

Gradient-Based Trajectory Optimization With Learned Dynamics

Bhavya Sukhija, Nathanael Köhler, Miguel Zamora, Simon Zimmermann,
Sebastian Curi, Andreas Krause, Stelian Coros

Abstract—Trajectory optimization methods have achieved an exceptional level of performance on real-world robots in recent years. These methods heavily rely on accurate analytical models of the dynamics, yet some aspects of the physical world can only be captured to a limited extent. An alternative approach is to leverage machine learning techniques to learn a differentiable dynamics model of the system from data. In this work, we use trajectory optimization and model learning for performing highly dynamic and complex tasks with robotic systems in absence of accurate analytical models of the dynamics. **We show that a neural network can model highly nonlinear behaviors accurately for large time horizons, from data collected in only 25 minutes of interactions on two distinct robots: (i) the Boston Dynamics Spot and an (ii) RC car. Furthermore, we use the gradients of the neural network to perform gradient-based trajectory optimization.** In our hardware experiments, we demonstrate that our learned model can represent complex dynamics for both the Spot and Radio-controlled (RC) car, and gives good performance in combination with trajectory optimization methods.

I. INTRODUCTION

Robots are expected to perform complex and highly dynamic maneuvers in unknown environments [1]–[5]. Traditional trajectory-based optimal control approaches are often used for this purpose [6]. Trajectory optimization methods are well established and give physically accurate trajectories which exhibit complex and dynamic behaviors [7]–[9].

However, trajectory optimization methods require an accurate dynamics model of the system. Traditional modeling approaches either rely on simplified models and/or invest immense engineering effort in selecting relevant features for system identification [10]–[13]. For highly dynamic and complex systems, it is difficult—if not impossible—to derive models following these approaches. For example, the Boston Dynamics Spot robot¹ has an on-board inaccessible low-level controller, and only allows control of high-level commands. This makes the Spot a complete black box and possibly non-Markovian system. Thus, deriving a dynamics model for the Spot’s high-level behavior is very challenging. However, understanding its behavior is extremely essential for planning, especially when operating the robot in unknown environments (Fig. 1).

The goal of this work is to **leverage gradient-based trajectory optimization methods for dynamic robotic systems**, such as the Spot and the RC car, whose dynamics are unknown



Fig. 1: We evaluate two distinct robots Boston Dynamics Spot and a dynamic RC car. For the Spot, we consider a normal and slippery terrain. We simulate the slippery terrain by using socks at its feet. This causes the spot to slip and lose balance, as depicted on the top three images.

and challenging to model. Towards that goal, we use **data-driven methods to obtain the dynamics model**. Specifically, we record **short spans of data (around 25 minutes) directly on the system and learn a model from the recorded data**. We employ parametric models such as multilayer feedforward neural networks [14] and Recurrent Neural Networks (RNN) [15] to capture the dynamics. This allows us to model the system while requiring no first principles or engineered solutions. We then leverage the learned model to perform complex and dynamic maneuvers through trajectory optimization. This approach is evaluated on two distinct mobile robots: the Spot, and a drifting RC car (Fig. 1). For the Spot, we consider **two different terrains; normal/nominal terrain, and slippery terrain** that we simulate by putting socks at the robot’s feet (see the top half of Fig. 1). In our results, we demonstrate that we can reliably learn the robot’s dynamics with data collected in around 25 minutes, and leverage the model to execute a complex trajectory using trajectory optimization. Furthermore, when operating on more slippery terrain, we show that by recording just **15 minutes of additional data, we can adapt our model to the new surface successfully**. Similarly, on the RC car, we show that after collecting a small corpus of data consisting of **15 minutes** of interactions, we can perform highly complex and dynamic maneuvers like **drifting**, which are generally difficult to model [16], [17].

Our contributions are as follows, we demonstrate that (i) our learned model works successfully for trajectory optimization, (ii) can be adapted for different operating conditions, (iii) is capable of achieving complex drifting

The authors are with the Department of Computer Science, ETH, Zürich, Switzerland. nate@striking.ch; (bhavya.sukhija; miguel.zamora; simon.zimmermann; sebastian.curi, stelian.coros)@inf.ethz.ch; krausea@ethz.ch

¹<https://www.bostondynamics.com/products/spot>

maneuvers, and (iv) gives significant performance gain in lap-time for the RC car compared to a human expert. Lastly, to the best of our knowledge, we are the first to demonstrate on the example of the Spot, a dynamic, black box, and closed-loop system, that a RNN dynamics model can be learnt from scratch in 25 minutes and enable agile control using gradient-based trajectory optimization. Refer to the accompanying video² for more details.

II. RELATED WORK

There has been a considerable amount of research in learning-based control for robotic systems [18]–[27]. Most works typically use Gaussian processes (GPs) [28] to learn the system dynamics [29]. GPs are powerful non-parametric machine learning models that can exhibit strong theoretical guarantees, but they scale poorly for large datasets [28]. Alternatively, neural networks have been suggested as an expressive class of parametric models [30], [31]. Specifically, Multilayer Perceptron (MLP), and RNN have shown promising results in modeling unknown nonlinear dynamical systems [32]–[34]. Neural networks can capture complex behaviors, and therefore are often used in deep model-based reinforcement learning (MBRL) [24]–[27]. Here, generally, the methods either also learn a control policy, i.e., end-to-end control (for instance, an MLP that outputs control signals for a given state), or use population-based search heuristics such as the cross-entropy method [35] for trajectory optimization. Nonetheless, there are notable exceptions such as [36], [37] that deploy traditional trajectory optimization solvers. In [36], local time-varying linear dynamics are learned and then integrated into an iLQG [38] based trajectory optimizer, which is finally used to learn a parametric policy. In contrast, we learn a global dynamics model and use a **gradient-based direct shooting trajectory optimization approach** for simplicity. Our approach is straightforward to implement and works successfully on two distinct and challenging mobile robots. A similar approach is used in [37], where trajectory optimization is performed using the out-of-the-box Adam optimizer [39]. Specifically, their work focuses on regularizing trajectory optimization to prevent the exploitation of model inaccuracies using denoising auto-encoders. The proposed scheme is then tested in simulation. However, the objective in our work is different since we focus on the successful deployment on real hardware. Particularly, we want to demonstrate how our learned models can be successfully leveraged to optimize trajectories and deploy them on real, and dynamic robots. Most closely related to our approach are [23], [40]. The approach in [40] uses a neural network to learn the dynamics of a RC car and perform dynamic maneuvers through a model predictive controller with a sampling-based optimization scheme. However, in our work, we can achieve similar dynamic behavior using our gradient-based trajectory optimization approach, which is fast, especially in high dimensions, and known to **have strong**

