Energy optimization system of Electric vehicles — A case study for applications of deep reinforcement learning in the real world

A Preprint

Binjian Xin[*]
Shanghai
binjian.xin@gmail.com

July 15, 2024

Abstract

We present the application of deep reinforcement learning in the optimization of energy efficiency of driving an electric vehicle. The optimziation is modeled as Markov Decision Process and the design choice of the state, action and reward is provided. We set up a flexible data pipeline for capturing, processing, storing and sampling time sequences which enables both online and offline reinforcement learning. The training and inferring setup of the agent are provided with respect to the practical technicalities. Our system is scalable by leveraging the cloud infrucstructure therefore is capable of handling a massive fleet in scope of the application asynchronously in training and infering deployment. We observe the manifest increase of the energy efficiency on the real road condition within a short time range of online training. We assume the driver behavior is stationary and montior the driving style during training and give an analysis of the interaction between the human driver and the agent. We verify the transferability and multimodality of the trained model on different road conditions and give a review of safety, sample efficiency, data synchronicity of the system concerning the applications of deep reinforcement learning in the real world. Solutions are provided for solving partial observability, long-term stragtegy and efficient training of ragged episode length by leveraging sequential models. Deployment of reward-driven learning methods can promote the industry to leverage abundant available online or offline domain data and interfaces to achieve continuous and dynamic optimization in many complex industrial processes which require the large capacity of deep neural networks.

Keywords deep reinforcement learning · electric vehicle · time series · dataflow

1 Introduction

Drivers with diverse driving experience tend to have quite different fuel or electricity consumption on the same vehicle and the same driving route. The general common sense is that the driving styles, i.e., how drivers operate the vehicle through acceleration and brake, have an impact on the vehicle energy consumption. We would expect there exists an experienced driver with a specific driving style can handle various road conditions to achieve the optimal energy efficiency. This leads us to the assumption that if we could apply an agent which observe the driving dynamics and adjusts the driving operation, we could reduce the energy consumption. If we choose the right observation and action, we could optimize the energy consumption online or offline for driving the vehicle with a general paradigm based on learning methods so that we can leverage a large mount of easily available driving data.

Nevertheless, the application of deep reinforcement learning in the real world is generally difficult (Irpan, 2018). The main reasons include the sample inefficiency, reward shaping challenge, local optima, overfitting to rare patterns, unstable training, hyperparameter sensitivity. Most known successful applications of deep reinforcement learning

[*]

can be found in games (Mnih et al., 2013), (Silver et al., 2016), (Brown & Sandholm, 2017), (OpenAI et al., 2019), (Bakhtin et al., 2022), where the system dynamics are deterministic or closely deterministic. If the system dynamics is deterministic or very well known and a good simulation is available, then an enormous amount of samples can be generated for training. Furthermore for games specifically, self-play can be used for learning (Silver et al., 2018).

In recent years, robotics is the domain with growing successful deployment of deep reinforcement learning in the real world, particularly in manipulation, grasping and legged locomotion. Diverse techniques are used. Usually simulation is leveraged to have a good model for transfer learning in the real world. In particular privileged information of the system dynamics is used to have dedicated modules to explicitly deal with environment changes (Kumar et al., 2021), to learn the alignment of proprioception with exteroception (Miki et al., 2022), or to just use a specialized initialization strategy, randomized physical parameters and intentional delay in simulation for efficient exploration (Song et al., 2023). Furthermore, (Hoeller et al., 2024) focuses on limited skill sets to have a modular model structure to increase the learning efficiency. (Smith et al., 2023) constrains the policy in a principled way to the familiar system dynamics to increase the learning efficiency in the real world. (Wu et al., 2022) leverages the world model and efficient sequential latent encoding of the system dynamics. Overall robotics learning tends to use more imitation learning and utilize offline data, as it's still hard to search for the generic reinforcement learning and extra pre-training and supervised fine-tuning is cheaper (Irpan, 2024).

