have fostered continual spiritual and emotional growth, as well as provided financial support throughout my studies.

To my supervisors, Dr. J.C. Schoeman and Dr. H.W. Jordaan, I would like to thank you for the guidance that you have provided, as well as the patience and kindness you have shown towards me during my degree. Thank you for the many meetings, comments, corrections, and encouragement.

Friends, thank you for your continual support, prayer, and encouragement throughout my studies.

And to my colleagues at the Electronic Systems Laboratory, thank you for making my studies a pleasant experience.

Contents

List of Figures

List of Algorithms

List of Tables

Nomenclature

Acronyms and abbreviations

LiDAR	light detection and ranging
IMU	inertial measurement unit
DNN	deep neural network
RL	reinforcement learning
MPC	model predictive control
DARPA	Defense Advanced Research Projects Agency
IL	imitation learning
CNN	convolutional neural network
BNN	bayesian neural networks
GTS	Gran Turismo Sport
TORCS	The Open Source Car Simulator
CAPS	conditioning for action policy smoothness
F1	Formula 1
ANN	artificial neural network
ReLU	rectified linear unit
FNN	feedforward neural network
Adam	adaptive moment estimation
MPD	Markov decision process
TD3	twin delay deep deterministic policy gradient
RC	remote controlled
DC	direct current
CoG	centre of gravity

Notation

x	Scalar
$\boldsymbol{x}$	Vector
$\boldsymbol{x}^\intercal$	Transpose of vector $\boldsymbol{x}$

Chapter 1

End-to-end autonomous racing

Having introduced our simulation environment, we formulate an end-to-end solution method in which a reinforcement learning (RL) agent directly predicts controller commands based on observation information. We employ this end-to-end agent as a baseline to compare our partial end-to-end algorithm against, as similar end-to-end methods are commonly used to solve the racing problem [1; 2; 3; 4; 5; 6; 7; 8; 9; 10; 11].

We begin this chapter by discussing the design of the end-to-end racing algorithm. Subsequently, we show how the TD3 RL algorithm is used to train an end-to-end agent, followed by a detailed exposition of evaluation procedures. We then experimentally determine the optimal values for each hyper-parameter for an agent racing on a relatively simple race track, before presenting agents capable of driving on more complex race tracks, The performance of end-to-end agents under conditions where vehicle modelling errors are present si also investigated, along with the effectiveness of domain randomisation as a technique to improve performance under these conditions.

1.1 End-to-end racing algorithm

Our end-to-end autonomous racing algorithm is composed of an RL agent and a velocity constraint. The agent maps an observation sampled from the simulator to desired longitudinal acceleration ($a_{\text{long},d}$) and steering angle (δ_d) control commands. The velocity constraint then modifies the acceleration commands to ensure that the vehicle remains within safe velocity bounds. The steering angle from the agent and acceleration from the velocity constraint component are passed to the simulator described in Chapter **??**. This end-to-end framework is depicted in Figure 1.1.

Importantly, the simulator and velocity constraint components are grouped together in the environment as per the definition of the MDP given in Section **??**, because this definition solely encompasses an agent and an environment. To ensure conformity between the end-to-end algorithm and the MDP definition, all of the racing algorithm components apart from the agent are considered as part of the environment, and executed in unison with the rest of the environment. Furthermore, due to the simulator's time step being chosen as 0.01 seconds, the environment components are sampled at a frequency of 100 Hz. The agent is sampled at a slower rate of f_{agent}.

The end-to-end agent, which comprises a neural network, is shown in Figure 1.2. To ensure uniformity across all observation vector elements, each element in the input vector is normalized to the range $(0, 1)$. The neural network's design consists of three fully connected layers, with m_1, m_2, and 2 neurons in the input, hidden, and output layers,

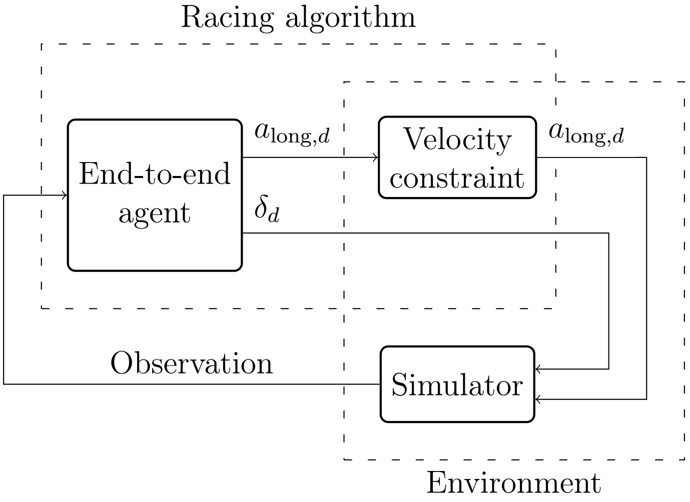


Figure 1.1: The end-to-end racing algorithm, which is comprised of an RL agent which outputs control actions, and a velocity constraint. The velocity constraint and simulator are both considered part of the environment.

respectively. The first two layers are ReLU-activated, while the output layer is activated by a hyperbolic tangent function to normalize the neural network output to the range $(-1, 1)$. While the number of neurons in the first two layers are determined empirically, the two neurons in the output layer correspond to the steering and acceleration actions. Scaling factors are applied to their outputs so that the selected steering and acceleration actions fall within the range $(\underline{\delta}, \overline{\delta})$ and $(\underline{a}, \overline{a})$ from Table **??**, respectively.

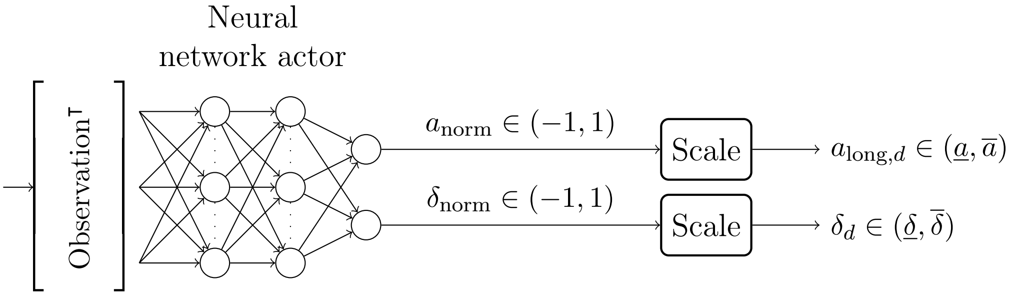

Neural network actor
Observation$^\intercal$
$a_{\text{norm}} \in (-1, 1)$
$\delta_{\text{norm}} \in (-1, 1)$
Scale
Scale
$a_{\text{long},d} \in (\underline{a}, \overline{a})$
$\delta_d \in (\underline{\delta}, \overline{\delta})$

Figure 1.2: The end-to-end agent. The outputs of the neural network are scaled to the ranges of a_{long} and δ in Table **??**.

While the steering angle is passed directly to the simulator, the longitudinal action is first modified by the velocity constraint component to ensure that the velocity of the vehicle remains within safe bounds,

$$a_{\text{long},d} \leftarrow \begin{cases} 0 & \text{for } v \geq v_{\max}, \\ 0 & \text{for } v \leq v_{\min}, \\ a_{\text{long},d} & \text{otherwise}, \end{cases} \tag{1.1}$$

before being passed to the simulator. $v_{\max}$ and $v_{\min}$ are the imposed maximum and minimum allowable velocities.

1.2 Applying TD3 to end-to-end autonomous racing

We applied the TD3 RL algorithm from Section **??** to train the end-to-end agent. Several adaptations to the original TD3 algorithm were made to ensure its compatibility with the end-to-end racing algorithm. The adapted TD3 is shown in Algorithm 1.

Algorithm 1: Twin delay deep deterministic policy gradient

```
Input: Table 1.1 parameters
Output: Trained actor DNN $\pi_{\boldsymbol{\phi}}$

Initialise critic networks $Q_{\boldsymbol{\theta}_1}, Q_{\boldsymbol{\theta}_2}$ and actor network $\pi_{\boldsymbol{\phi}}$ with
parameters $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2$ and $\boldsymbol{\phi}$
Initialise target networks $Q_{\boldsymbol{\theta}'_1}$, $Q_{\boldsymbol{\theta}'_2}$ and $\pi_{\boldsymbol{\phi}'}$ with weights
$\boldsymbol{\theta}'_1 \leftarrow \boldsymbol{\theta}_1$, $\boldsymbol{\theta}'_2 \leftarrow \boldsymbol{\theta}_2$, and $\boldsymbol{\phi}' \leftarrow \boldsymbol{\phi}$
Initialise experience replay buffer $\mathcal{B}$

while MDP time steps $< M$ do
    Reset simulator (start new episode)
    for $t$=0, $T$ do
        Sample action with exploration noise from end-to-end agent,
        $a_t = [a_{\text{long},d}, \delta_d] \leftarrow \text{scale}(\pi_{\boldsymbol{\phi}}(o_t) + \epsilon),\ \ \epsilon \sim \mathcal{N}(0, \sigma_{\text{action}})$
        for $n$=0, $N$ do
            Modify $a_{\text{long},d}$ according to Equation 1.1 to limit velocity
            Simulator executes action $a_t$
            Observe environment step reward $r_{t,n}$
            Update MDP one step reward: $r_t = r_t + r_{t,n}$
            Sample observation with noise,
            $o_t \leftarrow o_t + \epsilon,\ \ \epsilon \sim \mathcal{N}(0, \sigma_{\text{observation}})$
        end
        Store transition tuple $(o_t, a_t, r_t, o_{t+1})$ in $\mathcal{B}$

        Sample mini-batch of $B$ transitions $(o_i, a_i, r_i, o_{i+1})$ from $\mathcal{B}$
        Select target actions:
        $\tilde{a} \leftarrow \pi_{\boldsymbol{\phi}'}(o_{t+1}) + \epsilon, \quad \epsilon \sim \text{clip}(\mathcal{N}(0, \tilde{\sigma}), -c, c)$
        Set TD target: $y_i \leftarrow r_i + \gamma \min_{q=1,2} Q_{\boldsymbol{\theta}'_q}(o_{i+1}, \tilde{a})$
        Update critics by minimising the loss: $J_{\boldsymbol{\theta}_q} \leftarrow \frac{1}{N}\sum_i^N (y_i - Q_{\boldsymbol{\theta}_q}(o_i, a_i))^2$

        if $t \mod d$ then
            Update $\boldsymbol{\phi}$ by the deterministic policy gradient:
            $\nabla_{\boldsymbol{\phi}} J(\boldsymbol{\phi}) = \frac{1}{N}\sum_i^N \nabla_a Q_{\boldsymbol{\theta}_1}(o_i, a)|_{a=\pi_{\boldsymbol{\phi}}(o_i)} \nabla_{\boldsymbol{\phi}} \pi_{\boldsymbol{\phi}}(o_i)$
            Update target networks:
            $\boldsymbol{\theta}'_q \leftarrow \tau\boldsymbol{\theta}_q + (1-\tau)\boldsymbol{\theta}'_q$
            $\boldsymbol{\phi}' \leftarrow \tau\boldsymbol{\phi} + (1-\tau)\boldsymbol{\theta}'$
        end
    end
end
```

Note that end-to-end agents in the context of racing receive only partial information about the state of the environment, because the pose and LiDAR scan do not fully capture the environment state. Hence, the racing environment is only partially observable. As such, the input to a racing agent is therefore an observation, denoted as o, rather than the complete environment state s. This notation conforms to the notation used for the output of the simulator in Chapter **??**. However, we use the notation for time steps in Chapter **??**, denoting a time step as t, rather than k.