local convergence guarantees [41]. In [23] a model for the Spot is learned and leveraged for gradient-based trajectory optimization. However, [23] considers a parametric model with hand-picked features. For a black-box system like the Boston Dynamics Spot, hand-picking features is a time-consuming, and often unintuitive process. We overcome this limitation by employing RNNs instead of hand-picking features. RNNs are suitable for sequence modeling by design [42], and have been successfully used to learn dynamic models for predicting longer time dependencies [32], [34], [43]–[45]. However, most of these works focus on simulation setups, and do not consider complex and dynamical real-world systems like the Spot. Especially on the case of the Spot, due to its low-level controller and gait cycle, the system might not be non Markovian [46]. However, the RNN allows us to capture its dynamics by a learning the hidden state.

III. METHOD

The goal of this paper is to find an optimal finite-horizon control sequence for our dynamical system. We formulate this as a *time-discretized* trajectory optimization problem: Let $\mathbf{X} := (\mathbf{x}_1, \dots, \mathbf{x}_n)$ and $\mathbf{U} := (\mathbf{u}_0, \dots, \mathbf{u}_{n-1})$ be the stacked state and control input vectors for a total of n trajectory steps. Given a known initial state \mathbf{x}_0 , we write the trajectory optimization problem as

$$\min_{\mathbf{U}} \quad \ell(\mathbf{X}, \mathbf{U}) \quad (1)$$

$$\text{s.t.} \quad \mathbf{x}_{i+1} = \mathbf{x}_i + f(\mathbf{x}_i, \mathbf{u}_i), \quad \forall i = 0, \dots, n-1$$

with total cost ℓ and a deterministic state transition function f . The latter models the dynamics of the physical system we want to control.

A. Trajectory Optimization

We solve the trajectory optimization problem as stated in Equation (1) through a gradient based method. Hereby, we are interested in finding the optimal control parameters that minimize the total cost $\ell(\mathbf{X}, \mathbf{U}) := \ell(\mathbf{X}(\mathbf{U}), \mathbf{U})$. By following the chain rule, we can compute the gradient as

$$\frac{d\ell}{d\mathbf{U}} = \frac{\partial \ell}{\partial \mathbf{X}} \frac{d\mathbf{X}}{d\mathbf{U}} + \frac{\partial \ell}{\partial \mathbf{U}}. \quad (2)$$

We then perform gradient-based optimization either using standard optimizers such as Adam [39] or a simple line search for the step size. **To avoid convergence to bad local optimas, we run the optimization with random initialisations and pick the best sequence.**

The Jacobian $\frac{d\mathbf{X}}{d\mathbf{U}}$ depends on the state transition function. Specifically, it is a **lower diagonal matrix that can also be computed via the chain rule:**

$$\begin{aligned} \left[\frac{d\mathbf{X}}{d\mathbf{U}} \right]_{i,j} &= \frac{\partial x_i}{\partial u_j}, \\ \frac{\partial x_i}{\partial u_j} &= \begin{cases} \frac{\partial x_i}{\partial x_{i-1}} \frac{\partial x_{i-1}}{\partial u_j}, & \forall j < i-1, \\ \frac{\partial f}{\partial \mathbf{u}} \Big|_{(\mathbf{x}_{i-1}, \mathbf{u}_{i-1})} & j = i-1, \\ 0 & \text{else,} \end{cases} \end{aligned} \quad (3)$$

²<https://polybox.ethz.ch/index.php/s/tKTAncxW9hdkVv>

$$\frac{\partial x_i}{\partial x_j} = \begin{cases} \frac{\partial x_i}{\partial x_{i-1}} \frac{\partial x_{i-1}}{\partial x_j}, & \forall j < i-1, \\ 1 + \frac{\partial f}{\partial x} \Big|_{(x_{i-1}, u_{i-1})} & j = i-1, \\ 1 & j = i, \\ 0 & \text{else.} \end{cases}$$

1) Control Costs

We construct the cost function, $\ell(\mathbf{X}, \mathbf{U})$ (Equation (1)), as a **sum over input penalties and state-wise immediate costs**. In particular, we encourage smooth trajectories by penalizing both **high magnitudes and high changes in the control inputs** throughout the entire time horizon. The corresponding cost can be written as

$$\ell_{\text{reg}}(\mathbf{U}) = w_u \sum_{i=0}^{n-1} \|\mathbf{u}_i\|^2 + w_{\text{jerk}} \sum_{i=1}^{n-1} \|\mathbf{u}_i - \mathbf{u}_{i-1}\|^2. \quad (4)$$

