2024 IEEE International Conference on Robotics and Automation (ICRA)
May 13-17, 2024. Yokohama, Japan

Generalizing Cooperative Eco-driving via Multi-residual Task Learning

Vindula Jayawardana[1*], Sirui Li[1], Cathy Wu[1], Yashar Farid[2], and Kentaro Oguchi[2]

2024 IEEE International Conference on Robotics and Automation (ICRA) | 979-8-3503-8457-4/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICRA57147.2024.10610586

***Abstract*— Conventional control, such as model-based control, is commonly utilized in autonomous driving due to its efficiency and reliability. However, real-world autonomous driving contends with a multitude of diverse traffic scenarios that are challenging for these planning algorithms. Model-free Deep Reinforcement Learning (DRL) presents a promising avenue in this direction, but learning DRL control policies that generalize to multiple traffic scenarios is still a challenge. To address this, we introduce Multi-residual Task Learning (MRTL), a generic learning framework based on multi-task learning that, for a set of task scenarios, decomposes the control into nominal components that are effectively solved by conventional control methods and residual terms which are solved using learning. We employ MRTL for fleet-level emission reduction in mixed traffic using autonomous vehicles as a means of system control. By analyzing the performance of MRTL across nearly 600 signalized intersections and 1200 traffic scenarios, we demonstrate that it emerges as a promising approach to synergize the strengths of DRL and conventional methods in generalizable control.**

I. INTRODUCTION

Autonomous vehicles (AVs) are surging in popularity due to rapid technological advancements. Lately, AVs also have been used as Lagrangian actuators for system-level traffic control. Lagrangian control describes microscopic level traffic control techniques based on mobile actuators (e.g., vehicles) rather than fixed-location actuators (e.g., traffic signals). Therefore, it involves planning a fleet of AVs to accomplish a given objective at the system level, which involves both AVs and human drivers. These objectives include mitigating traffic congestion [1], curbing emissions [2], and promoting smoother traffic flows [3].

In particular, recent work explores the use of AVs as Lagrangian actuators for fleet-wide emission reduction [4]. As illustrated in Figure 1, the goal is to reduce emissions of the fleet by controlling and coordinating AVs and exerting control over human-driven vehicles. Methods from heuristics [5], to model-based methods [6], to model-free [2] methods, have been used in tackling this challenge.

Although model-based control strategies like model-predictive control are frequently employed [4], they rely on the assumption of having a precise vehicle dynamics model. Yet, devising such a model, including the multitude of factors influencing driving dynamics, is challenging. In the absence of them, these methods often fall short in terms of adapting to various traffic scenarios. The deployment of simplified models tends to result in poor generalization and, in some cases, even increasing emissions levels.

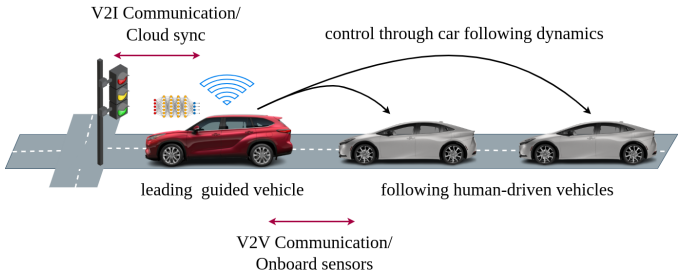


Fig. 1: In a signalized intersection, AVs lead platoons of human-driven vehicles. As Lagrangian actuators, they reduce fleet emissions by controlling their own acceleration and shepherding the human drivers through car following dynamics.

On the other hand, DRL operates without the need for a predefined dynamics model, hence model-free. It has the capacity to address control challenges that prove challenging for conventional methods as it specifies control objectives indirectly within a reward function rather than explicit control actions. Although limited in success, DRL has demonstrated the capacity to adapt to changes in the underlying environmental conditions [7], underscoring its potential to generalize across various traffic scenarios.

However, developing DRL algorithms for Lagrangian control in diverse traffic settings still remains a challenge. Real-world roads are complex, with complexities including vehicle interactions, varied road topologies, and external controls like traffic signals and stop signs. These complexities are often unpredictable, introducing uncertainty. Devising eco-driving planning algorithms capable of handling multiple scenarios is thus demanding, and many existing studies focus on a few scenarios [4], leading to less meaningful insights and potential overfitting in evaluations [8].