In line 1 of Algorithm 1, the actor ($\pi_{\boldsymbol{\phi}}$) and critics ($Q_{\boldsymbol{\theta}_1}$ and $Q_{\boldsymbol{\theta}_2}$) are initialised. The end-to-end agent shown in Figure 1.2 is the actor $\pi_{\boldsymbol{\phi}}$. The design of the critics $Q_{\boldsymbol{\theta}_1}$ and $Q_{\boldsymbol{\theta}_2}$ are now described. For simplicity, our critics have an identical structure which is analogous to that of the actor. We therefore describe the details of only one critic, which is depicted in Figure 1.3. The critic DNN receives a vector input comprised of observation and control actions normalised to the range $(-1, 1)$. It comprises three fully connected layers: an input layer with m_1 neurons, a hidden layer with m_2 neurons, and an output layer with a single neuron. The output of this neuron is the action-value, denoted $Q_{\boldsymbol{\theta}}(o, a)$. The first two layers utilize the Rectified Linear Unit (ReLU) activation function, while the final layer uses a linear activation function. Additionally, the target networks are initialised identically to their counterparts.

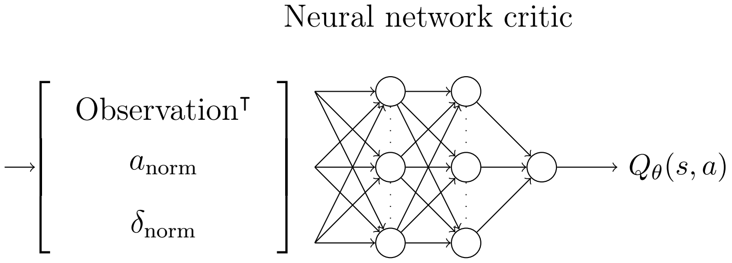


Figure 1.3: The critic DNN. Its input is a vector containing a normalised observation and action pair, and its output is an action-value.

After initializing the replay buffer in line 3, the TD3 algorithm enters a while loop which executes a number of episodes (lines 5-22). However, rather than limiting the number of episodes, we set a limit on the number of MDP time steps, denoted as M, as it is a more accurate indicator of the training time and the number of actor and critic updates. Each episode starts by resetting the simulator, and ends when the simulator indicates that the vehicle has crashed or finished.

In line 7 of Algorithm 1, an action is sampled from the end-to-end agent by forward passing the observation through the actor DNN. Gaussian noise with a zero mean and a standard deviation of σ_{action} is added to the normalised output (a_{norm} and δ_{norm}), which is then scaled to generate a longitudinal acceleration and a steering angle.

The action sampled from the agent in line 7 is executed by repeatedly sampling the environment (lines 8-13). Our implementation for sampling the environment differs from the standard implementation of TD3 given in Algorithm **??**. This is because the sample rate of the components inside the environment is higher than the agent, whereas the definition of the MDP given in Section **??** requires that the agent and environment is sampled at the same rate. As such, the environment components will be sampled multiple times in-between agent sampling periods. We therefore define an MDP step as N environment samples, where

$$N = \frac{100}{f_{\text{agent}}}. \tag{1.2}$$

In Equation 1.2, N is a whole number. The environment is sampled by applying the velocity constraint from Equation 1.1 to limit the agent's selected longitudinal action, and then executing one simulator step. This is followed by sampling the reward signal from the simulator in line 11.

The reward signal is designed to closely approximate the objective of minimizing lap time for high reward discount rates. Specifically, we reward the agent for the distance it travels along the centerline between the current and previous time step, while penalizing it a small amount on every time step, as described by Fuchs et al. [2]. In addition, we impose a large penalty if the agent collides with the track boundary. As a result, we obtain a piece-wise reward signal expressed as

$$r(s_t, a_t) = \begin{cases} r_{\text{collision}} & \text{if collision occurred} \\ r_{\text{dist}}(D_t - D_{t-1}) + r_{\text{time}} & \text{otherwise.} \end{cases} \tag{1.3}$$

Here, $r_{\text{collision}}$, r_{dist}, and r_t represent the penalty for collisions, the reward for distance traveled, and the penalty for each time step, respectively. Notably, this reward signal is similar to those used in numerous prior works [1; 3; 4].

The reward signal is accumulated over the sequence of N steps during which the environment is sampled in line 12. In line 15, the transition tuple is stored, which consists of the observation before sampling the environment N times, as well as the observation and accumulated reward after sampling the environment N times.

The remaining steps of Algorithm 1 are identical to the standard implementation of TD3, as described in Algorithm **??**. Specifically, we first sample a mini-batch of B transitions, from the replay buffer. Next, we employ the target actor network to select actions for each observed sample, which in turn are used to update the critics. To ensure the stability of the learning process, we update the actor and the target networks every d steps. Additionally, the target networks are updated via a soft update which is controlled by the target update rate parameter τ.

After the training procedure is completed, we utilise Algorithm 2 to evaluate the trained agents. Under evaluation conditions, learning was halted and the weights of the DNNs were not updated. Furthermore, no exploration noise was added to the agents selected actions. However, Gaussian noise was added to the observation vector to mimic practical sensor data in simulation. This added Gaussian noise had standard deviations of 0.025 m for x and y coordinates, 0.05 rads for heading, 0.1 m/s for velocity, and 0.01 m for LiDAR scan. Each agent completed 100 laps under these evaluation conditions.

1.3 Empirical design and hyper-parameter values

The previous section introduced the design of the end-to-end algorithm and the TD3 algorithm with symbolic hyper-parameter values. The optimal values for these hyper-parameters cannot be derived, and require experimentation to be determined empirically. Furthermore, hyper-parameters are sensitive to the track. The following five sections of this chapter detail the experiments that were undertaken to determine a locally optimal set of hyper-parameters for agents racing on a relatively simply track named Porto. The selected hyper-parameters are listed in Table 1.1. Additionally, the average learning curve for 10 agents racing on this track using this set of hyper-parameter is shown in Figure 1.4.

Algorithm 2: Evaluating the end-to-end algorithm without exploration noise, and with observation noise.

```
Input: Trained actor DNN $\pi_\phi$
Output: Lap times, collisions over 100 laps
Initialise actor DNN $\pi_\phi$ with weights $\phi$ from training
for episode = 1, 100 do
    for $t$=0, $T$ do
        Select control action $a_t = [a_{\text{long,d}}, \delta_d] \sim \text{scale}(\pi_\phi(o_t + \epsilon))$
        for $n$=0, $N$ do
            Modify $a_{\text{long},d}$ according to Equation 1.1 to limit velocity
            Simulator executes action $a_t$
        end
    end
end
```

Hyper-parameter	Symbol	Value
Algorithm	$1.5 \cdot 10^5$	
Maximum time steps	M	
Target update rate	τ	$5 \cdot 10^{-3}$
Replay buffer size	$\mathcal{B}$	$5 \cdot 10^5$
Replay batch size	B	400
Exploration noise standard deviation	σ_{action}	0.1
Reward discount rate	γ	0.99
Agent samples between network updates	d	2
Agent sample rate	f_{agent}	5 Hz
Target action noise standard deviation	$\tilde{\sigma}$	0.2
Target action noise clip	c	0.5
Reward signal		
Distance reward	r_{dist}	0.25
Time step penalty	r_{time}	0.01
Collision penalty	$r_{\text{collision}}$	-10
Observation		
Number of LiDAR beams	L	20
Neural network		
Learning rate	α	10^{-3}
Neurons in input layer	m_1	400
Neurons in hidden layer	m_2	300
Velocity constraints		
Minimum velocity	v_{min}	3 m/s
Maximum velocity	v_{max}	5 m/s

Table 1.1: Selected values of hyper-parameters for the end-to-end racing algorithm on the Porto track.

To select each hyper-parameter value listed in this Table, we repeatedly trained agents using Algorithm 1 varied values of the hyper-parameter under consideration while keeping all other hyper-parameters fixed at the values listed in Table 1.1. When evluating agents, we are particularly interested in the rate at which they successfully complete laps, as well as their lap time during and after training. Furthermore, to ensure consistency in the results, we trained and evaluated multiple agents for each set of hyper-parameters. Specifically, we chose to train three agents for each hyper-parameter set due to time constraints.

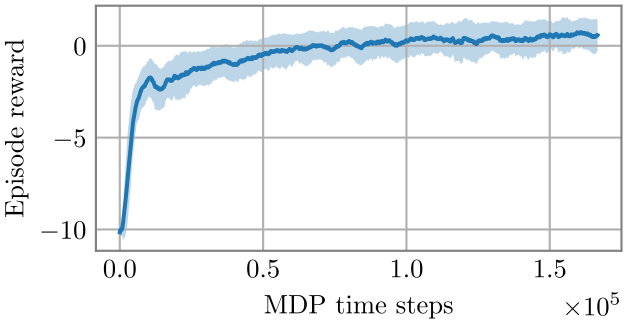


Figure 1.4: Average learning curve for 10 end-to-end agents trained on the simple Porto track.

1.4 TD3 hyper-parameters

We performed experiments to determine values for the TD3 algorithm hyper-parameters that result in good performance for the Porto track. These hyper-parameters were the number of MDP time steps M, agent sample rate f_{agent}, target update rate τ, replay buffer size $\mathcal{B}$, replay batch size B, standard deviation of exploration noise σ_{action}, reward discount rate γ, and agent samples in-between DNN updates d.

We begin this section by discussing the experiment by which the appropriate number of time steps to train the agent for was determined. The objective for determining the length of the training time was to ensure that the agent would demonstrate satisfactory performance under evaluation conditions by racing quickly and consistently avoiding crashes. Ending training too soon may result in poor agent performance, while training for too many time steps long may result in unnecessarily prolonged training times. To achieve this, a set of three agents with the hyper-parameters listed in Table 1.1 were trained. These agents were evaluated using Algorithm 2 at 100 episode intervals during training. The percentage of failed laps and lap time under evaluation conditions are depicted as a function of training time in Figure 1.5.