Here, w_{jerk} and w_u , are weights used to penalize jerks, and large control magnitudes respectively. We define the state-wise immediate costs as

$$\ell_{\text{target}}(\mathbf{X}) = \sum_{i \in I} \|\mathbf{x}_i - \bar{\mathbf{x}}_i\|^2, \quad (5)$$

Here, $\bar{\mathbf{x}}_i$ is the predefined target state at time step $i \in I$. We select the target state according to a reference trajectory we want the system to follow. The overall cost is a weighted sum of the two objectives,

$$\ell(\mathbf{X}, \mathbf{U}) = w_{\text{target}} \ell_{\text{target}}(\hat{\mathbf{X}}) + w_{\text{reg}} \ell_{\text{reg}}(\mathbf{U}). \quad (6)$$

B. Learning the Dynamics

Though trajectory optimization in itself is well studied, the main challenge for us stems from the unknown dynamics f . To this end, we represent f as a parametric model f_θ , and learn the parameters θ . Specifically, we **record a dataset of transitions $\{x_k, u_k, x_{k+1}\}$ directly on the robots and use the collected data to learn the dynamics in a supervised manner by maximizing the data likelihood**. The learned model f_θ is then used for trajectory optimization. Then, we fix the learned model f_θ , and leverage it to perform trajectory optimization to find the optimal control input \mathbf{U}^* :

$$\min_{\mathbf{U}} \ell(\hat{\mathbf{X}}, \mathbf{U}) \quad (7)$$

$$\text{s.t.} \quad \hat{\mathbf{x}}_{i+1} = \hat{\mathbf{x}}_i + f_\theta(\hat{\mathbf{x}}_i, \mathbf{u}_i), \quad \forall i = 0, \dots, n-1$$

where $\hat{\mathbf{X}} := (\hat{\mathbf{x}}_1, \dots, \hat{\mathbf{x}}_n)$ is the concatenation of predicted n -step trajectory using the learned model.

1) Learning for the Spot

In this work, we consider a three-dimensional input space for the Spot, which corresponds to the forward, sideward, and angular velocities. This allows us to move the robot's base in a 2D plane only. Therefore, the states of the Spot we consider are the position, orientation, and velocities in the local frame, i.e.,

$$\mathbf{x} = [p_x^l, p_y^l, \psi, v_x^l, v_y^l, \dot{\psi}]^T.$$

Furthermore, when considering mobile robots operating on homogeneous terrains, we can assume that the dynamics are invariant with respect to the robot's global position. Therefore, we do not consider the global positions as input for f when learning the dynamics. This reduces the input space for f and potentially helps in faster generalization.

We control the Spot at a **20 Hz frequency and fix its gait to trot**. For the state measurements, we use the on-board Spot

state estimator. From initial experiments, we notice that the leg joint configuration and gait cycle of the Spot influences its behavior. This influence is not captured by description of our state space. To this end, we learn a hidden state by using an RNN, specifically a gated RNN (GRU) [47], with the hope that the hidden state can represent the true dynamics of the robot better. Moreover, we deliberately choose a simple state-space representation of the Spot and leverage the learned hidden state to compensate for other unaccounted influences.

2) Learning for the RC car

We use an RC car with a high torque motor, which allows us to perform dynamic maneuvers that involve loss of traction and drifting. The state of the car consists of three degrees of freedom, two for its position, and one for its orientation. This corresponds to the same state space as the Spot. The inputs for the car are the **forward velocity and steering angle**. We use the Optitrack for robotics motion capture system³. We use a **feed-forward neural network to capture the dynamics**.

3) Regularization and Continuous Activation Functions

Since we use our learned model for gradient-based trajectory optimization, we prefer smooth derivatives. Smooth derivatives not only ease the trajectory optimization itself but also result in smoother action sequences. To this end, we choose continuous activation functions, such as **Gaussian Error Linear Units (GELU)** [48] for our neural networks and apply a L2 regularization to avoid overfitting.

Our approach is summarized in Algorithm 1.

Algorithm 1 Learned Model for Trajectory Optimization

Require: Data: \mathcal{D} , Initial Input Sequence U_0
Train model from dataset \mathcal{D} : $\theta^* \leftarrow \min_{\theta} L(\mathcal{D}|f_\theta)$.
 $k \leftarrow 0$
while Not Converged and $k < N_{\text{max}}$ **do**
 $U_{k+1} \leftarrow U_k - \eta_k \frac{d\ell_\theta}{dU}$
 $k \leftarrow k + 1$
return U_k

IV. RESULTS

This section presents experimental results on hardware achieved through trajectory optimization using learned models. The aim of our experiments is to demonstrate that our learned models can (i) capture nonlinear dynamics of the robots well, and (ii) be successfully used for trajectory optimization. Therefore, we learn models for two mobile robots, the Spot and the RC car. **To demonstrate the strength of our models, we perform open-loop trajectory optimization**. Lastly, we demonstrate how our learned model can be successfully integrated into closed-loop trajectory optimization on the example of the RC car on a race track. A summary table of our open-loop trajectory optimization results is presented in Table I. We also provide a video (see Section I) of our dynamic motions on hardware.