In the automotive industry, one of the most evident applicaion would be autonomous driving. But there's no available public work of autonomous driving in the real world that's based on deep reinforcement learning framework in a principled way. More attention is paid to leveraging simulation to generate sufficient training samples or using prior knowledge in pretrained foundational models and scaling up.

In general, there's no generic end-to-end deep reinforcement learning method for real world applications, but a paradigm to combine deep reinforcement learning with domain specific techniques to alleviate the aforementioned challenges. As each problem in the real world have its specific prior information and exploitable inductive biases, it's only natural to exploit them in an engineering way.

This paper gives an example that progress in real world applications can be achieved with deep reinforcement learning. While research concerns might be alleviated with design choices, the practical deployment should still comply with the basic requirements of the theories stringently without loss of validity. Eventually in long term, issues of multimodality and out-of-distribution matter in complex and very long time horizon. On the practical side, deployment of reward-driven learning methods can promote the industry to leverage abundant available online or offline domain data and interfaces to achieve continuous and dynamic optimization in many complex industrial processes which require the large capacity of deep neural networks and have the potential to help reshaping the industry into the data-driven paradigm.

1.1 Related work

Electric vehicles (EVs) have been growing in popularity in the automotive industry with sales increasing globally. The share of electric cars in total sales has increased from negligibly less than 0.1% in 2010 to 14% in 2022 (Carlier, 2023). The deployment of EVs is largely due to the rising fuel economy standards and the required reduction in greenhouse gas emissions, which leads to the increasing complexity of powertrains in the form of additional actuators and control systems. The energy efficiency of EVs has been the focus of powertrain electrification. In (Egan et al., 2023) an overview of the reinforcement learning methods applied in powertrain controllers in hybrid electric vehicles (HEVs), fuel cell electric vehicles (FCEVs), plug-in hybrid electric vehicles (PHEVs) is given. They reviewed the state, action space and the reward function of their design choices.

For example, (P. Wang & Northrop, 2020) exploits the physical models of extended range electric electric vehciles (EREVs) in the investigation so that the chosen state consists of the traveled distance and time, fuel use and the GPS coordinates of the vehicle and in particular the "energy compensated expected trip distance" which by their definition is the expected total trip distance multiplied by a ratio of trip energy intensity and the expected one. The chosen action is the range of the constructed state. The reward function is designed as weighted sum of the penaltes of internal combustion engine (ICE) operation, low state of charge (SOC), change of the constructed state and a specific terminal state. In (Hou et al., 2022) the state is defined as the power demand and SOC, the action is the fuel cell power, while the reward is a weighted sum of the change rate of the instantaneous hydrogen consumption, a qudratic poloynomial of the battery power with cooefficients containing open-circuit voltage, cell capacity and battery internal resistance in order to reduce unnecessary energy storation while keep the action feasiblity of the electric motor and the ICE. Similarly, (Hu et al., 2018) selects for HEVs the total required torque and SOC as state and the output torque of ICE as action which is equivalent to the power-split of HEV between the electric motor and ICE, while the reward is defined as the reciprocal of instantaneous fuel consumption of the ICE conditioned on the range of itself and the instantaneous SOC.

It can be seen that these methods exploit the domain specific knowledge for each of the vehicle model under investigation to select the state and action. The constructed and estimated expectation state signals which are not directly measured contain further biases and noise. Besides, the measureable state signals in those applications require specific sensors or processing and are thus expensive to acquire. Furthermore, it should be noted that even the SOC cannot be directly measured in real time while the vehicle is driving so it's estimated from the measured voltage and a lookup table acquired by calibration. The selected actions serve their specific objective domains and therefore cannot be generalized to other vehicle models. The reward shaping in those applications are never the energy consumption directly, since the objective is to optimize the co-operation of eletric motor and ICE for HEV. It's usually a heuristic mixture of electricity and fuel consumption with constraints which are required to regularized the system behavior but in general detrimental to the optimization performance. The weights in the reward brings extra hyperparameters which are sensitive to the system performance and difficult to analyze.