We observe that during the early stages of training, both lap time and success rate improved rapidly. However, it takes a considerable amount of time before the agent consistently completes all of its laps under evaluation conditions. Considering these results, we determined that $1.5 \cdot 10^5$ MDP time steps is an appropriate length for training an end-to-end agent.

The optimal value for the rate at which actions are sampled from the agent, denoted as f_{agent}, was then determined. We investigated agent sample rates in the range of 3 Hz to 50 Hz. For each sample rate investigated, three agents were trained with the remaining hyper-parameters set equal to those listed in Table 1.1. Figure 1.6 shows the average failed laps and lap time of agents racing under evaluation conditions as a function of f_{agent}.

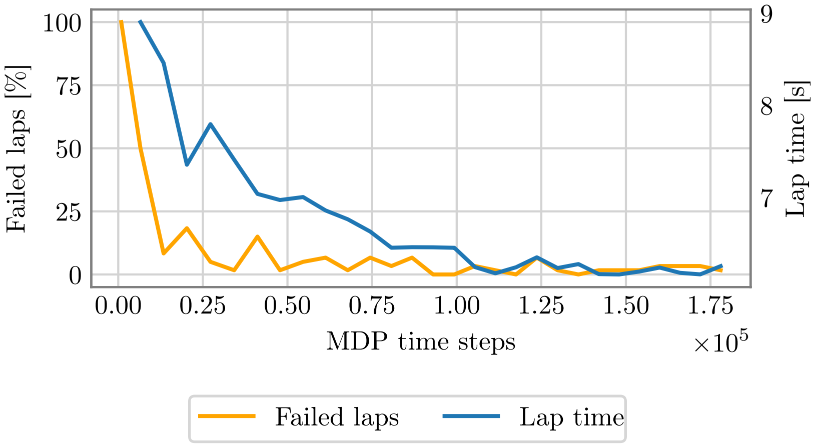


Figure 1.5: Percentage failed laps (left vertical axis) and lap time (right vertical axis) of three agents tested under evaluation conditions at 100 episode intervals during training.

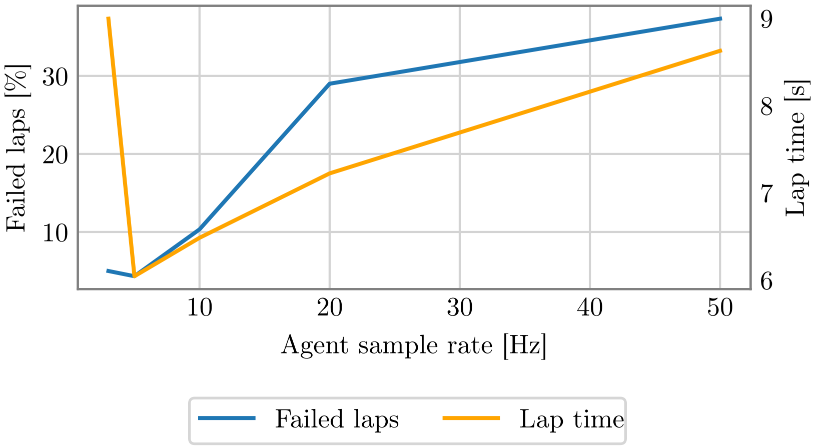


Figure 1.6: Training time (left vertical axis) and percentage failed laps (right vertical axis) of three trained end-to-end agents racing under evalution conditions on the Porto track, with sampling rates ranging from 3 Hz to 50 Hz.

From this figure, we observe that agents trained with sampling rates higher than 5 Hz tend to crash, as well race slowly. This outcome may be attributed to the fact that when a higher sampling rate is used, the agent needs to learn longer action sequences to complete a lap, leading to a more complex learning problem. We set the value of f_{agent} to be 5 Hz as it resulted in the minimum number of failed laps during evaluation, despite taking a relatively long 26.2 minutes to train each agent. It is notable that 5 Hz is a relatively slow sampling rate compared to classical controllers. For instance, Li et al. [12] develop path tracking controllers with sampling rates up to 100 Hz.

The optimal value for the batch size B was determined by training and evaluating agents with batch sizes of 50, 100, 150, 200, 400, 600 and 1000 samples. Three agents were trained for every batch size, while holding the remaining hyper-parameters constant at the values listed in Table 1.1. The average lap time and percentage of failed laps of agents under evaluation conditions are shown as a function of batch size in Figure 1.7. From this figure, we observe that lap time and failed laps under evaluation conditions are minimised when the batch size is set to 400. Based on these results, we selected a batch size of 400 samples for our agents.

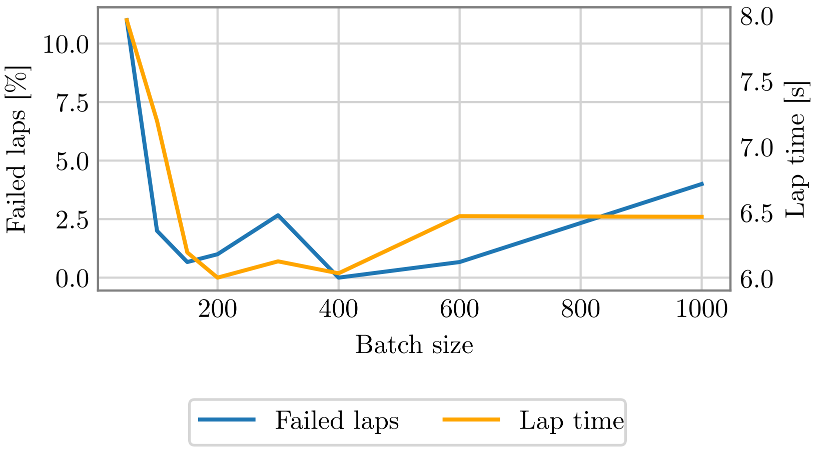


Figure 1.7: Training time and lap time under evaluation conditions of end-to-end agents with batch sizes from 50 to 1000. The percentage failed laps is mapped onto the left vertical axis, while the lap time is mapped onto the right vertical axis.

An experimental analysis was then conducted to select the reward discount rate, denoted γ. To determine the value for γ, we assessed the performance of agents with reward discount rates of 0.95, 0.98, 0.99 and 1 during training. For each of these reward discount rate values, three agents were trained with their remaining hyper-parameters set equal to those listed in Table 1.1. The percentage failed laps and lap times during training, was well as the learning curves for these agents are shown in Figure 1.8.

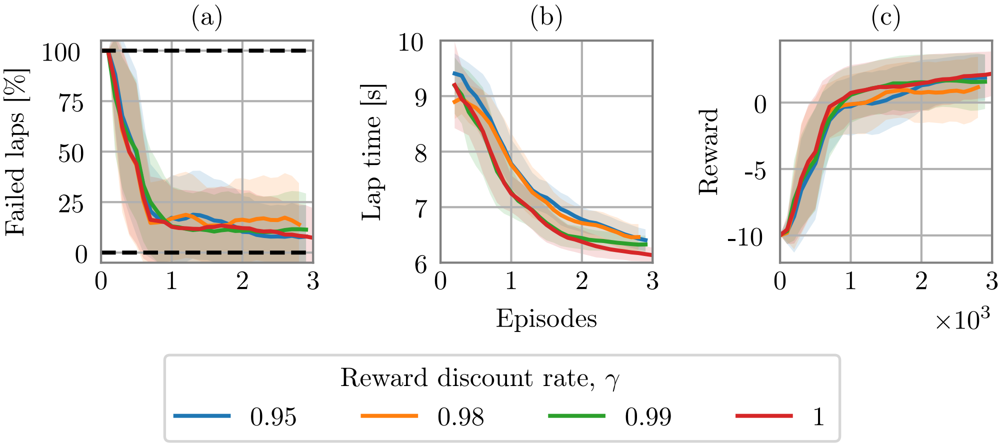


Figure 1.8: (a) The percentage failed laps and (b) lap time of completed laps during training, as well as (c) the learning curves of three end-to-end agents with reward discount rates ranging from 0.9 to 1.

By only assessing the performance during training, these agents appear to be perform similarly. However, the TD3 algorithm has no mechanism for decreasing the exploration noise added to every action with training time. Figure 1.8 is therefore an indicator of the performance of each agent with added exploration noise. As such, we also considered the performance of each agent under evaluation conditions where no exploration noise is present. The percentage failed laps and lap times for agents trained with each learning rate is shown in Table 1.2. The table show that a discount rate of 0.99 yields agents that successfully complete all of their laps. Based on this finding, a discount rate of 0.99 was selected for our agents.

Reward discount rate (γ)	**Failed laps [%]**	**Average lap time [s]**	**Standard deviation of lap time [s]**
0.95	2.33	6.12	0.28
0.98	0.33	6.51	0.28
0.99	0.00	6.07	0.20
1	0.33	5.94	0.11

Table 1.2: Percentage failed laps and lap times under evaluation conditions for agents trained with reward discount rates ranging from 0.9 to 1.

Values for the hyper-parameters τ, σ_{action}, and d were determined by repeating the tuning procedure used for γ. That is, one hyper-parameter was varied while holding the others constant. For each hyper-parameter set, three agents were trained, and the selected was based on the agents' average performance during training and evaluation. For conciseness, the experimental results for these hyper-parameters are presented in Appendix A. Values of $5 \cdot 10^{-3}$ for τ, 0.1 for σ_{action}, and 2 for d yielded agents with the best performance.

After determining locally optimal hyper-parameters for TD3, we compared the performance of our implementation of the TD3 algorithm to a standard implementation of the popular Deep Deterministic Policy Gradient (DDPG) algorithm [3; 13; 7]. The percentage failed laps, lap time and learning curves of agents trained using both algorithms are depicted in Figure 1.9. The results reveal that TD3 outperforms DDPG by a substantial margin in terms of both crashes and lap time. Moreover, we have observed that the training stability of TD3 is superior to that of DDPG, as evidenced by the smoother learning curve of TD3 in contrast to the more erratic curve of DDPG.