In this study, we address this challenge of algorithmic generalization for DRL across various scenarios, such as different traffic scenarios in cooperative eco-driving. We introduce *Multi-residual Task Learning (MRTL)*, a generic framework that combines DRL and conventional control methods. MRTL divides each scenario into parts solvable by conventional control and residuals solvable by DRL. The final control input for a scenario is thus the superposition of these two control inputs.

We employ MRTL in eco-driving at signalized intersections, achieving better generalization in nearly 600 signalized intersections across 1200 traffic scenarios. While many existing works focus on a few eco-driving scenarios [4], to the

[1]Massachusetts Institute of Technology, Cambridge, MA, USA {vindula, siruil, cathywu}@mit.edu
[2] Toyota InfoTech Labs and Toyota Motor North America, Mountain View, CA, USA {yashar.zeiynali.farid, kentaro.oguchi}@toyota.com
* Work done during the author's internship at Toyota InfoTech Labs.

best of our knowledge, we are the first to solve this problem on a large scale.

Our key contributions are:

- We present a generic learning framework called Multi-residual Task Learning to enable algorithmic generalization of DRL algorithms.
- We employ the MRTL framework to devise generalizable control policies for cooperative eco-driving.
- Analyzing nearly 1200 traffic scenarios across 600 signalized intersections, we demonstrate that our MRTL framework enables control generalization, outperforming baseline control methods by a large margin.

II. Related Work

The use of residuals in learning has previously been explored both in supervised learning and reinforcement learning. He et al. [9] first propose residual neural networks where they reformulate the layers of a neural network as learning residual functions. In reinforcement learning, the use of residuals takes a slightly different form. The closest to our work is residual reinforcement learning (RRL), which was first introduced simultaneously by Silver et al. [10] and Johannink et al. [11] for robot control.

Silver et al. [10] primarily look at how RRL can improve the performance of robotics manipulation tasks with various nominal policies as backbones. They demonstrate that RRL performs well when the environment is partially observable and when there is sensor noise, model misspecifications, and controller miscalibrations. Johannink et al. [11] further show that RRL can be used to deal with the sim-to-real transfer. In particular, it can be used to modify the learned controller such that it can react to the modeling inaccuracies observed in the simulations to better adapt to the real world.

While these works lay a foundation, the focus is single-agent, single-task robotic manipulations characterized by relatively simple control. In contrast, we look at multi-agent, multi-task scenarios, necessitating not only optimizing performance on individual tasks but also extending to robust and generalization control synthesis.

In autonomous driving, Zhang et al. [12] use RRL and reduce the lap time in autonomous racing. They further show the transferability of learned policies by transferring a policy from one track to another and to new tracks. However, this work primarily looks at single-agent racing, and multi-agent racing poses different challenges. Furthermore, while some transferability results have been shown, it is still limited to a few select racing tracks instead of a diverse set of scenarios. Moreover, autonomous racing and Lagrangian control have contrasting dynamics: racing involves competition, while Lagrangian control involves cooperation.

RRL is further utilized in synthesizing generalizable reinforcement learning controllers for robotics manipulations. Hao et al. [13] introduce a meta-residual policy learning method that performs in unseen peg-in-hole assembly tasks. It improves adaptation and sample efficiency but is limited to specific robotics skills and lacks task diversity. Further, it operates in low-dimensional state spaces, leaving its suitability for high-dimensional spaces uncertain. It's worth noting that our approach differs fundamentally, focusing on multi-task learning compared to their meta-learning approach. Nevertheless, this work underscores the potential of RRL in enhancing control policy generalization.

On the other hand, combining model-based methods with learning (model-free) has been a topic of interest for some time [14]. On the one hand, these methods often involve learning a dynamics model and then using the model for trajectory optimization or model predictive control [15, 16] to simulate experience [17] or to compute gradients for model-free updates [18]. In another line of work, learning is used for capturing the objectives [19] or constraints [20]. Recently, there has also been a line of work that looks at learning a corrective term to analytical physical models [21] with the purpose of performing better predictive control.

In summary, while RRL has proven effective in single-agent, single-task robotic manipulations, none of the existing studies have showcased its application to multi-agent cooperative control with the capacity for generalization across a range of scenarios, let alone in Lagrangian control. Alternatively, combining model-based and model-free methods has exhibited mixed results, and none of them adequately tackle the challenges of algorithmic generalization. In this study, we aim to bridge this gap in the field.