The review of the applications of deep reinforcement learning in EVs makes us wonder whether we can use the energy consumption directly as reward, while taking observation directly from the vehicle dynamics and driver behavior. The objective would not be the optimization of the co-operation of the mixed powertrain components, but the optimization of overall powertrain dynamic behavior conditioned on the road environment and driver operation. We'd call such a system "Energy Optimization System of Electric Vehcicle " (EOSEV). Such a system would be generic and applicable for any type of EVs since all the required observations, actions and rewards would be available and easily measureable. The challenge is to learn the complex system dynamics of driving in the real world condition.

It shoud be noted that the method would applicable for regular ICE vehicles as well. We experiment on a BEV, since it has a simplistic yet powerful electric powertrain. The electric motor has a much faster torque response than ICE, therefore the measurements are more precise and have better real-time performance which is crucial to guarantee the causality of actions, states and rewards.

Empirically, the optimization space of the driving operation comes from the way to adaptively adjust the vehicle speed by acceleration or deceleration to reach the target position with the target speed. Usually a smooth speed transition is better than abrupt acceleration or braking which causes large currents in the electric motor and results in more electricity consumption. One particular feature of EVs is regenerative braking system (RBS) (Wikipedia contributors, 2024), which lets the vehicle recapture energy from momentum by running the electric motor in reverse to recharge the battery with a negative torque request. The regenrative braking system improves the energy efficiency and is now a standard part of many electrified vehicle including HEVs and BEVs. The parameters of the RBS are contained in the electironic control unit (ECU) of the vehicle powertrain controller which is called vehicle control units (VCU). Usually the RBS doesn't depend on road conditions or driver operations and is implicitly static. However, if we choose the action so that it can impact the RBS, under the EOSEV the RBS can be exploited by the agent like the other acceleration or braking strategies. We observe in the experiment that even without explicitly modeling the RBS behaviour in the system, the RBS is actively and legitmately exploited by the agent to reduce the energy consumption without any explicit built-in rules or modeling.

2 Preliminary

In order to increase the energy efficiency while driving, we consider the following structure in Fig.1. Without the agent and its connection to the input (observation) and the output (action), the parameters of the powertrain controller are kept static, the depicted process is a regular powertrain control with the driver in the loop. The driver controls the vehicle speed through acceleration and brake pedal while observing the road condition.

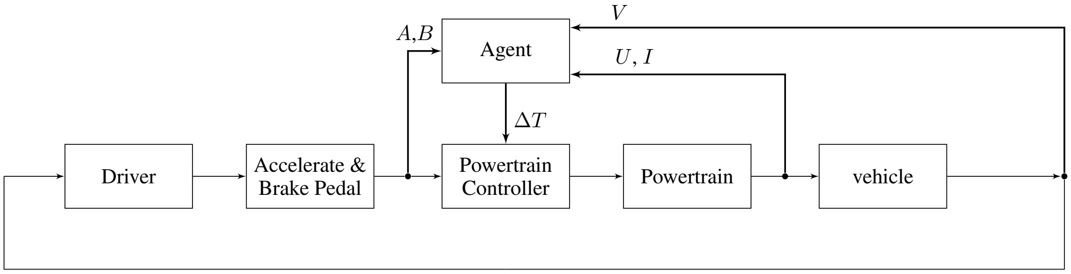

Figure 1: Add an agent to the conventional EV powertrain control loop.

When the agent is connected to the regular system as depicted, it will get the vehicle speed V, the driver's operation on the acceleration pedal A and the brake pedal B from the on-board sensors as its observation. (V, A, B) is defined as to have an optimal policy. To deal with this partial observability, an RDPG agent (Heess et al., 2015) is implemented with the LSTM networks (Hochreiter & Schmidhuber, 1997). With truncated BPTT (Sutskever, 2013) and the stateful feature of LSTM cell (Chollet et al., 2015), we can handle arbitrary ragged episode lengths and do inference efficiently with long and short term policies.