³<https://optitrack.com/applications/robotics/>

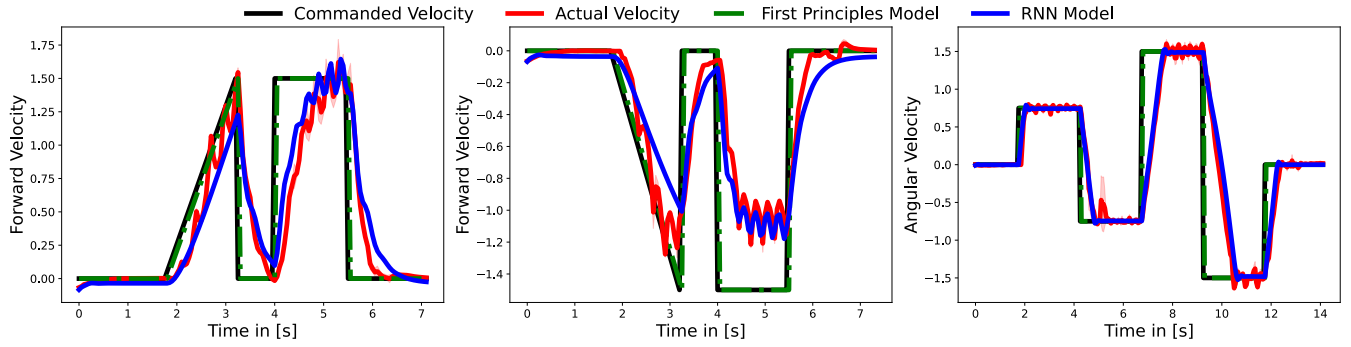


Fig. 2: Spot's behavior on normal terrain (red) vs predictions using the first principle model (green), and the learned model (blue) over 3 independent runs for forward, backward, and turning motion commands (black-dashed line).

A. Boston Dynamics Spot Experiment

Due to the black-box nature of the Spot, it is difficult to derive a model from the first principles. A simple model one may consider is $v_{x,k+1}^l = u_k^{Forward}$, $v_{y,k+1}^l = u_k^{Sideward}$, and $\dot{\psi}_{k+1} = u_k^{Turning}$, i.e., desired/commanded velocities are equal to the actual velocities of the robot. From this, the positions can be determined using Euler forward integration. This model is compared to our learned model on test trajectories with forward and turning motions (Fig. 2). From Fig. 2, we can deduce that our learned model is considerably better than the simple first principles model. For instance, it is noticeable from the figure that given the low-level controller, the robot cannot walk backward as fast as commanded. Particularly, it can walk forward faster than backward. While the simple model cannot capture this behavior, our learned model can. This highlights the importance of learning a dynamics model for the Spot and also showcases the limitations of our first principle model. On the left in Fig. 4, we compare the test-error accumulation over open-loop predictions for varying horizons between (i) simple model, (ii) neural network model, and (iii) RNN (GRU) model. The errors of the simple model increase drastically with the horizon length. Nonetheless, the neural network model and the GRU model show better performance, with the GRU giving better results.

To further demonstrate the benefits of learning for the robot, we simulate a slippery terrain by putting socks at the feet of the Spot. The socks cause the Spot to slide and therefore slip (Fig. 1). As depicted in Fig. 3, this leads to a slight difference in Spot's tracking performance. We capture this change in dynamics, by recording another dataset for the slippery case for 15 minutes and adapting the learned model from before by retraining on the new dataset. On the right-hand side of Fig. 4, we compare the adapted model with the unadapted one for test data recorded on the slippery surface. From the figure, we can conclude that the adapted model performs slightly better than the unadapted one as the prediction horizon increases.

1) Trajectory Optimization

We leverage our learned model to perform trajectory optimization. In order to quantify the prediction strength of our model, we execute an open-loop rollout and measure the deviation between the expected and observed trajectory.

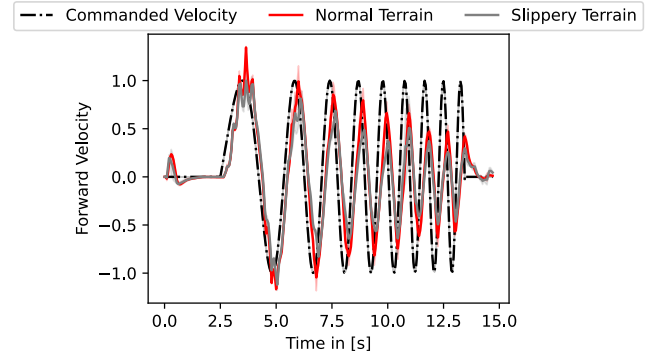


Fig. 3: Tracking performance of the Spot on normal and slippery terrain (Spot with socks).

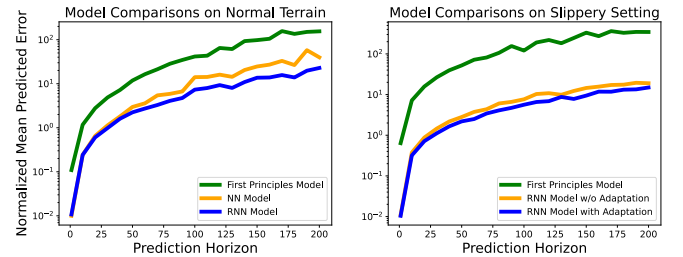


Fig. 4: Left: Normalized mean prediction error accumulation, in log-scale, over multiple horizon lengths for simple model, neural network model, and RNN model for the normal terrain. Right: Normalized mean prediction error accumulation, in log-scale, over multiple horizon lengths for the simple, unadapted, and adapted model for slippery terrain.

The Spot has a very good low-level controller, however because we consider an open-loop input sequence, its motion deviates considerably from the desired trajectory. For our experiments, we provide a sinusoidal motion as a reference to execute a dynamic zig-zag drill with the Spot (Fig. 5). The horizon for this trajectory is 150. Therefore, for the open-loop execution, an accurate model is required to avoid the accumulation of errors over the horizon length. We execute the same trajectory four times. Furthermore, we perform trajectory optimization using the first principles model and compare its performance to our learned one. In Fig. 6, we compare the performance of the two trajectories. Specifically, we depict the error between the predicted and real trajectory.