III. Preliminaries

A. Reinforcement Learning

In reinforcement learning, an agent learns a control policy by interacting with its environment, typically modeled as a Markov Decision Process (MDP) denoted as $M = \langle \mathcal{S}, \mathcal{A}, p, r, \rho, \gamma \rangle$. Here, $\mathcal{S}$ represents the set of states, $\mathcal{A}$ denotes the possible actions, $p(s_{t+1}|s_t, a_t)$ denotes the transition probability from the current state s_t to the next state s_{t+1} upon taking action a_t, the reward received for action a_t at state s_t is $r(s_t, a_t) \in \mathbb{R}$, a distribution over the initial states is ρ, and $\gamma \in [0, 1]$ is a discounting factor that balances immediate and future rewards.

Given the MDP, we seek to find an optimal policy $\pi^* : \mathcal{S} \rightarrow \mathcal{A}$ over the horizon H that maximizes the expected cumulative discounted reward over the MDP.

$$\pi^*(s) = \underset{\pi}{\operatorname{argmax}}\, \mathbb{E}\left[\sum_{t=0}^{H} \gamma^t r(s_t, a_t)|s_0, \pi\right] \tag{1}$$

B. Multi-task Reinforcement Learning

In multi-task reinforcement learning, we extend the single-MDP (single task) reinforcement learning in Section III-A to multiple MDPs (multiple tasks). Accordingly, our objective in finding optimal policy thus becomes,

$$\pi^*(s) = \underset{\pi}{\operatorname{argmax}}\, \mathbb{E}\left[\sum_{\tau \in \mathcal{T}} \sum_{t=0}^{H} \gamma^t r_\tau(s_t, a_t)|s_0, \pi\right] \tag{2}$$

where $\mathcal{T}$ is the set of MDPs (tasks). Also, note that what we seek in multi-task reinforcement learning is a unified policy that is performant over all MDPs (tasks).

IV. METHOD

In this section, we formalize the concept of algorithmic generalization in DRL and detail our generic Multi-Residual Task Learning framework.

A. Problem Formulation

In this work, we study the algorithmic generalization of DRL algorithms across a family of MDPs (scenarios) that originate from a single task, such as eco-driving. To formalize this exploration, our primary focus revolves around solving Contextual Markov Decision Processes (cMDPs) [22].

A cMDP expands upon the MDP discussed in Section III-A by incorporating a 'context'. Context serves as a means to parameterize the environmental variations encountered, such as changes in lane lengths at different intersections, among other factors in eco-driving. Mathematically, we denote a cMDP as $\mathcal{M} = \langle \mathcal{S}, \mathcal{A}, \mathcal{C}, p_c, r_c, \rho_c, \gamma \rangle$. Compared to MDPs, a context space $\mathcal{C}$ is introduced, and the action space A and state space S remain unchanged. The transition dynamics p_c, rewards r_c, and initial state distribution ρ_c are changed based on the context $c \in \mathcal{C}$. Essentially, a cMDP $\mathcal{M}$ defines a collection of MDPs, each differing based on the context, such that $\mathcal{M} = \{M_c\}_{c \sim \mathcal{C}}$.

Solving a given cMDP leads to solving the problem of algorithmic generalization within that task (i.e., finding a policy that performs well in the cMDP overall). The generalization objective where the goal is to find a unified policy $\pi^*(\cdot)$ that performs well on all $M_c \in \mathcal{M}$ is as follows.

$$\pi^*(s) = \underset{\pi}{\operatorname{argmax}} \mathbb{E}\left[\sum_{c \in \mathcal{C}} \sum_{t=0}^{H} \gamma^t r_c(s_t, a_t) | s_0^c, \pi\right] \quad (3)$$

The multi-task learning framework introduced in Section III-B emerges as a natural approach to tackle cMDPs. Here, the contexts themselves define the different tasks, effectively aligning with the notion that a specific context $c \in \mathcal{C}$ in Equation 3 corresponds to a task $\tau \in \mathcal{T}$ in Equation 2.