The result of transferablitiy in Sec.3.4.1 with frozen model indicates that while the training must be episodic in order to have meaningful reward signals, the inference-only mode doesn't need to be episodic. A trained frozen model can be deployed for non-episodic situations in inference-only mode in a wide range of applicaitions without the episodic constraint.

The data in Tab.1 has a low density and is light-weight for collection and storage. In order to leverage the large volume of offline data with ongoing training, an offline reinforcement learning is implemented with IDQL (Hansen-Estruch et al., 2023). With the offline data, training can occur in the cloud with uploaded observation data from vehicles. The updated local model can be dispatched onto the vehicle through OTA communication. In this framework, the training and inference can be done flexibly either locally or in the cloud, see Fig.17. Training and inference locally are fast, have better realtime performance and signal quality, but need extra compute resources on the vehicle. The DDPG agent has a size of less than 1MB as exported tflite models which can be easily stored on the embedded system, while it's difficult to accommodate the RDPG agent with its closely 100 MB. On the cloud it's more scalable, doesn't need much compute locally but OTA communication with less realtime performance, larger signal latency and lower signal quality. In large scale deployment, training and inferrence on the cloud also provide an opportunity for utilization of parallel training of asynchronous poclicy gradient methods (Mnih et al., 2016) or federated learning (Konečný et al., 2015).

Diffusion models have been studied actively in recent offine RL researches, in particular in behavior cloning for policy learning (Janner et al., 2022) (Z. Wang et al., 2022) and (Hansen-Estruch et al., 2023) (Psenka et al., 2023). Besides utilized by the offline reinforcement learning, the diffusion models can be used to recover complex and multimodal behavior data. With the diffusion model to fit the behavior policy, the drift of the driving style in Fig.11 is expected to be better tracked by the implicit actor with policy extraction (Hansen-Estruch et al., 2023). However, the training and sampling in diffusion models require extra noising and denoising time steps, which reduce its realtime performance and constrains their application. All these improvements require more compute resources and need further verification with large scale deployment.

5 Conclusion

We present the application of deep reinforcement learning in the optimization of energy efficiency of driving an electric vehicle. We demonstrate that the design choice of the state, reward and action is crucial. If the reward is rich and dense, the application in real world should be sample efficient. If the observation data are of low density and generally accessible like vehicle speed, acceleration and braking pedal opening but contain complex behaviors or system models, we could harvest a large volume of offline data leverage the offline reinforcement learning to improve the optimization process. As reward-driven learning methods are machine learning method which requires no supervision or labeling, but learns from large amount of raw data, they are appealing to the industrial applications. Our experiment is still far from being scaled up. With deployment with large amount of vehicles in the future, we hope to track complex system dynamics, road scenarios and driving behaviors by more efficient RL methods.

The purpose of our work is twofold. On the engineering side, we hope to advocate in the industry to leverage abundant available online or offline domain data and interfaces to achieve continuous and dynamic optimization in complex industrial processes which require the large capacity of deep neural networks and have the potential to help reshaping the industry into the data-driven paradigm.

For research, we intend to provide an application of deep reinforcement learning in the real world that provides interesting and challenging optimization goals. Unlike games or robotics, we cannot always leverage simulation in the real world, but we can have abundant data and achieve sample efficiency with prudent design choice of state, action and reward and by careful implementation which maintains the signal causality and avoids adding noise to state and reward signal. Furthermore, system and task persistency are sometimes guaranteed. More importantly, with large scale deployment of data-driven method and the improvement of the industrial process optimization, a virtuous cycle would contribute to providing abundant data, applying new research results and finding new interesting research challenges.