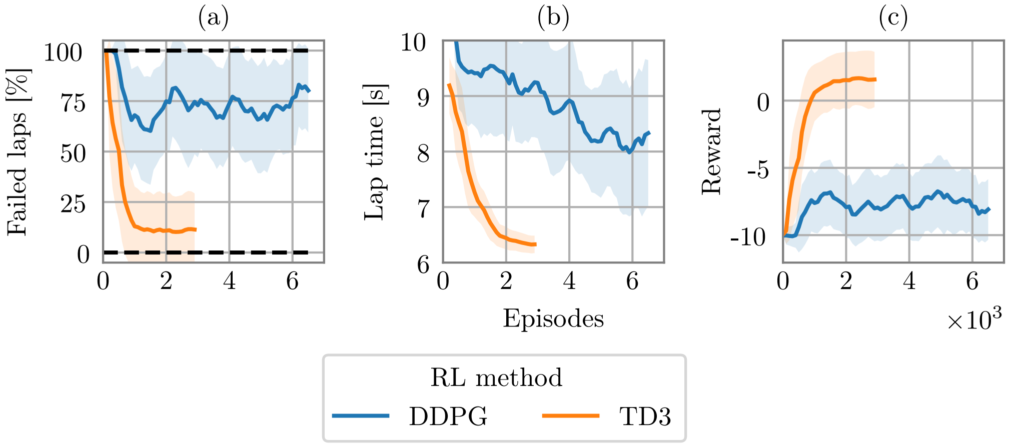


Figure 1.9: (a) The percentage failed laps and (b) lap time of completed laps during training, as well as (c) the learning curves of three end-to-end agents that were trained using TD3 and DDPG.

1.5 Reward signal

Having experimentally determined locally optimal values for the TD3 hyper-parameters, we investigated the reward signal. Our objectives were to choose reward signal parameter values that yielded agents that (a) race safety while (b) minimising lap time. This was a challenging task, considering that these two objectives are in conflict with each other. Further complicating the task is that the lap time alone is too sparse a signal to allow the agent to learn effectively [4; 5]. The reward signal from Equation 1.3 was therefore designed to approximate a signal that minimises lap time, while also providing continuous rewards to the agent. Specifically, the reward signal described in Equation 1.3 rewards the agent for the distance it travelled along the centerline between the current and previous time step, and penalises the agent a small amount on every time step [2]. Additionally, the agent receives a large penalty for colliding with the track boundary.

To motivate the use of a time step penalty r_{time} and to prove that our reward signal approximates a minimisation of lap time, we examine the total reward accumulated over a successful episode,

$$R_{\text{total}} = \sum_{t=1}^{T} \left(r_{\text{dist}}(D_t - D_{t-1}) + r_{\text{time}} \right), \tag{1.4}$$

which is the quantity that the agent learns to maximize when no reward discounting is assumed. In this equation, the subscript t indicates a time step, T is the final time step of the episode, and D_t is the distance travelled along the centerline at time t. Expanding the summation yields

$$\begin{aligned} R_{\text{total}} &= r_{\text{dist}} \left((D_1 - 0) + \ldots + (D_T - D_{T-1}) \right) + \sum_{t=1}^{T} r_{\text{time}} \\ &= r_{\text{dist}} D_T + T r_{\text{time}}. \end{aligned} \tag{1.5}$$

To simplify the expression for total reward, r_{time} was set equal to $-\Delta t$, or -0.01. Additionally, T was substituted as

$$T = \frac{\text{lap time}}{\Delta t}. \tag{1.6}$$

By substituting Equation 1.6 into Equation 1.5, we get R_{total} as

$$R_{\text{total}} = r_{\text{dist}} D_T - \text{lap time}. \tag{1.7}$$

From this equation, we can see that the agent must minimise lap time to maximise the total reward because D_t is constant. Therefore, the reward signal from Equation 1.3 approximates a signal that minimises lap time for sufficiently large reward discount factors. Furthermore, if no time step penalty is applied, then every successful lap yields the same reward regardless of lap time.

To experimintally confirm the result from Equation 1.7, we trained and evaluated three agents with r_{time} set to -0.01, then repeated the training procedure for three agents without the time step penalty. For each of these agents, the remaining reward signal components and hyper-parameters were set equal to the values listed in Table 1.1. Setting r_{time} to -0.01 improves the average evaluation lap time of agents from 9.26 seconds to 6.07 seconds, compared to agents that did not receive the penalty. Furthermore, we tuned the other reward signal terms are relative to the r_{time} value of -0.01.

We now present the tuning procedure for the distance reward r_{dist}, as well as the collision penalty $r_{\text{collision}}$. We initially determined a plausible range of r_{dist} values to train our agents with. Intuitively, a lower bound for r_{dist} exists that results in a policy that completes laps. If r_{dist} is set beneath this lower bound, the agent can only accumulate negative reward by continuing to race, and the optimal action is to crash immediately. We estimated this lower bound by considering that the agent should be able to achieve positive reward at every time step, such that

$$r_{\text{dist}}(D_t - D_{t-1}) + r_{\text{time}} > 0. \tag{1.8}$$

Solving the inequality in Equation 1.8 for r_{dist} gives

$$r_{\text{dist}} > \frac{-r_{\text{time}}}{(D_t - D_{t-1})}, \tag{1.9}$$

where D_t and D_{t-1} are unknown. To obtain the smallest value for r_{dist}, we estimate the largest value possible for $(D_t - D_{t-1})$ by considering a case whereby the vehicle travels at maximum speed parallel to the centerline, such that

$$(D_t - D_{t-1}) = v_{\text{max}}\Delta t. \tag{1.10}$$

After substituting the expression from Equation 1.10 into Equation 1.9 and setting r_{time} equal to $-\Delta t$, the minimum value for r_{distance} is found to be

$$r_{\text{dist}} > \frac{1}{v_{\text{max}}}. \tag{1.11}$$

Substituting the value for v_{max} as 5 m/s into Equation 1.11 yields an estimated minimum r_{dist} of 0.2.

Using this value as a guide for the region in which to search for r_{dist}, we trained agents with r_{dist} values of 0.1, 0.25, 0.3 and 1. For each r_{dist} value, three agents were trained with their remaining hyper-parameters equal to those listed in Table 1.1. The percentage failed laps and average lap time of completed laps during training for these agents are shown in Figure 1.10. Unsurprisingly, the agent with r_{dist} set to 0.1 (i.e., less than the estimated minimum) learns that terminating the episode immediately is the optimal behaviour, as its failure rate remains at 100 percent.

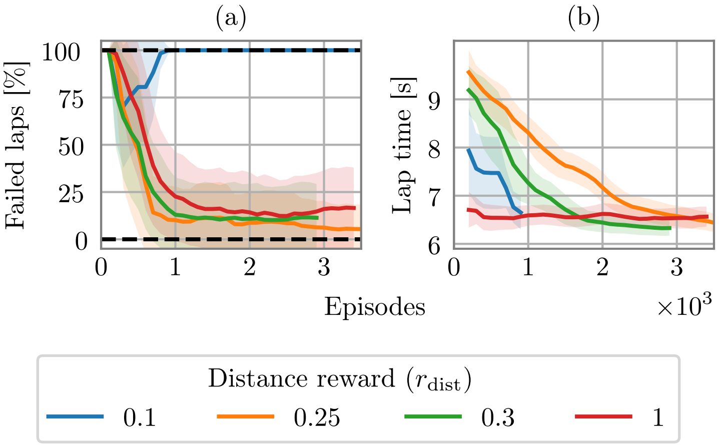


Figure 1.10: (a) The percentage failed laps and (b) lap times of completed laps during training of end-to-end agents with r_{dist} values ranging from 0.1 to 1.

Figure 1.10 also reveals that larger values of r_{dist} result in worse performance in terms of failed laps and lap time. When r_{dist} is set to a larger value, the time step penalty becomes less significant. As such, the agent is less incentivised to minimise lap time. Conversely, when r_{dist} is set close to the estimated minimum value, the time step penalty becomes significant, and the agent must optimise lap time to receive positive rewards. The value for r_{dist} was chosen as 0.25, as agents that were trained with this value had the lowest crash rate while also achieving competitive lap times.

After setting the value for r_{dist}, the penalty imposed on the agent when it collides with the track boundary was fine-tuned. Initially, we investigated whether the agent could acquire the racing skills without facing any penalties for collisions. However, agents trained with such a reward signal crashed on 4% of their laps during evaluation. Consequently, we conducted further experiments by considering negative $r_{\text{collision}}$ values.

To identify a suitable range within which we could conduct experimental searches for an optimal value, we operated on the premise that $r_{\text{collision}}$ should be substantial compared to the positive reward an agent can receive in an episode. As shown in Figure 1.4, agents attain an average reward value of 2 in episodes where crashes do not occur. Consequently, we trained agents with collision penalties ranging from -2 to -10. As before, three agents with hyper-parameter set equal to those listed in Table 1.1 were trained for each $r_{\text{collision}}$ value. The percentage failed laps and average lap times of agents trained with these values for $r_{\text{collision}}$ are presented in Table 1.3. We selected $r_{\text{collision}}$ as -10, as it is the only penalty that results in no failed laps.

$\mathbf{r_{collision}}$	**Failed laps [%]**	**Average lap time [s]**	**Standard deviation of lap time [s]**
0	4.00	5.69	0.16
-2	1.33	5.63	0.17
-4	1.00	5.69	0.17
-8	1.33	6.11	0.47
-10	0.00	6.07	0.20

Table 1.3: Percentage failed laps and lap times under evaluation conditions for agents trained with $r_{\text{collision}}$ values from ranging from 0 to -10.

Interestingly, the effect of increasing the collision penalty can be seen in the path taken by agent. Figure 1.11 shows the paths taken by agents with $r_{\text{collision}}$ set to -4 and -10. The agent with the lower collision penalty races close to the edge of the track, while the agent that is penalised more heavily takes a much more conservative path by staying clear of the track boundaries, instead preferring to drive near the centerline of the track.

1.6 Observation space

We also investigated the optimal the observation space vector. Initially, the observation vector components that yield agents with the best performance were determined , by training and evaluating agents that received (a) only the pose, (b) only a LiDAR scan and (c) a combination of vehicle pose and LiDAR scan. For each of these observation space combinations, three agents with hyper-parameters listed in Table 1.1 were trained. The performance of these agents during training, in terms of percentage failed laps, average lap time and average reward, is shown in Figure 1.12.

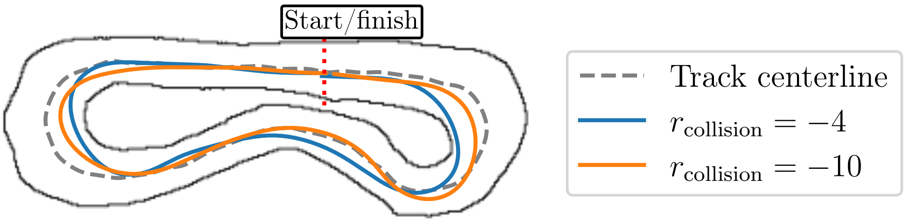


Figure 1.11: The paths taken by agents trained with $r_{\text{collision}}$ values of -4 and -10.