B. Cooperative Eco-driving cMDP

In cooperative eco-driving at signalized intersections, a wide array of context factors come into play, including lane lengths, speed limits, lane count, vehicle inflows, and the timings of green and red traffic signals. These factors collectively shape the contexts within the eco-driving cMDP, which encompasses a spectrum of signalized intersections (MDPs). Then, we seek a unified AV control policy that adeptly curbs emissions of the fleet across these signalized intersections.

MDPs within a cMDP can manifest in both single-agent and multi-agent configurations. However, cooperative eco-driving adopts multi-agent control as coordination between AVs to reduce emissions is required. This characteristic amplifies the complexity of solving eco-driving cMDP, necessitating the implicit modeling of vehicle interactions and addressing the challenges posed by partial observability.

The overall objective of the cooperative eco-driving at signalized intersections is to minimize the emissions of a fleet of vehicles (both AVs and human-driven vehicles) while having a minimal impact on travel time across all signalized intersections. Given an instantaneous emission model $E(\cdot)$ that measures vehicular emission, we seek an AV control policy such that,

$$\pi^* = \underset{\pi}{\operatorname{argmin}} \mathbb{E}\left[\sum_{c \in \mathcal{C}} \sum_{i=1}^{n_c} \int_0^{T_i} E\left(a_i(t), v_i(t)\right) dt + T_i\right] \quad (4)$$

Here, n_c represents the total number of indexes of both AV and human-driven vehicles in intersection defined by c, T_i denotes the travel time of vehicle i, $v_i(t)$, and $a_i(t)$ denote the speed and acceleration of vehicle i at time t and $\mathcal{C}$ denote the contexts defining a set of signalized intersections.

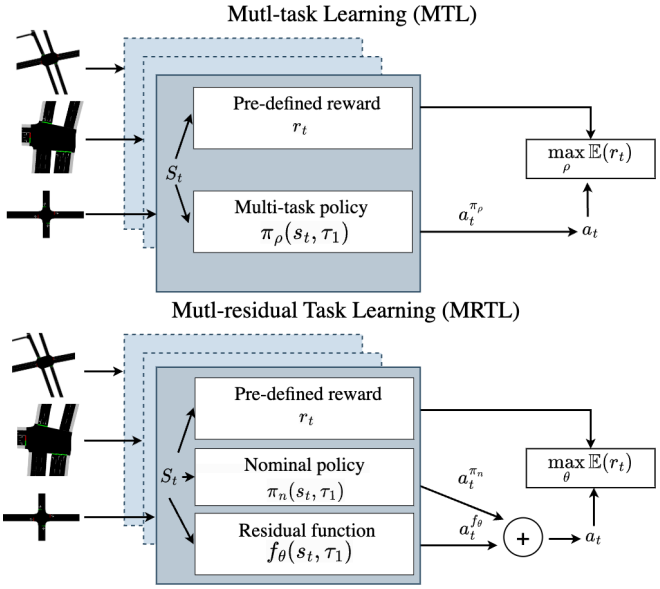


Fig. 2: Multi-task learning trains a unified policy directly with environments (intersections) sampled from a distribution of environments (top figure). Multi-residual task learning building on multi-task learning decomposes the cMDP into parts solved by a nominal policy and residual parts solved by DRL, as shown in the bottom figure.

C. Multi-residual Task Learning

While multi-task reinforcement learning can be used for solving cMDPs, it struggles when multiple MDPs are combined within one learning framework. Simultaneous training can lead to competition among MDPs for the learning agent's limited capacity, making it hard to balance MDP-specific and shared knowledge. Moreover, varying dynamics across MDPs challenge robust adaptation and generalization. Catastrophic interference risk, where new learning disrupts prior performance, hinders effectiveness further.

In addressing these issues, we propose a generic learning framework designed to enhance the algorithmic generalization of DRL algorithms, ultimately enabling solving cMDPs. We introduce Multi-Residual Task Learning, a unified learning approach that harnesses the synergy between multi-task learning and residual reinforcement learning [10, 11] (Figure 2). At its core, MRTL operates by augmenting any given

nominal policy, which exhibits sub-optimal and varying performance in each MDP in a cMDP, by learning residuals on top of it. These residuals serve as corrective measures to address suboptimalities within the nominal policy.