- multiple models: with DDPG and RDPG for time sequences with arbitrary length;
- offline reinforcement learning with "Implict Diffusion Q-Learning" (IDQ);
- compatible data pipelines to both ETL and ML dataflow;
- multiple data sources (local CAN or remote cloud object storage);
- stateful time sequence processing with sequential model;
- support of both NoSQL database, local and cloud data storage.

The diagram shows the basic architecture. The dataflow of the pipelines

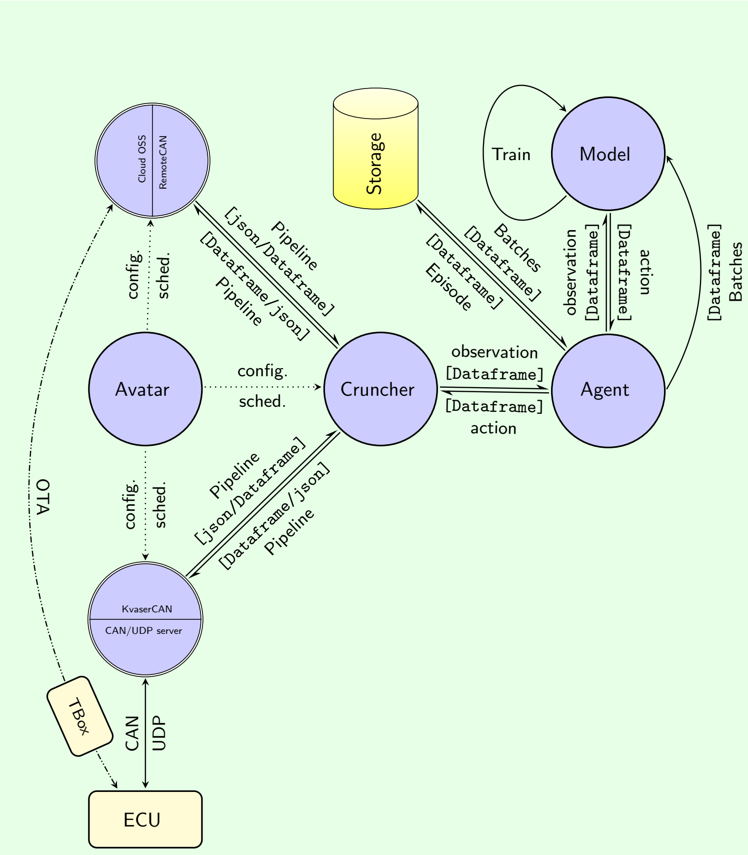

Figure 17: Software Overview

References

Bakhtin, A., Brown, N., Dinan, E., Farina, G., Flaherty, C., Fried, D., Goff, A., Gray, J., Hu, H., Jacob, A. P., Komeili, M., Konath, K., Kwon, M., Lerer, A., Lewis, M., Miller, A. H., Mitts, S., Renduchintala, A., Roller, S., . . . Zijlstra, M. (2022). Human-level play in the game of diplomacy by combining language models with strategic reasoning. *Science*, *378*(6624), 1067–1074. https://doi.org/10.1126/science.ade9097 (cit. on p. 2).

Brown, N., & Sandholm, T. (2017). Libratus: The superhuman AI for no-limit poker. *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI 2017, Melbourne, Australia, August 19-25, 2017*, 5226–5228. https://doi.org/10.24963/IJCAI.2017/772 (cit. on p. 2).

Carlier, M. (2023, May). Global market share of electric vehicles within passenger car sales between 2010 and 2022. https://www.statista.com/statistics/1371599/global-ev-market-share/ (cit. on p. 2).

Chollet, F., et al. (2015). Keras. (Cit. on p. 12).

Egan, D., Zhu, Q., & Prucka, R. (2023). A review of reinforcement learning-based powertrain controllers: Effects of agent selection for mixed-continuity control and reward formulation. *Energies*, *16*(8), 3450. https://doi.org/10.3390/en16083450 (cit. on p. 2).