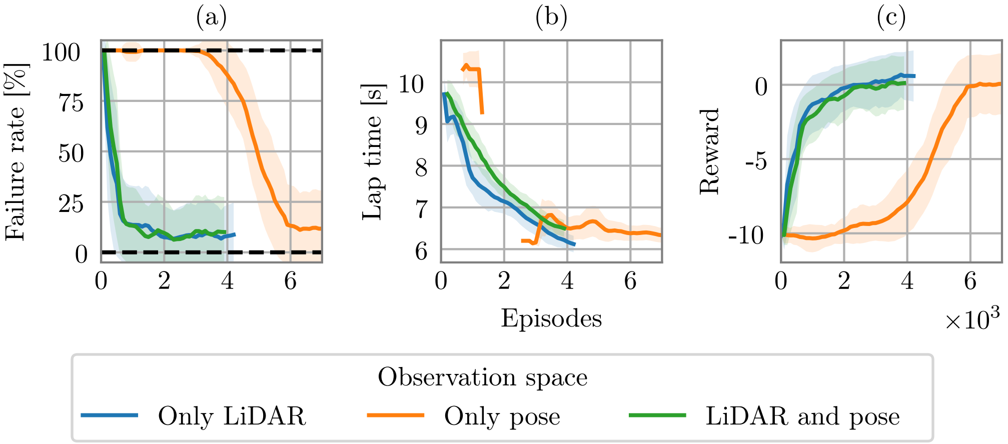
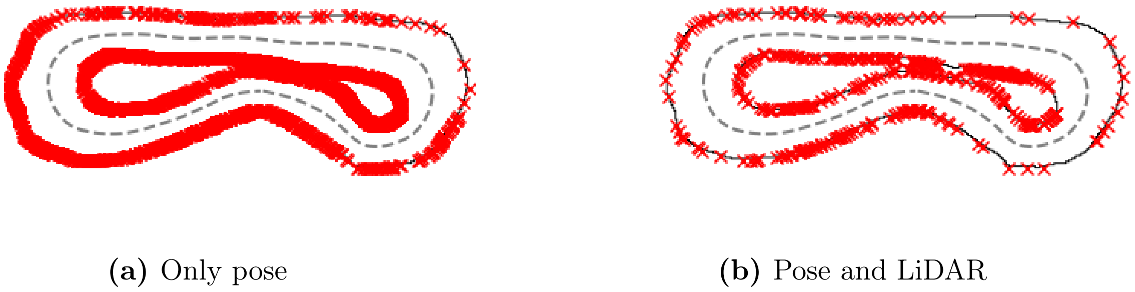


Figure 1.12: (a) the percentage failed laps and (b) lap time of completed laps during training, as well as (c) the learning curves showing episode reward for end-to-end agents with different observation spaces.

From this figure, agents utilising each of the observation spaces converge to a similar values for all three evaluation metrics. However, agents that receive a LiDAR scan train significantly faster than agents without a LiDAR scan in their observation. Specifically, the LiDAR scan allows the agent to learn to avoid track boundaries without needing to sample collision experiences at every point along the track boundary. This is clearly demonstrated in Figure 1.13, which shows all of the locations where an agent observing only the pose, and an agent observing both pose and LiDAR scan crashed during training. Agents without LiDAR scans crashed 5183 times, whereas agents observing LiDAR scans crashed only 464 times during the same training period.

After determining that including the LiDAR scan in the observation improves training performance, we assessed agents utilising each observation space under evaluation conditions. The agent that utilised both the LiDAR scan and pose in the observation did not crash during evaluation, whereas agents with either only a LiDAR scan or pose failed to complete laps 0.67% and 6.00% of the time, respectively. Based on these results, we chose to include both a LiDAR scan and the vehicle pose into the observation.

Another parameter to consider when choosing the observation space is the number



(a) Only pose

(b) Pose and LiDAR

Figure 1.13: Crash locations of agents with (a) only the pose and (b) both the pose and LiDAR scan during training.

of LiDAR beams. To determine the number of LiDAR beam that results in optimal performance, agents with LiDAR scans consisting of 5, 10, 20 and 50 were trained and tested. These beams are equally spaced, and have a field of view of 180°. As before, three agents with hyper-parameters from Table 1.1 were trained for every value of L LiDAR beams. Figure 1.14 displays the percentage failed laps and lap times during training, as well as the learning curves of these agents. The results indicate that increasing the number of LiDAR beams above 20 does not significantly impact the performance of agents in terms of any of the measured criteria. We therefore chose to incorporate 20 LiDAR into the observation space.

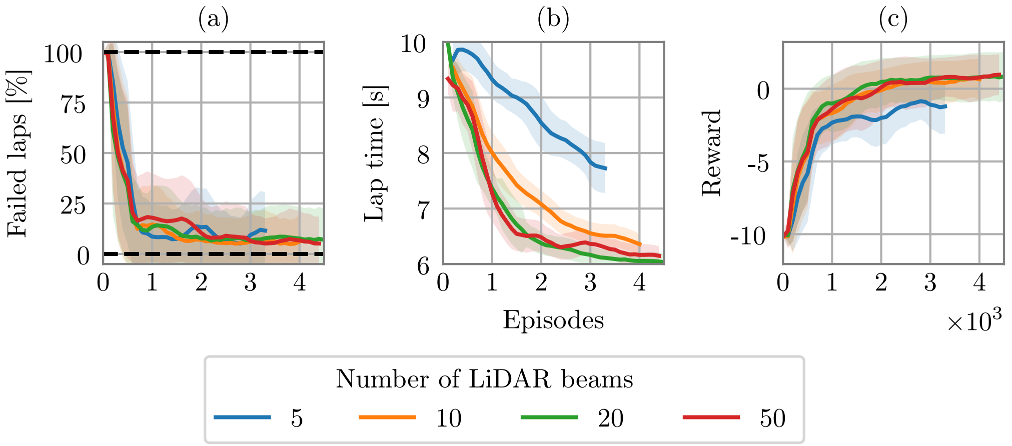


Figure 1.14: (a) the percentage failed laps and (b) lap time of completed laps during training, as well as the (c) learning curves showing episode reward for end-to-end agents with different numbers of LiDAR beams during training.

The simulation environment described in Section **??** allows for the addition of noise to the observation vector. The tests conducted thus far have not included noise in the agents' observations. However, noise is added to the observation vector to increase the realism of the simulation when racing under evaluation conditions,. It is therefore important to determine whether adding noise to the observation elements during training benefits the performance of the agent under evaluation conditions. Specifically, we trained three agents without any noise in the observation vector, and another three with added Gaussian noise which had standard deviations of 0.025 m for x and y coordinates, 0.05 rads for heading, 0.1 m/s for velocity, and 0.01 m for LiDAR scan.

The agents trained with noise achieved an average lap time of 6.77 seconds while completing 98.67% of the laps under evaluation conditions. In comparison, agents trained without noise completed all laps with an average time of 6.09 seconds. It is noteworthy that the agents trained with observation noise completed laps in a more erratic manner than agents trained without noise. Examples of paths completed by agents trained with and without noise are shown in Figure 1.15. This was despite the presence of noise under evaluation conditions. We therefore chose to train agents without observation noise.

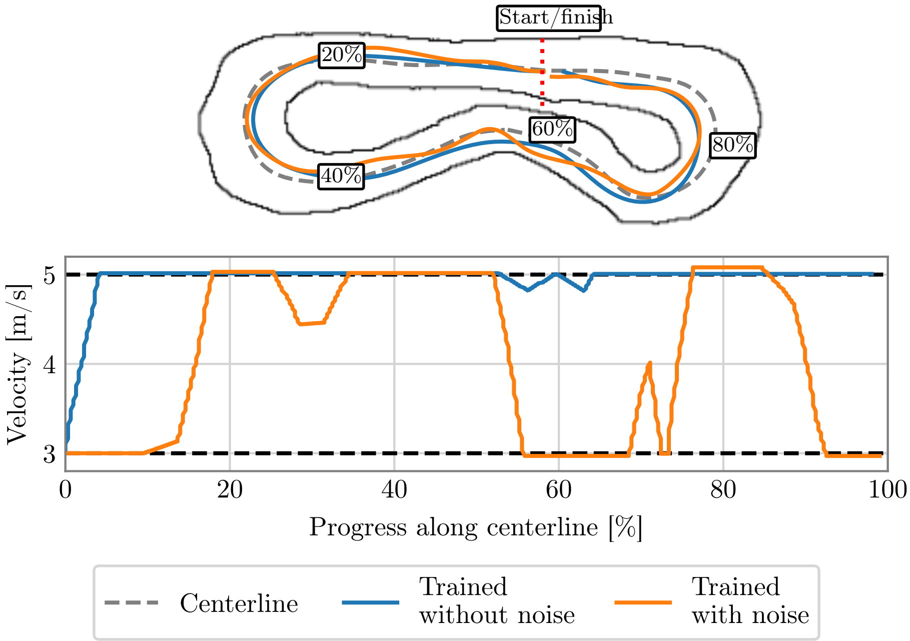


Figure 1.15: Path and velocity profiles of end-to-end agents that were trained with and without noise added to the observation vector.

1.7 Neural network hyper-parameters

Next, an investigation was conducted to determine the optimal DNN layer configuration. The layer configuration of the actor and both critics were varied together, so that they remained identical in structure. The input and hidden layers of these DNNs were initially specified to be 400 and 300 units, respectively. In this experiment, three agents were trained with input and hidden layers that were 100 units larger and smaller than the initial DNN configuration. The remaining hyperparmeters of these agents were set equal to those listed in Table 1.1. The percentage failed laps, lap times, and learning curves while training these agents are depicted in Figure 1.16.

The experimental results from Figure 1.16 indicates that increasing or decreasing the number of units in the input and hidden layers led to a deterioration in performance, particularly in terms of lap time. As a result, we have selected an input layer size of 400 units and a hidden layer size of 300 units for the actor and critic DNNs in our algorithm.

Additional experiments were conducted to determine the optimal learning rate α. The same value for α was used for the actor and critic DNNs. During this experiment, we trained three agents with learning rates of 10^{-4}, 10^{-3} and $2 \cdot 10^{-3}$, and remaining hyper-parameters set equal to Table 1.1. The performance of these agents under evaluation

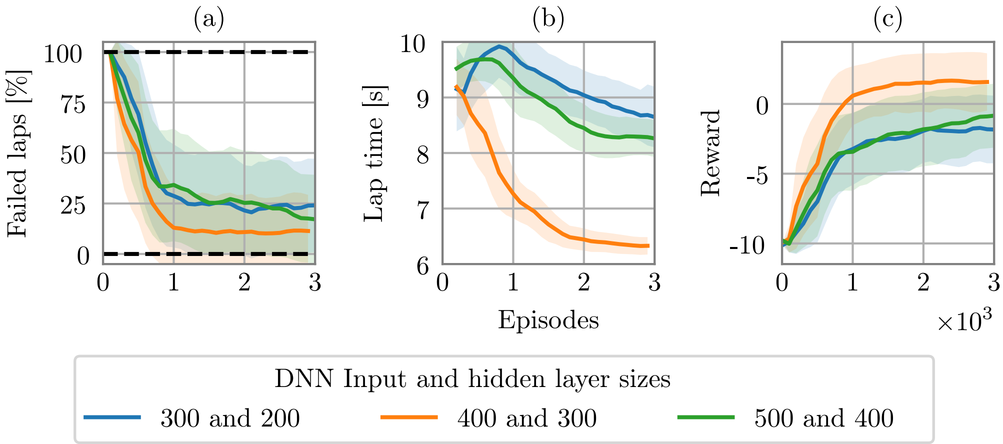


Figure 1.16: (a) the percentage failed laps and (b) lap time of completed laps during training, as well as the (c) learning curves showing episode reward for end-to-end agents with different DNN layer sizes.

conditions in terms of percentage successful laps and lap time is shown in Table 1.4. The percentage successful laps and lap time of agents were maximised and minimised, respectively, when the learning rate was set to 10^{-3}.