Consider eco-driving at signalized intersections. An AV's overall emission-reduction reward, r, can be split into r_a and r_b. r_a rewards AV gliding at red signals, a known approach for emission reduction [4], and can be achieved through a straightforward model-based controller [5]. Conversely, r_b rewards adaptive gliding behaviors to environmental changes (e.g., other vehicles lane changing), a challenge for model-based controllers due to system complexity. Training a DRL policy for r_b is feasible, allowing the nominal policy to fix r_a while learning targets the residual r_b.

To put this formally, MRTL is concerned with augmenting a given nominal policy $\pi_n(s,c) : \mathcal{S} \times \mathcal{C} \to \mathcal{A}$ by learning residuals on top of it. In particular, we aim to learn the MRTL policy $\pi(s,c) : \mathcal{S} \times \mathcal{C} \to \mathcal{A}$ by learning a residual function $f_\theta(s,c) : \mathcal{S} \times \mathcal{C} \to \mathcal{A}$ on top of a given nominal policy $\pi_n(s,c) : \mathcal{S} \times \mathcal{C} \to \mathcal{A}$ such that,

$$\pi(s,c) = \pi_n(s,c) + f_\theta(s,c)$$

where $s \in \mathcal{S}$ and $c \in \mathcal{C}$. The gradient of the π does not depend on the π_n. This enables flexibility with nominal policy choice. The effectiveness of π_n can vary among different MDPs within a cMDP. Hence, the role of the residual function f_θ in each MDP depends on MDP characteristics and nominal policy performance in that MDP. In some MDPs, π_n acts as a starting point for better exploration for the residual function. In others, it can be nearly optimal, requiring fewer improvements by the residual function.

V. MRTL for Cooperative Eco-driving

In this section, we discuss the application of the MRTL framework on eco-driving at signalized intersections. We procedurally generate a synthetic dataset with nearly 600 MDPs, which represent incoming approaches at signalized intersections. We simplify the eco-driving task to focus on these incoming approaches since traffic signals coordinate conflicting approaches [2].

Approaches are described by six features with diverse ranges: lane length (75-400 m), vehicle inflow (675-900 veh/hour), speed limit (10-15 m/s), lane count (1-3), and green and red signal phase times (25-30s). These features define the context space for the eco-driving cMDP, and each environment is a realization of these features.

A. Nominal Policy

As the nominal policy, we design a model-based heuristic controller inspired by the GLOSA algorithm for eco-driving [5]. While our nominal policy doesn't perform real-time optimizations, its low computational demands and pre-deployment verification appeal to practical applications. We detail the nominal policy in Algorithm 1.

The nominal policy operates on a simple set of criteria aimed at reducing idling and thereby reducing emissions [4]. First, it checks if the ego-vehicle can pass the intersection at its current speed; if yes, it maintains that speed (lines 5 and 6). If the time remaining to reach the intersection is greater than the time until the traffic light turns green, the policy initiates a gliding maneuver to arrive precisely when the light changes (lines 7, 8, and 9). If neither condition applies, it defaults to natural driving behavior, following the IDM car-following model [23] (lines 10 and 11).

Algorithm 1 Nominal policy π_n for eco-driving

1: **procedure** GLIDE OR KEEP SPEED(ego-vehicle speed $v(t)$, ego-vehicle distance to intersection $d(t)$, traffic signal timing plan T and green light duration T_g)
2: Calculate time to intersection $T_I \leftarrow \frac{d(t)}{v(t)}$
3: Calculate time to green light T_G from T
4: Calculate time to end green light $T_E \leftarrow T_G + T_g$
5: **if** $T_G \leq T_I \leq T_E$ **then**
6: Target speed $v_{target} \leftarrow v(t)$
7: **else if** $T_G \geq T_I$ **then**
8: Calculate target speed based on gliding principle
9: $v_{target} \leftarrow \frac{d(t)}{T_G}$
10: **else**
11: Target speed $v_{target} \leftarrow v_{IDM}$
12: **return** v_{target}
13: **end procedure**

1) What makes the nominal policy suboptimal?: The nominal policy has inherent limitations due to the simplifications made for real-time feasibility. First, it focuses solely on the ego vehicle's dynamics, ignoring nearby vehicles, which compromises its effectiveness, especially in scenarios with human-driven vehicles, lane changes, or intersection queues.

Furthermore, the policy doesn't account for appropriate vehicle behaviors during unprotected left turns, leading to suboptimal control results and undermining intersection queues and lane changes. Dedicated left turn lanes and traffic signal phases introduce unmodeled lane changes, rendering the policy ineffective when these conditions arise, impacting its emission reduction objective.