Fujimoto, S., van Hoof, H., & Meger, D. (2018). *Addressing Function Approximation Error in Actor-Critic Methods*. arXiv: 1802.09477v3 [cs.AI]. (Cit. on p. 7).

Hansen-Estruch, P., Kostrikov, I., Janner, M., Kuba, J. G., & Levine, S. (2023). *IDQL: Implicit Q-Learning as an Actor-Critic Method with Diffusion Policies*. arXiv: 2304.10573v2 [cs.LG]. (Cit. on p. 12).

Heess, N., Hunt, J. J., Lillicrap, T. P., & Silver, D. (2015). *Memory-based control with recurrent neural networks*. arXiv: 1512.04455v1 [cs.LG]. (Cit. on pp. 7, 12).

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, *9*(8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735 (cit. on p. 12).

Hoeller, D., Rudin, N., Sako, D., & Hutter, M. (2024). Anymal parkour: Learning agile navigation for quadrupedal robots. *Science Robotics*, *9*(88). https://doi.org/10.1126/scirobotics.adi7566 (cit. on p. 2).

Hou, S., Liu, X., Yin, H., & Gao, J. (2022). Reinforcement learning-based energy optimization for a fuel cell electric vehicle. *2022 4th International Conference on Smart Power and Internet Energy Systems (SPIES)*. https://doi.org/10.1109/spies55999.2022.10082644 (cit. on p. 2).

Hu, Y., Li, W., Xu, K., Zahid, T., Qin, F., & Li, C. (2018). Energy management strategy for a hybrid electric vehicle based on deep reinforcement learning. *Applied Sciences*, *8*(2), 187. https://doi.org/10.3390/app8020187 (cit. on p. 2).

Ibarz, J., Tan, J., Finn, C., Kalakrishnan, M., Pastor, P., & Levine, S. (2021). How to train your robot with deep reinforcement learning: Lessons we have learned. *Int. J. Robotics Res.*, *40*(4-5). https://doi.org/10.1177/0278364920987859 (cit. on p. 11).

Irpan, A. (2018, February). Deep reinforcement learning doesn't work yet. https://www.alexirpan.com/2018/02/14/rl-hard.html (cit. on p. 1).

Irpan, A. (2024, January). My ai timelines have sped up (again). https://www.alexirpan.com/2024/01/10/ai-timelines-2024.html (cit. on p. 2).

Janner, M., Du, Y., Tenenbaum, J. B., & Levine, S. (2022). *Planning with Diffusion for Flexible Behavior Synthesis*. arXiv: 2205.09991v2 [cs.LG]. (Cit. on p. 12).

Konečný, J., McMahan, B., & Ramage, D. (2015). *Federated Optimization:Distributed Optimization Beyond the Datacenter*. arXiv: 1511.03575v1 [cs.LG]. (Cit. on p. 12).

Kumar, A., Fu, Z., Pathak, D., & Malik, J. (2021). *RMA: Rapid Motor Adaptation for Legged Robots*. arXiv: 2107.04034v1 [cs.LG]. (Cit. on p. 2).

Lillicrap, T. P., Hunt, J. J., Pritzel, A., Heess, N., Erez, T., Tassa, Y., Silver, D., & Wierstra, D. (2015). *Continuous control with deep reinforcement learning*. arXiv: 1509.02971v6 [cs.LG]. (Cit. on p. 4).

Miki, T., Lee, J., Hwangbo, J., Wellhausen, L., Koltun, V., & Hutter, M. (2022). Learning robust perceptive locomotion for quadrupedal robots in the wild. *Science Robotics*, *7*(62). https://doi.org/10.1126/scirobotics.abk2822 (cit. on p. 2).

Mnih, V., Badia, A. P., Mirza, M., Graves, A., Lillicrap, T. P., Harley, T., Silver, D., & Kavukcuoglu, K. (2016). *Asynchronous Methods for Deep Reinforcement Learning*. arXiv: 1602.01783v2 [cs.LG]. (Cit. on p. 12).