Learning rate, α	**Successful evaluation laps [%]**	**Average evaluation lap time [s]**	**Standard deviation of test lap time [s]**
$1 \cdot 10^{-4}$	100	6.09	0.17
$1 \cdot 10^{-3}$	100	6.07	0.20
$2 \cdot 10^{-3}$	98.67	7.29	0.53

Table 1.4: Evaluation results of end-to-end agents with actor and critic DNN learning rates between $1 \cdot 10^{-4}$ and $2 \cdot 10^{-3}$.

1.8 Velocity constraint

In our final hyper-parameter tuning investigation, we conduct experiments to determine the minimum and maximum allowable velocities. Limiting the velocity is a common technique to ensure safe operation of the vehicle. For example, Ivanov et al. [3] restrict the torque applied to the vehicle's driving motors, thereby limiting its maximum speed to 2.4 m/s. Hsu et al. [8] adopt less conservative bounds, enforcing minimum and maximum speed limits of 1.125 and 9.3 m/s, respectively.

To determine the maximum safe velocity, we trained and evaluated the behaviour of agents with v_{max} values of 5, 6, 7 and 8 m/s. Figure 1.17 illustrates the velocity and slip angle profiles of the agents as they complete one lap under evaluation conditions. Interestingly, we observed that the agents tended to maintain the maximum velocity around the track. This behaviour likely occurs because even small values of $a_{\text{long},d}$ result in large changes in velocity in-between agent samples. Further exacerbating this effect is the slow rate at which actions can be sampled from the agent.

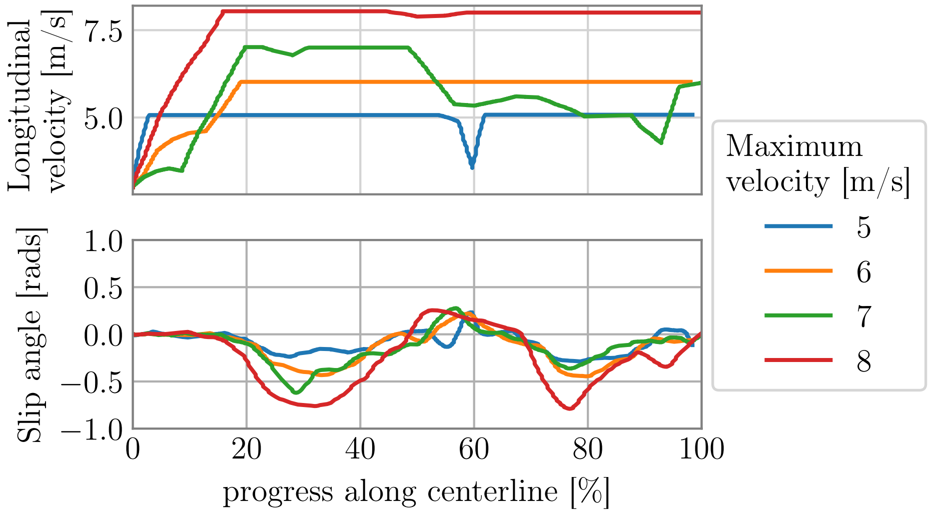


Figure 1.17: The velocity profile and slip angle of agents with different maximum velocities during one test lap.

We observed from Figure 1.17 that agents with a maximum velocity greater than 5 m/s experienced slip angles larger than 0.2 radians, which is considered both dangerous and unrealistic drifting behavior. Furthermore, the dynamic bicycle model from Chapter **??** makes assumption that tire stiffness varies linearly with lateral force. This assumption is only valid for slip angles below 0.2 radians [14]. Allowing the agent to select large velocities enables it to exploit the simulation in an unrealistic manner to achieve fast lap times. Therefore, the vehicle's maximum speed was set to 5 m/s, which was the fastest velocity that did not result in the agent driving dangerously and exploiting the simulator by operating the car at large slip angles.

We also observed that when there was no minimum velocity constraint in place, the agent would often choose to bring the car to a standstill during training, resulting in excessively long training times. The minimum speed was therefore set to to 3 m/s to prevent this behaviour. Importantly, we found that this constraint did not significantly affect the agent's performance.

1.9 End-to-end racing without model uncertainty

Having determined a set of hyper-parameters that yield optimal performance in terms of both safety and lap time for the end-to-end agent racing on the Porto track in Table 1.1, we trained and evaluated ten agents with these hyper-parameters. These agents completed 98.9% of evaluation laps with an average lap time of 6.05 seconds, achieving better performance than any other hyper-parameter set we tested. In fact, varying any of the hyper-parameters resulted in decreased performance, showing that the selected hyper-parameter values are at least locally optimal. Figure 1.18 provides a visualization of one of these agents' laps, highlighting the path taken with a color map representing the agent's velocity. Notably, the agent maintained maximum velocity for the majority of the track length. Nevertheless, the trajectory is smooth and the agent successfully navigated around the Porto circuit.

So far, our focus has been on the relatively simple Porto track. However, we further

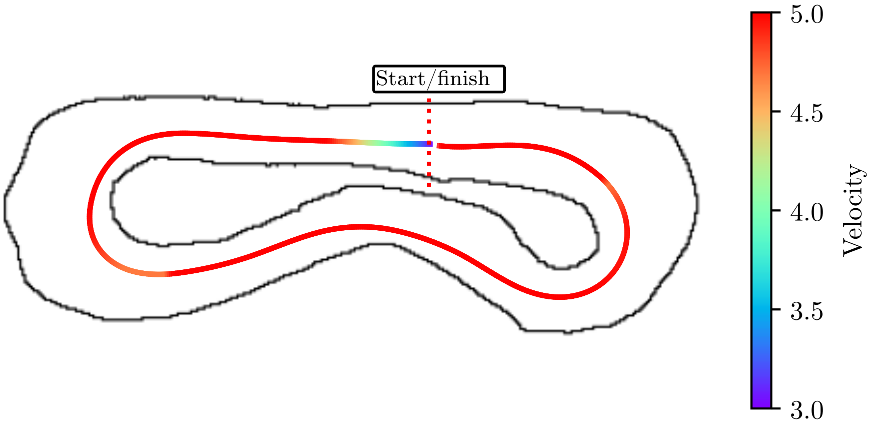


Figure 1.18: The path and velocity profile taken by an end-to-end agent completing Porto in the anti-clockwise direction.

expanded the analysis to encompass more realistic racing scenarios by training agents to navigate scaled versions of actual Formula 1 tracks. Specifically, Circuit de Barcelona-Catalunya in Spain and the Circuit de Monaco in Monaco were selected. These tracks are not only considerably larger, but also feature sharper corners and more complex geometries compared to the Porto track.

When selecting hyper-parameters for the larger tracks, a tuning procedure similar to the one presented for the Porto track was utilised. That is, the hyper-parameter were systematically varied one at a time, while keeping the other hyper-parameters constant. This hyper-parameter tuning procedure resulted in the following adjustments to Table 1.1 for agents racing on these longer tracks: the number of MDP time steps (M) was increased to $3.5 \cdot 10^5$, agent sample rate (f_{agent}) was increased to 10 Hz, and the reward signal values for r_{dist} and $r_{\text{collision}}$ were changed to 0.3 and -2, respectively. The learning curves for 10 agents trained on all three tracks using the given hyper-parameters are shown in Figure 1.19. Importantly, we observe that these agents maximise reward on each of their respective tracks.

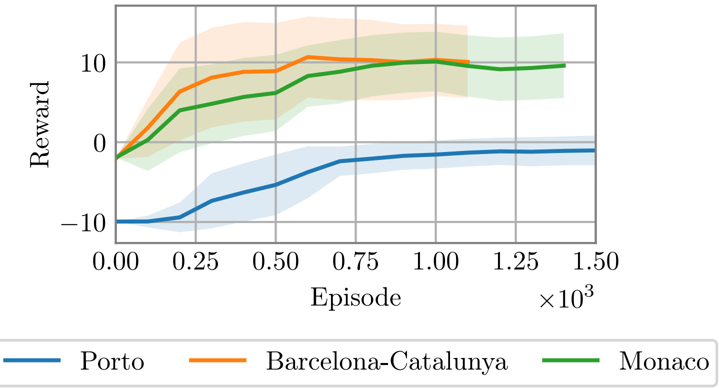


Figure 1.19: Learning curves for end-to-end agents trained and tested on Porto, Circuit de Barcelona-Catalunya and Circuit de Monaco.

Agents trained to race on Circuit de Barcelona-Catalunya completed their laps 79.0% of the time, and achieved an average lap time of 47.47 seconds under evaluation conditions.

Figure 1.20 shows the path and velocity profile taken by an agent completing Circuit de Barcelona-Catalunya under evaluation conditions. Similar to the findings on the Porto track, agents racing on the Circuit de Barcelona-Catalunya selected maximum velocity for the majority of the track, even when navigating sharp corners. Furthermore, an interesting phenomenon emerged on the Circuit de Barcelona-Catalunya that was not present on the shorter Porto track: agents tend to exhibit a slaloming behavior, which is characterized by a winding path. This slaloming effect is quite severe, occurring at nearly every section of the track.