Our results show that the learned model performs considerably better than the first principle one, i.e., has considerably (around a factor of five) lower errors.



Fig. 5: Spot open-loop zig-zag trajectory obtained through trajectory optimization with the learned model.

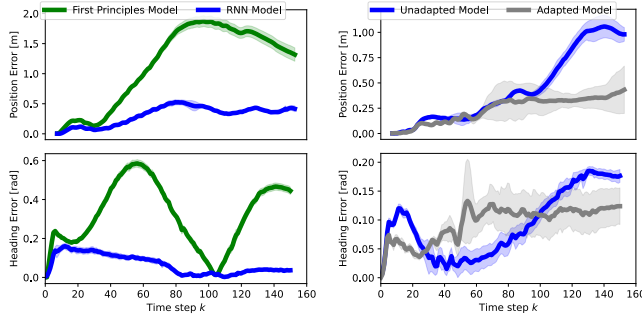


Fig. 6: Comparisons for Spot experiments. Left: Prediction error using the first principles model (green) and the learned model (blue). Right: Prediction error of the unadapted model (blue) and the model adapted for the slippery floor (grey). For both cases, we average over four independent trajectories and also depict the standard deviation.

We perform the same experiment for the slippery setting. Here, we compare the trajectory of our adapted model to the unadapted one, i.e., the model solely trained on nominal/normal terrain. Fig. 6 compares the performance of the two trajectories. Our results show that overall we achieve smaller errors when using our adapted model. Furthermore, we notice that the standard deviation in our executed trajectories on the slippery terrain is higher. We believe this is due to the slipping of the robot.

B. RC Car

For the RC car, we execute multiple open-loop rollouts. Furthermore, we evaluate the performance of our model in closed-loop using model predictive control [49].

1) Trajectory Optimization

We perform trajectory optimization for three different scenarios (i) parallel parking (Fig. 7), (ii) dynamic reverse, and (iii) drifting turn (Fig. 9). All three scenarios include dynamic drifting maneuvers. For each scenario, we repeat the same experiment 20 times for slightly different starting positions. The results are summarized in table I.

TABLE I: The error between the planned and achieved position for open-loop trajectory optimization.

	Spot zig-zag normal terrain	Spot zig-zag slippery terrain	Parallel Parking	Dynamic Reverse	Drifting Turn
Trajectory Length	150	150	40	40	60
Mean L2	0.31	0.24	0.38	0.21	0.37
Position Error [m]					
STDev L2	0.03	0.09	0.10	0.074	0.21
Position Error [m]					
Mean Absolute	0.07	0.01	0.13	0.50	0.10
Heading Error [rad]					
STDev Absolute	0.01	0.03	0.08	0.10	0.10
Heading Error [rad]					

2) Closed-Loop Trajectory Optimization

We perform experiments in an online trajectory optimization setting by driving on a predefined race track (Fig. 10). Closed-loop trajectory optimization is the more conventional approach to controlling robots. Thus, in this experiment, we demonstrate how it can be successfully achieved with our learned model. To this end, we apply trajectory optimization in a receding horizon fashion. The overall objective (introduced as costs $\ell(\mathbf{X}, \mathbf{U})$) is to advance along the race track with a predefined high velocity for the entire time horizon of the trajectory. Additionally, we add a cost for track excursions. Hereby, we choose a horizon of $n = 20$ and control the robot at a frequency of 20 Hz. We leverage parallelization to estimate the derivatives $\frac{d\hat{\mathbf{X}}}{d\mathbf{U}}$ with finite differences for this particular experiment. Additional performance metrics of the RC car on the racetrack are given in Fig. 10. As shown in the video, the car is able to race through the track with high velocity while performing dynamic maneuvers.

C. Discussion

In our open-loop trajectory optimization experiments, we notice that model inaccuracies accumulate over the trajectory horizon (Table I). Even though these inaccuracies are small, they can still have an impact on the robot's performance. Nonetheless, we can compensate for these inaccuracies using feedback control, as we demonstrate this on the example of the RC car Section IV-B.2. Furthermore, during our model selection process (Section III-B.3), we hypothesized that paying close attention to regularization and selection of activation functions would help in obtaining smoother action sequences. Clearly, smoother action sequences are preferred when deploying directly on real-world hardware. We validate this hypothesis on the RC car example. As depicted in Fig. 11, the control sequence resulting from the network with GELU activations is considerably smoother than the one obtained using Rectified Linear Units (ReLU) activations. We trace this back to the gradient used for trajectory optimization, which is much noisier for ReLU as well. Lastly, for the Spot experiments, we simulate a drastic shift in the robot's operating condition in form of the slippery terrain. However, given the robot's state-of-the-art low-level controller, we expect it to perform reasonably well in settings where the shift in the operating conditions is not drastic. To this end, we test the Spot on dry grass and notice that at least for forward, backward, and turning motions, the



Fig. 7: Executed trajectory for the parallel parking scenario.