Mnih, V., Kavukcuoglu, K., Silver, D., Graves, A., Antonoglou, I., Wierstra, D., & Riedmiller, M. (2013). *Playing Atari with Deep Reinforcement Learning*. arXiv: 1312.5602v1 [cs.LG]. (Cit. on p. 2).

OpenAI, : Berner, C., Brockman, G., Chan, B., Cheung, V., Dębiak, P., Dennison, C., Farhi, D., Fischer, Q., Hashme, S., Hesse, C., Józefowicz, R., Gray, S., Olsson, C., Pachocki, J., Petrov, M., d. O. Pinto, H. P., Raiman, J., . . . Zhang, S. (2019). *Dota 2 with Large Scale Deep Reinforcement Learning*. arXiv: 1912.06680v1 [cs.LG]. (Cit. on p. 2).

Psenka, M., Escontrela, A., Abbeel, P., & Ma, Y. (2023). *Learning a Diffusion Model Policy from Rewards via Q-Score Matching*. arXiv: 2312.11752v2 [cs.LG]. (Cit. on p. 12).

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., van den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., Dieleman, S., Grewe, D., Nham, J., Kalchbrenner, N., Sutskever, I., Lillicrap, T. P., Leach, M., Kavukcuoglu, K., Graepel, T., & Hassabis, D. (2016). Mastering the game of go with deep neural networks and tree search. *Nat.*, *529*(7587), 484–489. https://doi.org/10.1038/NATURE16961 (cit. on p. 2).

Silver, D., Hubert, T., Schrittwieser, J., Antonoglou, I., Lai, M., Guez, A., Lanctot, M., Sifre, L., Kumaran, D., Graepel, T., Lillicrap, T., Simonyan, K., & Hassabis, D. (2018). A general reinforcement learning algorithm that masters chess, shogi, and go through self-play. *Science*, *362*(6419), 1140–1144. https://doi.org/10.1126/science.aar6404 (cit. on p. 2).

Singh, H. (2020). Deep deterministic policy gradient (ddpg). https://keras.io/examples/rl/ddpg_pendulum/ (cit. on p. 4).

Smith, L., Cao, Y., & Levine, S. (2023). *Grow Your Limits: Continuous Improvement with Real-World RL for Robotic Locomotion*. arXiv: 2310.17634v1 [cs.RO]. (Cit. on p. 2).

Song, Y., Romero, A., Müller, M., Koltun, V., & Scaramuzza, D. (2023). Reaching the limit in autonomous racing: Optimal control versus reinforcement learning. *Science Robotics*, *8*(82). https://doi.org/10.1126/scirobotics.adg1462 (cit. on p. 2).

Sutskever, I. (2013). *Training recurrent neural networks*. University of Toronto Toronto, ON, Canada. (Cit. on p. 12).

Wang, P., & Northrop, W. (2020). Data-driven framework for fuel efficiency improvement in extended range electric vehicle used in package delivery applications. *SAE Technical Paper Series*. https://doi.org/10.4271/2020-01-0589 (cit. on p. 2).

Wang, Z., Hunt, J. J., & Zhou, M. (2022). *Diffusion Policies as an Expressive Policy Class for Offline Reinforcement Learning*. arXiv: 2208.06193v3 [cs.LG]. (Cit. on p. 12).

Wikipedia contributors. (2024). Regenerative braking — Wikipedia, the free encyclopedia [[Online; accessed 8-July-2024]]. (Cit. on p. 3).

Wu, P., Escontrela, A., Hafner, D., Goldberg, K., & Abbeel, P. (2022). *DayDreamer: World Models for Physical Robot Learning*. arXiv: 2206.14176v1 [cs.RO]. (Cit. on p. 2).

Xin, B. (2024). Tspace. https://github.com/Binjian/tspace/ (cit. on p. 13).