B. MRTL Implementation Details

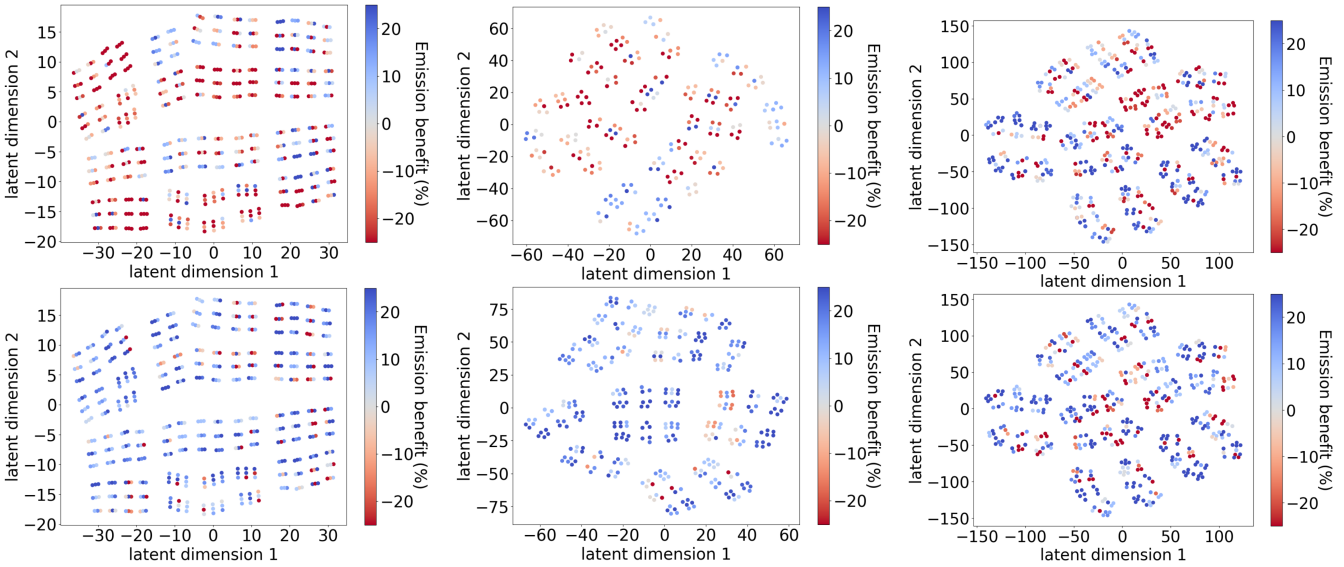
We employ centralized training and decentralized execution for training MRTL policies. We use actor-critic architecture with three hidden layers, each with 128 neurons in both the actor and critic, with a learning rate of 0.005. 1200 traffic scenarios are modeled in 600 intersections and two AV penetration levels (20% and 100%) in the SUMO simulator. PPO algorithm [24] is used as the DRL algorithm with 12 workers running for 400 iterations. We use a neural surrogate emission model [25] replicating MOVES [26] as our emission model to measure vehicular emissions.

Each MDP in eco-driving cMDP is defined as follows.

- **States**: speed, position of the ego-vehicle, leading and following vehicles in the same lane and adjacent lanes, the current traffic signal phase with remaining time, and context features including lane length, speed limit, green and red phase times, and approach phase count.

Method	20% penetration			100% penetration		
	Emission ↓	Speed ↑	Throughput ↑	Emission ↓	Speed ↑	Throughput ↑
Multi-task learning	64.08%	-27.70%	-34.70%	95.86%	-30.87%	-68.11%
Nominal policy	13.13%	-21.11%	-30.07%	-25.09%	11.72%	-3.90%
Multi-residual task learning (Ours)	**-13.95%**	**12.35%**	**7.95%**	**-29.09%**	**17.10%**	**5.72%**

TABLE I: Performance comparison of MRTL with other baselines for eco-driving at 20% and 100% AV penetration. The percentages are calculated compared to the naturalistic human-like driving denoted by the IDM baseline. Evaluation metrics involve emissions reduction and speed improvement of vehicles and throughput improvement at the intersection - where lower emissions, higher speed, and higher throughput percentages indicate better performance.