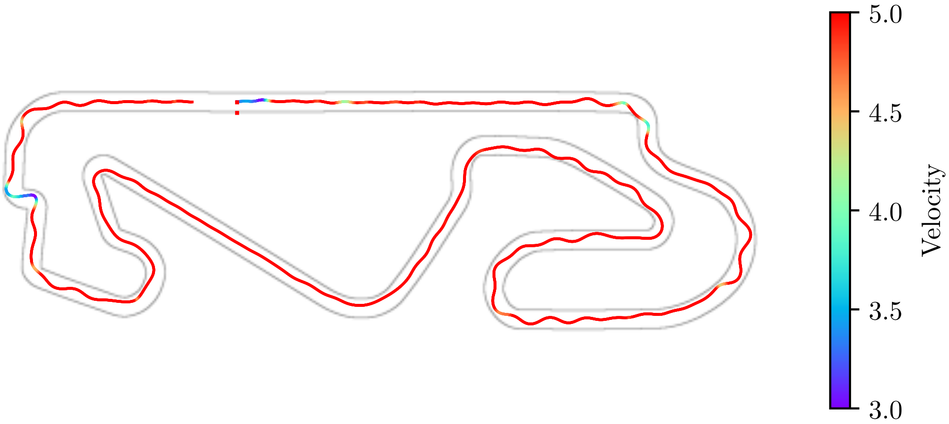


Figure 1.20: The path and velocity profile taken by an end-to-end agent completing Circuit de Barcelona-Catalunya

When assessing agents trained to race on Circuit de Monaco, we found that agents successfully completed their laps on 61.67% of their attempts, achieving an average lap time of 35.48 seconds. Figure 1.20 depicts one example of the path and velocity profile taken by an agent taht successfully completed the Circuit de Monaco under evaluation conditions. Interestingly, the slaloming is also present on the Circuit de Monaco, indicating that slaloming tends to be a common issue for end-to-end agents navigating long tracks.

1.10 End-to-end racing with model uncertainty

Up to now, results have been presented for end-to-end agents that were trained and evaluated in identical environments. However, it is important to assess the performance of agents tasked with driving in situations where the vehicle model does not match the one utilised during training. During this initial investigation, we introduced model mismatches by modifying the vehicle model parameters after training, but prior to executing the evaluation process outlined in Algorithm 2. This adjustment allows us to gain insights into how the agent performs in a more realistic setting where variations in the vehicle model are present.

Our initial focus was on investigating the impact of altering the road surface friction coefficient on the evaluation performance of trained agents. Friction is influenced by various dynamic factors, including temperature and precipitation, making it challenging to predict accurately. Consequently, it is likely that model mismatches in the road friction coefficient occur. Figure 1.22 presents a comparison of paths taken by agents evaluated

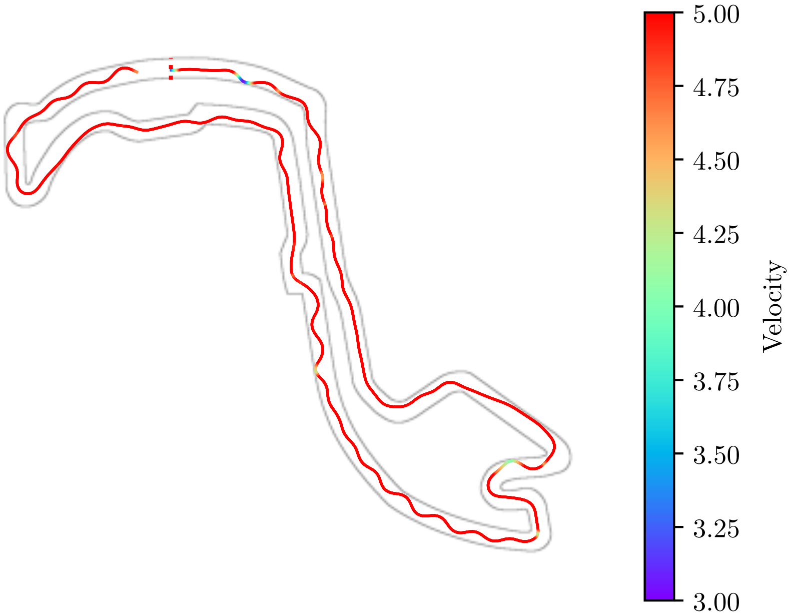


Figure 1.21: The path and velocity profile taken by an end-to-end agent completing Circuit de Monaco

with (a) the nominal friction value of 1.04, and (b) a friction value of 0.6 (equivalent to wet asphalt conditions) on a section of the Monaco track. The slip angles of the agents are visualized by color-mapping them onto their respective paths. When evaluated with the nominal friction value, the agents display slaloming behavior, resulting in maximum slip angles of approximately 0.2 radians throughout most areas of the track. In contrast, agents evaluated with decreased friction exhibit drifting behavior, characterized by slip angles exceeding 0.4 radians. The drastic increase in slip angle indicates that the learned policy of standard end-to-end agents becomes dangerous when model mismatches are present.

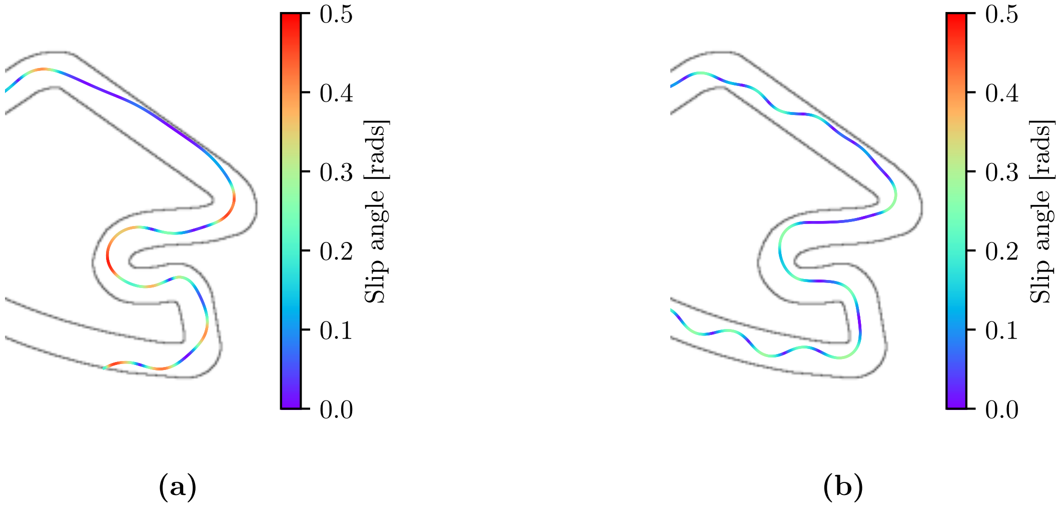


Figure 1.22: Trajectory and slip angle of an end-to-end agent racing on Circuit de Monaco with (a) the nominal road-surface friction value of 1.03, and (b) a decreased road-surface friction value of 0.6.

Notably, Figure 1.22 illustrates an instance where an agent with decreased friction crashes shortly after executing a drift maneuver. This observation emphasizes the impact of friction on the agents' handling capabilities and reinforces the significance of creating algorithms that are robust to errors in the vehicle model parameters.

1.11 Training with domain randomisation

A commonly used technique to enhance robustness against modelling errors is domain randomisation, which involves randomising simulation parameters during training. The agent is then tasked with finding a single policy that performs optimally across different parameter settings [15]. Previous autonomous racing studies have explored various approaches in this regard. Chisari et al. [9] introduced Gaussian noise to the lateral force experienced by the tires at each time step, while Ghignone et al. [16] initialized each episode by adding Gaussian noise with a standard deviation of 0.0375 to the road surface friction coefficient, which remained constant throughout the episode.

In this investigation, we adopted the approach of Ghignone et al. [16], and modified the training procedure by sampling the friction value used during every episode from a Gaussian distribution. This Gaussian distribution had a mean of 1.0489 (the nominal friction value). Two agents were trained to race on the Porto track; one with a friction coefficient standard deviation of 0.01 and another with a standard deviation of 0.05. These agents were then tasked with completing 100 laps under evaluation conditions, with the mean value of 1.0489 used in every episode.

While agents trained with friction coefficient standard deviation of 0.01 successfully completed 51% of their laps, agents trained with a standard deviation of 0.05 completed only 34% of their laps under evaluation conditions. These results indicate that domain randomisation has an adverse effect on the agents' performance, even when the agent is only tasked with racing under conditions where the average friction value is present. Figure 1.23 illustrates the paths taken by these agents during evaluation, and clearly indicates their inability to learn smooth driving behavior.

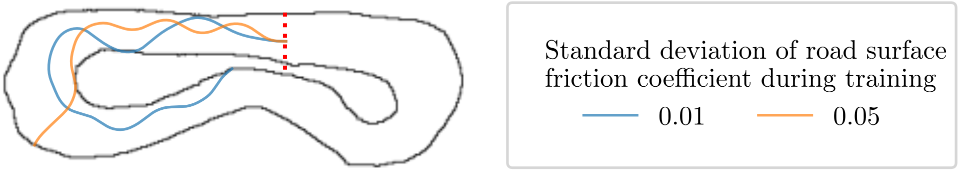


Figure 1.23: Paths taken by agents trained with randomised road-surface friction coefficients on the Porto track under evaluation conditions. During this evaluation lap, the friction coefficient was set to the nominal value of 1.0489.

Our findings suggest that the optimal policy for autonomous racing is highly sensitive to the friction coefficient of the road surface. Agents struggle to adapt their policies to changing friction values effectively, resulting in poorer performance. This sensitivity highlights the challenge of developing a single policy that performs optimally across a range of friction coefficients, demonstrating the limitations of domain randomisation in the racing context.

1.12 Summary

In this chapter, we have motivated the design of an end-to-end autonomous racing algorithm. Agents utilising this algorithm were trained to race effectively on the Porto track, successfully completing all of their laps under evaluation conditions. However, this performance did not scale to larger tracks such as Circuit de Barcelona-Catalunya or Circuit de Monaco. On these longer tracks, the performance of the agents was hindered by slaloming, and they did not complete all of their laps.

The presence of slaloming is particularly concerning when considering scenarios in which model mismatches are present. In fact, in a preliminary investigation into the effect of model mismatch on the performance of end-to-end agents, the vehicle experience a collision. This is indicative of the limitations of end-to-end algorithms under conditions where model mismatches are present, and emphasises the need for algorithms that exhibit robustness against modeling errors.

In the next chapter, we introduce our partial end-to-end solution, which aims to enhance robustness towards modeling errors and address the challenges posed by the sensitivity of the optimal policy to vehicle parameters.

Appendices

Appendix A

Supporting results

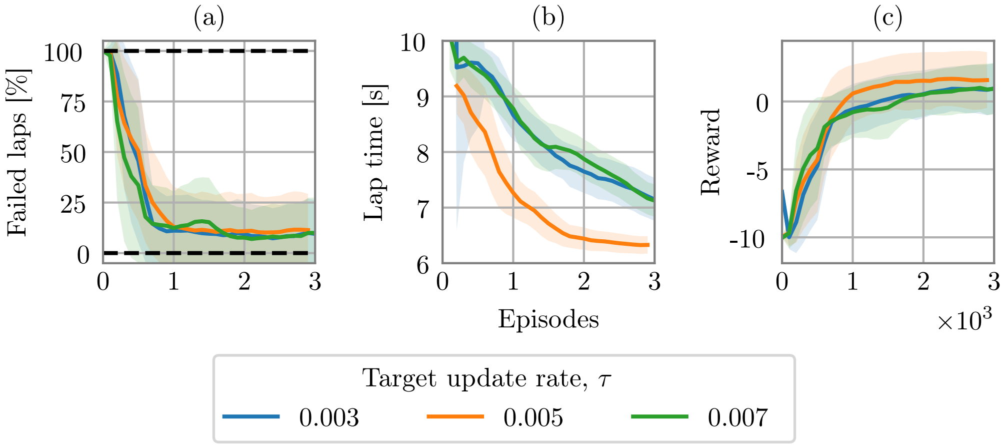


Figure A.1: Learning curves showing (a) the failure rate, i.e percentage of episodes that ended in a crash, (b) the lap time of completed laps, and (b) the episode reward for end-to-end agents with target update rates ranging from 0.003 to 0.007.