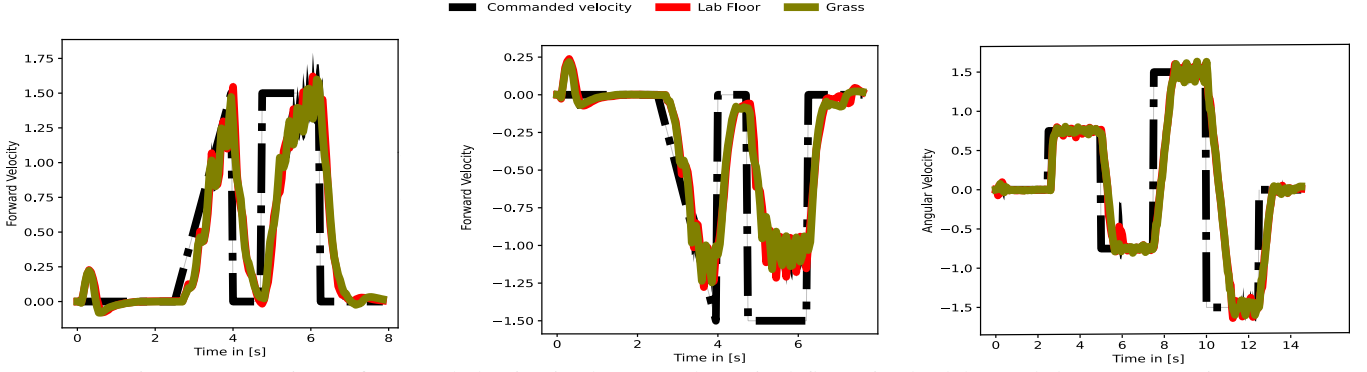


Fig. 8: Comparison of Spot's behavior in the normal/nominal floor (in the lab), and dry grass terrain.

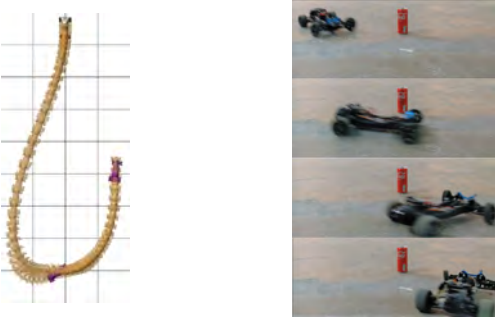


Fig. 9: The trajectory for the drifting turn scenario.

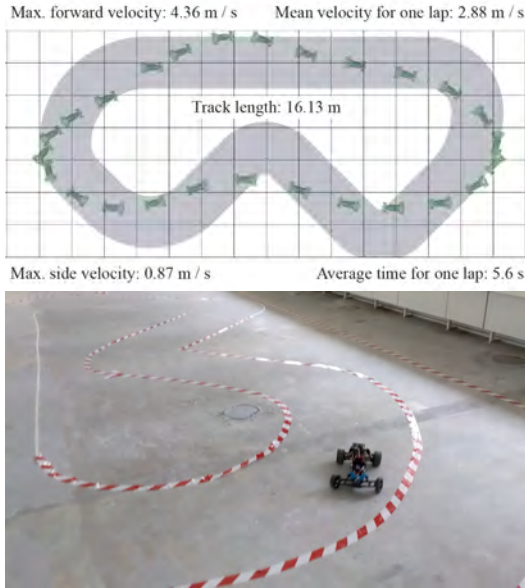


Fig. 10: The physical RC car (below) with feedback controller.

tracking performance on the lab floor and dry grass is equally good, see Fig. 8. Thus, we believe in such settings, we can still leverage the model learned in the lab environment.

V. CONCLUSION AND FUTURE WORK

The goal of this work is to leverage traditional trajectory optimization approaches for systems with unknown

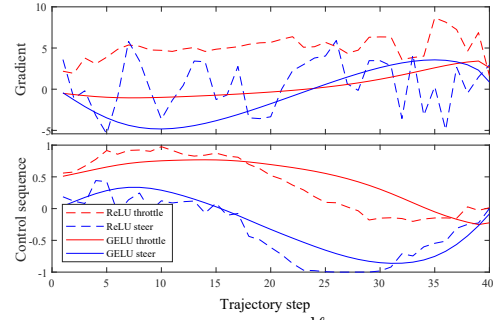


Fig. 11: Comparison of gradients $\frac{d\ell}{du}$ and control sequences u for neural network architectures with ReLU and GELU activations. ReLU activations results in noisier (i.e. more fluctuating) gradients and input sequences than the network with GELU activations.

dynamics. We demonstrate that this can be achieved through machine learning on two distinct and challenging robots. Specifically, our results show that we can capture the dynamics of the robots, adapt our learned model to new operating conditions, and perform dynamic maneuvers using trajectory optimization.

This work opens various avenues for further research. For instance, the exploitation of model inaccuracies by policy optimizers has been investigated in the literature [25], [50], [51]. Suggested strategies are the use of probabilistic ensembles [25], [27], [52]–[54], shorter task horizons [51] and denoising autoencoders [37]. Since in our work we observed an accumulation of model inaccuracies (Table I), in the future these approaches can be integrated to study their influence on performance. Additionally, in this work, the data used to learn the model was recorded offline. However, methods such as [27] automate the data acquisition by exploring the system dynamics in an episodic online learning setting. Future work may consider leveraging these advances. Furthermore, in this work, we were interested in finding smooth rather than accurate gradients. We think studying the influence of model selection on learning accurate dynamics and gradients, as well as leveraging structured learning techniques for capturing robot dynamics [55], is an exciting direction for future work.