(a) Emision benefits when 20% penetration of AVs are used with nominal policy (top) and multi-residual task learning (bottom)

(b) Emision benefits when protected left turns are present with nominal policy (top) and multi-residual task learning (bottom)

(c) Emision benefits when unprotected left turns are present with nominal policy (top) and multi-residual task learning (bottom)

Fig. 3: Visualization of t-SNE plots illustrating emission benefits (higher the better) in assessing the efficacy of MRTL policy in mitigating nominal policy limitations. t-SNE is used for dimensionality reduction of vectors describing incoming approaches to a two dimensional space (latent dimension 1 and 2 in the figures). Thus, each data point is an incoming approach, and the color denotes the emission benefits (a) with partial guided AV penetration (20%), (b) in the presence of protected left turns, and (c) when dealing with unprotected left turns. In all cases, the MRTL policy outperforms the nominal policy, evidenced by the predominance of blue data points in the lower-row figures as compared to the upper-row figures.

- **Actions**: longitudinal accelerations of the ego-vehicles. Lane changes are done by SUMO and not by the policy.
- **Rewards**: Ego-vehicle rewards are computed as $v_i(t) + w_1 e_i(t)$, where $w_1 = -7.57$ is a hyperparameter, and $v_i(t)$ and $e_i(t)$ represent the ego-vehicle's speed and emissions at time t, respectively. We use increasing $v_i(t)$ as a proxy for travel time reduction.

We adopt a neural network initialization method inspired by Silver et al. [10]. Initially, we set the last layer of the policy network to zero (and hence $f_\theta(s, c) = 0$ at the start). This prevents the MRTL policy from being worse than the nominal policy, especially when the nominal policy is close to optimal. We also include a 30-iteration pre-training phase for the critic to align better with the nominal policy, improving value estimates early on.

VI. Experimental Results

Here, we present the experimental results of employing MRTL for eco-driving at signalized intersections.

A. Baselines

In order to assess the benefit of the MRTL framework, we leverage three baselines to compare the performance.

1) **Intelligent Driver Model (IDM) [23]:** human-like driving baseline. The IDM [23] is used.
2) **Multi-task reinforcement learning:** Multi-task reinforcement learning from the scratch as introduced in Section III-B.
3) **Nominal policy:** policy in algorithm 1.

We do not use exhaustive training (training a different model on each intersection) as a baseline since it is prohibitively expensive given the large number of intersections and, hence, practically less useful for eco-driving.

B. Performance and generalization

In Table I, we analyze emission reduction and speed improvements of vehicles and throughput increase at the intersection at 20% and 100% AV penetration across 600

signalized intersections. Our findings highlight MRTL's effectiveness in enhancing emission reductions due to better generalization. Our MRTL policy improves benefits even in partial penetration scenarios when the nominal policy falls short. Furthermore, training multi-task reinforcement learning policies from scratch is challenging at both penetration levels. Suboptimal individual agent performances in multi-agent settings can lead to training collapse, especially in scenarios where vehicles follow one another. This can be seen in the significant emissions increase in multi-task reinforcement learning when comparing 20% to 100% penetration.

C. Nominal policy limitations

In Section V-A.1, we discussed limitations in our nominal policy design. Here, in Figure 3, we explore how the MRTL framework effectively addresses these limitations. We focus on three settings: partial penetration, intersections with protected left turns, and those with unprotected left turns. Through t-SNE [27] distribution plots as performance profiles, we show that MRTL significantly improves over the nominal policy performance in all three settings, with benefits extending across majority of intersections.

D. Control noise and bias noise

While conventional eco-driving controllers like GLOSA [5] struggle with noise from communication delays and sensor issues, MRTL policies can adapt well to such noise. Moreover, the nominal policies can be biased toward certain cities or conditions, but DRL can learn to adapt on top of them. To test this adaptability, we introduce control gaussian noise $\epsilon_c = \mathcal{N}(0, \sigma^2)$, varying σ^2 and a bias gaussian noise $\epsilon_b = \mathcal{N}(\mu, 0.3)$, varying μ to AV accelerations. In Figure 4 left, MRTL policies are more resilient to control noise, with only a 3% performance decrease compared to a significant 18% drop in the nominal policy. Similar results can also be seen under bias noise in Figure 4 right.