Target update rate, τ	Successful test laps [%]	Average test lap time [s]	Standard deviation of test lap time [s]
$3 \cdot 10^{-3}$	99	6.85	1.23
$5 \cdot 10^{-3}$	100	6.07	0.20
$7 \cdot 10^{-3}$	96	6.94	0.74

Table A.1: Evaluation results and training time of end-to-end agents with target update rates ranging from $3 \cdot 10^{-3}$ to $7 \cdot 10^{-3}$.

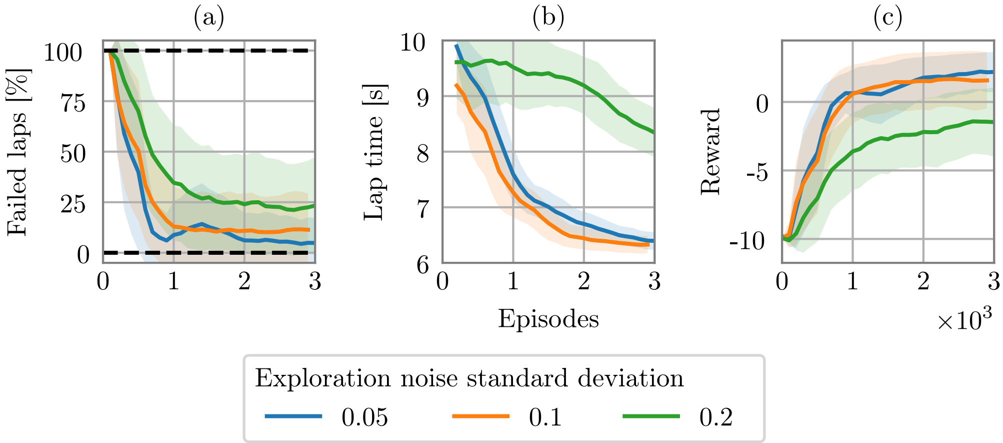


Figure A.2: Learning curves showing (a) the failure rate, i.e percentage of episodes that ended in a crash, (b) the lap time of completed laps, and (b) the episode reward for end-to-end agents with exploration noise standard deviations ranging from 0.05 to 0.2.

Exploration noise standard deviation, σ_{action}	**Successful test laps [%]**	**Average test lap time [s]**	**Standard deviation of test lap time [s]**
0.05	96	6.13	0.46
0.1	100	6.07	0.20
0.2	100	7.27	0.67

Table A.2: Evaluation results and training time of end-to-end agents with exploration noise varying from 0.05 to 0.15.

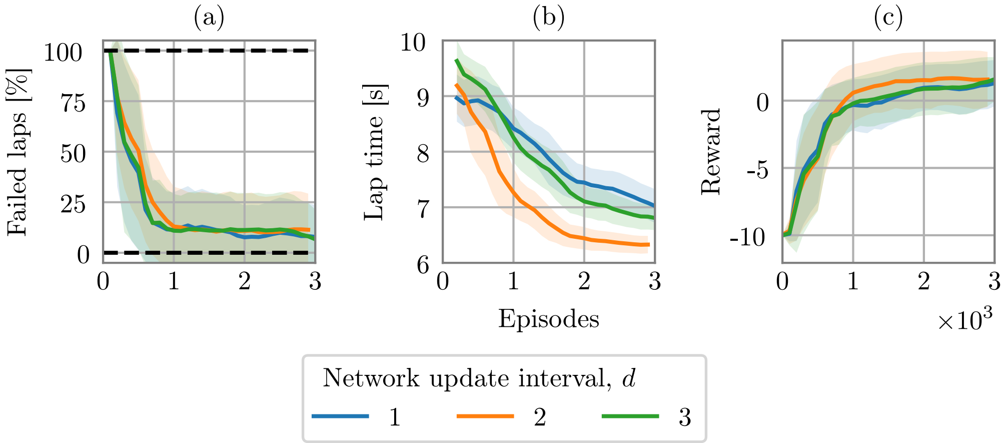


Figure A.3: Learning curves showing (a) the failure rate, i.e percentage of episodes that ended in a crash, (b) the lap time of completed laps, and (b) the episode reward for end-to-end agents with network update intervals d ranging from 1 to 3.

Number of action samples between network updates, d	**Successful test laps [%]**	**Average test lap time [s]**	**Standard deviation of test lap time [s]**
1	99	6.85	1.23
2	100	6.07	0.20
3	96	6.94	0.74

Table A.3: Evaluation results and training time of end-to-end agents with number of action samples between network updates ranging from 1 to 3.

List of References

[1] Song, Y., Lin, H., Kaufmann, E., Duerr, P. and Scaramuzza, D.: Autonomous overtaking in gran turismo sport using curriculum reinforcement learning. 2021.
Available at: `https://doi.org/10.48550/arXiv.2103.14666`

[2] Fuchs, F., Song, Y., Kaufmann, E., Scaramuzza, D. and Durr, P.: Super-Human Performance in Gran Turismo Sport Using Deep Reinforcement Learning. *IEEE Robotics and Automation Letters*, vol. 6, no. 3, pp. 4257–4264, 2021.
Available at: `http://doi.org/10.1109/LRA.2021.3064284`

[3] Ivanov, R., Carpenter, T.J., Weimer, J., Alur, R., Pappas, G.J. and Lee, I.: Case study: Verifying the safety of an autonomous racing car with a neural network controller. *HSCC 2020 - Proceedings of the 23rd International Conference on Hybrid Systems: Computation and Control ,part of CPS-IoT Week*, 2020.
Available at: `https://doi.org/10.1145/3365365.3382216`

[4] Perot, E., Jaritz, M., Toromanoff, M. and Charette, R.D.: End-to-End Driving in a Realistic Racing Game with Deep Reinforcement Learning. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, vol. 2017-July, pp. 474–475, 2017.
Available at: `https://doi.org/10.1109/CVPRW.2017.64`

[5] Jaritz, M., De Charette, R., Toromanoff, M., Perot, E. and Nashashibi, F.: End-to-End Race Driving with Deep Reinforcement Learning. *Proceedings - IEEE International Conference on Robotics and Automation*, pp. 2070–2075, 2018.
Available at: `https://doi.org/10.1109/ICRA.2018.8460934`

[6] Schwarting, W., Seyde, T., Gilitschenski, I., Liebenwein, L., Sander, R., Karaman, S. and Rus, D.: Deep latent competition: Learning to race using visual control policies in latent space. In: Kober, J., Ramos, F. and Tomlin, C. (eds.), *Proceedings of the 2020 Conference on Robot Learning*, vol. 155 of *Proceedings of Machine Learning Research*, pp. 1855–1870. PMLR, 16–18 Nov 2021.
Available at: `https://proceedings.mlr.press/v155/schwarting21a.html`

[7] Niu, J., Hu, Y., Jin, B., Han, Y. and Li, X.: Two-Stage Safe Reinforcement Learning for High-Speed Autonomous Racing. *Conference Proceedings - IEEE International Conference on Systems, Man and Cybernetics*, vol. 2020-Octob, pp. 3934–3941, 2020.
Available at: `https:doi.org/10.1109/SMC42975.2020.9283053`

[8] Hsu, B.-J., Cao, H.-G., Lee, I., Kao, C.-Y., Huang, J.-B. and Wu, I.-C.: Image-based conditioning for action policy smoothness in autonomous miniature car racing with reinforcement learning. 2022.
Available at: `https://arxiv.org/abs/2205.09658`

[9] Chisari, E., Liniger, A., Rupenyan, A., van Gool, L. and Lygeros, J.: Learning from Simulation, Racing in Reality. *Proceedings - IEEE International Conference on Robotics and*

Automation, vol. 2021-May, no. December, pp. 8046–8052, 2021.
Available at: `https://doi.org/10.1109/ICRA48506.2021.9562079`

[10] Brunnbauer, A., Berducci, L., Brandstätter, A., Lechner, M., Hasani, R., Rus, D. and Grosu, R.: Model-based versus model-free deep reinforcement learning for autonomous racing cars. 2021.
Available at: `https://arxiv.org/pdf/2103.04909v1.pdf`

[11] Remonda, A., Krebs, S., Veas, E., Luzhnica, G. and Kern, R.: Formula rl: Deep reinforcement learning for autonomous racing using telemetry data. 2021.
Available at: `https://arxiv.org/abs/2104.11106`

[12] Li, M., Wang, Y., Zhou, Z. and Yin, C.: Sampling rate selection for trajectory tracking control of autonomous vehicles. In: *2019 IEEE Vehicle Power and Propulsion Conference (VPPC)*, pp. 1–5. 2019.
Available at: `https://doi.org/10.1109/VPPC46532.2019.8952506`

[13] Capo, E. and Loiacono, D.: Short-Term Trajectory Planning in TORCS using Deep Reinforcement Learning. *2020 IEEE Symposium Series on Computational Intelligence, SSCI 2020*, pp. 2327–2334, 2020.
Available at: `https://doi.org/10.1109/SSCI47803.2020.9308138`

[14] Vorotic, G., Rakicevic, B. and Mitic, Sasa Stamenkovic, D.: Determination of cornering stiffness through integration of a mathematical model and real vehicle exploitation parameters. *FME Transactions*, vol. 41, pp. 66–71, 2013.
Available at: `https://www.mas.bg.ac.rs/_media/istrazivanje/fme/vol41/1/08_gvorotovic.pdf`

[15] Zhao, W., Queralta, P. and Westerlund, T.: Sim-to-real trainsfer in deep reinforcement learning for robotics: a survey. *IEEE Symposium Series on Computational Intelligence*, pp. 737–744, 2020.
Available at: `https://doi.org/10.48550/arXiv.2009.13303`

[16] Ghignone, E., Baumann, N., Boss, M. and Magno, M.: TC-Driver: Trajectory Conditioned Driving for Robust Autonomous Racing - A Reinforcement Learning Approach. In: *International conference on robotics and automation*. 2022. `2205.09370`.
Available at: `http://arxiv.org/abs/2205.09370`