REFERENCES

- [1] F. Rubio, F. Valero, and C. Llopis-Albert, "A review of mobile robots: Concepts, methods, theoretical framework, and applications," *International Journal of Advanced Robotic Systems*, vol. 16, no. 2, 2019.
- [2] S. Thrun, M. Montemerlo, H. Dahlkamp, D. Stavens, A. Aron, J. Diebel, P. Fong, J. Gale, M. Halpenny, G. Hoffmann, K. Lau, C. Oakley, M. Palatucci, V. Pratt, P. Stang, S. Strohband, C. Dupont, L.-E. Jendrossek, C. Koelen, C. Markey, C. Rummel, J. van Niek-erk, E. Jensen, P. Alessandrini, G. Bradski, B. Davies, S. Ettinger, A. Kaehler, A. Nefian, and P. Mahoney, *Stanley: The Robot That Won the DARPA Grand Challenge*. Springer Berlin Heidelberg, 2007.
- [3] M. Blösch, S. Weiss, D. Scaramuzza, and R. Siegwart, "Vision based mav navigation in unknown and unstructured environments," in *2010 IEEE International Conference on Robotics and Automation*, 2010, pp. 21–28.
- [4] C. Gehring, S. Coros, M. Hutter, C. Dario Bellicoso, H. Heijnen, R. Diethelm, M. Bloesch, P. Fankhauser, J. Hwangbo, M. Hoepfner, and R. Siegwart, "Practice makes perfect: An optimization-based approach to controlling agile motions for a quadruped robot," *IEEE Robotics Automation Magazine*, vol. 23, no. 1, pp. 34–43, 2016.
- [5] J. Lee, J. Hwangbo, L. Wellhausen, V. Koltun, and M. Hutter, "Learning quadrupedal locomotion over challenging terrain," *Science Robotics*, no. 47, 2020.
- [6] L. Biagiotti and C. Melchiorri, *Trajectory Planning for Automatic Machines and Robots*, 1st ed. Springer Publishing Company, Incorporated, 2008.
- [7] M. Geilinger, R. Poranne, R. Desai, B. Thomaszewski, and S. Coros, "Skaterbots: Optimization-based design and motion synthesis for robotic creatures with legs and wheels," in *Proceedings of ACM SIGGRAPH, A. T. on Graphics (TOG)*, Ed., vol. 37. ACM, August 2018.
- [8] J. M. Bern, P. Banzet, R. Poranne, and S. Coros, "Trajectory optimization for cable-driven soft robot locomotion," *Robotics: Science and Systems XV*, 2019.
- [9] S. Zimmermann, R. Poranne, J. M. Bern, and S. Coros, "PuppetMaster," *ACM Transactions on Graphics*, vol. 38, no. 4, pp. 1–11, 2019.
- [10] K. Åström and P. Eykhoff, "System identification - a survey," *Automatica*, vol. 7, no. 2, pp. 123–162, 1971.
- [11] L. Ljung, *System Identification*. John Wiley & Sons, Ltd, 1999.
- [12] K. Kozłowski, *Modelling and Identification in Robotics*. John Wiley & Sons, Ltd, 1998.
- [13] D. Nguyen-Tuong and J. Peters, "Model learning for robot control: a survey," *Cognitive Processing*, vol. 12, no. 4, pp. 319–340, 2011.
- [14] K. Hornik, M. Stinchcombe, and H. White, "Multilayer feedforward networks are universal approximators," *Neural Networks*, vol. 2, no. 5, pp. 359–366, 1989.
- [15] I. Goodfellow, Y. Bengio, and A. Courville, *Deep learning*. MIT press, 2016.
- [16] J. Z. Kolter, C. Plagemann, D. T. Jackson, A. Y. Ng, and S. Thrun, "A probabilistic approach to mixed open-loop and closed-loop control, with application to extreme autonomous driving," in *2010 IEEE International Conference on Robotics and Automation*, 2010, pp. 839–845.
- [17] A. Liniger, A. Domahidi, and M. Morari, "Optimization-based autonomous racing of 1:43 scale rc cars," *Optimal Control Applications and Methods*, vol. 36, no. 5, p. 628–647, Jul 2014.
- [18] T. M. Moerland, J. Broekens, and C. M. Jonker, "Model-based reinforcement learning: A survey," 2021.
- [19] M. P. Deisenroth and C. E. Rasmussen, "Pilco: A model-based and data-efficient approach to policy search," in *Proceedings of the 28th International Conference on machine learning (ICML-11)*, ser. ICML'11. Omnipress, 2011.
- [20] M. P. Deisenroth, D. Fox, and C. E. Rasmussen, "Gaussian processes for data-efficient learning in robotics and control," *IEEE transactions on pattern analysis and machine intelligence*, pp. 408–423, 2015.
- [21] S. Kamthe and M. P. Deisenroth, "Data-efficient reinforcement learning with probabilistic model predictive control," in *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2018.
- [22] J. Kabzan, L. Hewing, A. Liniger, and M. N. Zeilinger, "Learning-based model predictive control for autonomous racing," *IEEE Robotics and Automation Letters*, vol. 4, pp. 3363–3370, 2019.
- [23] S. Zimmermann, R. Poranne, and S. Coros, "Go fetch! - dynamic grasps using boston dynamics spot with external robotic arm," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*, 2021, pp. 4488–4494.
- [24] A. Nagabandi, G. Kahn, R. S. Fearing, and S. Levine, "Neural network dynamics for model-based deep reinforcement learning with model-free fine-tuning," in *2018 IEEE International Conference on Robotics and Automation (ICRA)*, 2018, pp. 7559–7566.
- [25] K. Chua, R. Calandra, R. McAllister, and S. Levine, "Deep reinforcement learning in a handful of trials using probabilistic dynamics models," in *Advances in Neural Information Processing Systems*, S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, Eds., vol. 31. Curran Associates, Inc., 2018.
- [26] A. Nagabandi, K. Konolige, S. Levine, and V. Kumar, "Deep dynamics models for learning dexterous manipulation," in *Proceedings of the Conference on Robot Learning*, ser. Proceedings of Machine Learning Research, L. P. Kaelbling, D. Kragic, and K. Sugiura, Eds., vol. 100. PMLR, 30 Oct–01 Nov 2020, pp. 1101–1112.
- [27] S. Curi, F. Berkenkamp, and A. Krause, "Efficient model-based reinforcement learning through optimistic policy search and planning," in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, Eds., 2020, pp. 14 156–14 170.
- [28] C. E. Rasmussen and C. K. I. Williams, *Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning)*. The MIT Press, 2005.
- [29] A. Chiuso and G. Pillonetto, "System identification: A machine learning perspective," *Annual Review of Control, Robotics, and Autonomous Systems*, vol. 2, no. 1, pp. 281–304, 2019.
- [30] J. SjÅllberg, H. Hjalmarsson, and L. Ljung, "Neural networks in system identification," *IFAC Proceedings Volumes*, vol. 27, no. 8, pp. 359–382, 1994.
- [31] K. Narendra and K. Parthasarathy, "Identification and control of dynamical systems using neural networks," *IEEE Transactions on Neural Networks*, vol. 1, no. 1, pp. 4–27, 1990.
- [32] J.-S. Wang and Y.-P. Chen, "A fully automated recurrent neural network for unknown dynamic system identification and control," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 53, no. 6, pp. 1363–1372, 2006.
- [33] O. Ogunmolu, X. Gu, S. Jiang, and N. Gans, "Nonlinear systems identification using deep dynamic neural networks," *arXiv preprint arXiv:1610.01439*, 2016.
- [34] J. Gonzalez and W. Yu, "Non-linear system modeling using lstm neural networks," *IFAC Conference on Modelling, Identification and Control of Nonlinear Systems MICNON*, vol. 51, no. 13, pp. 485–489, 2018.
- [35] Z. I. Botev, D. P. Kroese, R. Y. Rubinstein, and P. L'Ecuyer, "Chapter 3 - the cross-entropy method for optimization," in *Handbook of Statistics*. Elsevier, 2013, pp. 35–59.
- [36] S. Levine and P. Abbeel, "Learning neural network policies with guided policy search under unknown dynamics," in *Advances in Neural Information Processing Systems*, Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Q. Weinberger, Eds., vol. 27. Curran Associates, Inc., 2014.
- [37] R. Boney, N. Di Palo, M. Berglund, A. Ilin, J. Kannala, A. Rasmus, and H. Valpola, "Regularizing trajectory optimization with denoising autoencoders," in *Advances in Neural Information Processing Systems*, 2019.
- [38] E. Todorov and W. Li, "A generalized iterative lqg method for locally-optimal feedback control of constrained nonlinear stochastic systems," in *Proceedings of the 2005, American Control Conference, 2005.*, 2005, pp. 300–306 vol. 1.
- [39] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," *International Conference on Learning Representations*, 2014.
- [40] G. Williams, P. Drews, B. Goldfain, J. M. Rehg, and E. A. Theodorou. (2018) Information-theoretic model predictive control: Theory and applications to autonomous driving.
- [41] H. Bharadhwaj, K. Xie, and F. Shkurti, "Model-predictive control via cross-entropy and gradient-based optimization," in *Proceedings of the 2nd Conference on Learning for Dynamics and Control*, ser. Proceedings of Machine Learning Research, vol. 120. PMLR, 2020, pp. 277–286.
- [42] A. Graves, "Generating sequences with recurrent neural networks," *arXiv preprint arXiv:1308.0850*, 2013.
- [43] D. Ha and J. Schmidhuber, "Recurrent world models facilitate pol-

- icy evolution,” *Advances in neural information processing systems*, vol. 31, 2018.
- [44] D. Hafner, T. Lillicrap, I. Fischer, R. Villegas, D. Ha, H. Lee, and J. Davidson, “Learning latent dynamics for planning from pixels,” in *International conference on machine learning*. PMLR, 2019, pp. 2555–2565.
 - [45] D. Hafner, T. Lillicrap, J. Ba, and M. Norouzi, “Dream to control: Learning behaviors by latent imagination,” 2020.
 - [46] M. L. Puterman, *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons, 2014.
 - [47] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, “Empirical evaluation of gated recurrent neural networks on sequence modeling,” in *NeurIPS 2014 Workshop on Deep Learning, December 2014*, 2014.
 - [48] D. Hendrycks and K. Gimpel, “Gaussian error linear units (gelus),” *arXiv preprint arXiv:1606.08415*, 2016.
 - [49] C. E. García, D. M. Prett, and M. Morari, “Model predictive control: Theory and practice - a survey,” *Automatica*, pp. 335–348, 1989.
 - [50] S. Gu, T. Lillicrap, I. Sutskever, and S. Levine, “Continuous deep q-learning with model-based acceleration,” in *International conference on machine learning*. PMLR, 2016, pp. 2829–2838.
 - [51] M. Janner, J. Fu, M. Zhang, and S. Levine, “When to trust your model: Model-based policy optimization,” *Advances in Neural Information Processing Systems*, vol. 32, 2019.
 - [52] A. Rajeswaran, S. Ghotra, B. Ravindran, and S. Levine, “Epopt: Learning robust neural network policies using model ensembles,” *arXiv preprint arXiv:1610.01283*, 2016.
 - [53] T. Kurutach, I. Clavera, Y. Duan, A. Tamar, and P. Abbeel, “Model-ensemble trust-region policy optimization,” *arXiv preprint arXiv:1802.10592*, 2018.
 - [54] I. Clavera, J. Rothfuss, J. Schulman, Y. Fujita, T. Asfour, and P. Abbeel, “Model-based reinforcement learning via meta-policy optimization,” in *Proceedings of The 2nd Conference on Robot Learning*, ser. Proceedings of Machine Learning Research, vol. 87, 29–31 Oct 2018, pp. 617–629.
 - [55] A. R. Geist and S. Trimpe, “Structured learning of rigid-body dynamics: A survey and unified view from a robotics perspective,” *GAMM-Mitteilungen*, vol. 44, no. 2, 2021.