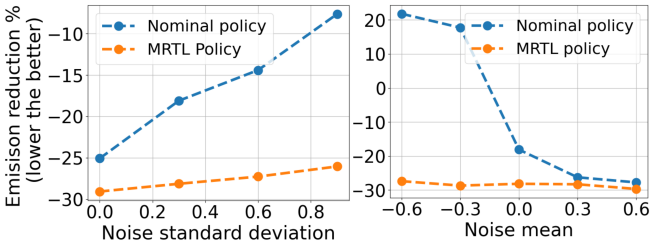


Fig. 4: Effect of control noise (left) and bias noise (right) on emissions.

E. Why does multi-residual task learning work?

Intuitively, MRTL simplifies policy search by fine-tuning from a nominal policy that is suboptimal yet not too far from the optimal policy, while learning a multi-task policy from scratch necessitates more computational effort due to possibly distant random initialization. This contrast is illustrated in Figure 5 (left) with π^* as the optimal policy, π_n as the nominal policy for MRTL, and π_0 as the random

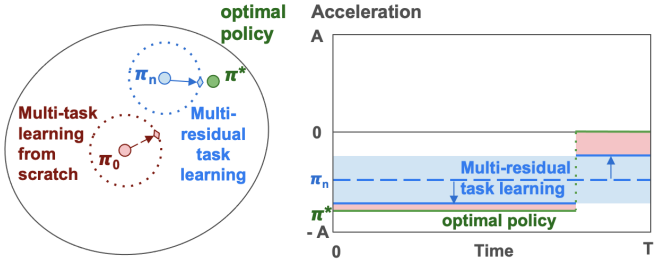


Fig. 5: Schematic interpretation of MRTL in policy search. Left: MRTL enables better policy search initialization compared to initializing from scratch. Right: A concrete example from eco-driving at signalized intersections.

initialization. MRTL allows searching within a high-quality ball around π_n, yielding a good solution near the optimal policy, while π_0 remains far from optimal under the same computational budget. In practice, the distance between policies and the performance landscape may be highly non-convex due to the nonconvexity of the objective function and the policy learning process. However, the nominal policy offers a favorable warm start to MRTL by initializing the search closer to the optimal policy compared to random initialization in learning from scratch.

As an example, consider a vehicle that eco-drives when encountering a red light. In Figure 5 (right), we contextualize the above interpretation with a commonly observed behavior in a subset of our traffic experiments by considering a general MDP that can lead to multiple MDPs based on traffic signal timing plans. We take the known strategy of gliding (constant deceleration throughout) as the nominal policy π_n [4]. With potential deviations like human vehicles at the traffic light, the optimal policy π^* may involve piecewise-constant acceleration (glide until the leading vehicle is met, then constant velocity to keep a constant headway, generally for a short time period before crossing the intersection).

MRTL from the gliding policy allows the search space (blue region) and the best policy within the search space (blue solid lines) to be close to the optimal policy. In contrast, random initialization from the entire action space $\prod_{t=1}^{T} a_t$, where $a_t \in [-A, A]$, on average results in the zero acceleration policy, which is further away from the optimal policy than the gliding residual initialization. Moreover, random initialization usually leads to non-smooth acceleration profiles in practice, potentially making policy search challenging, whereas the constant deceleration from the gliding policy for MRTL allows a smoother learning landscape.

VII. Conclusion

This study examines the algorithmic generalization of DRL in solving contextual Markov decision processes. We present MRTL as a generic framework for achieving this goal. MRTL uses DRL to acquire residual functions, improving upon conventional controllers. We apply MRTL to cooperative eco-driving, showing improved generalization in emission reductions. Potential future work includes analyzing MRTL to further understand the impact of different nominal policies on generalization.

REFERENCES

[1] Cathy Wu, Abdul Rahman Kreidieh, Kanaad Parvate, Eugene Vinitsky, and Alexandre M. Bayen. Flow: A modular learning framework for mixed autonomy traffic. *IEEE Transactions on Robotics*, 2022.

[2] Vindula Jayawardana and Cathy Wu. Learning eco-driving strategies at signalized intersections. In *European Control Conference*, 2022.

[3] Nathan Lichtlé, Eugene Vinitsky, Matthew Nice, Benjamin Seibold, Dan Work, and Alexandre M. Bayen. Deploying traffic smoothing cruise controllers learned from trajectory data. In *2022 International Conference on Robotics and Automation (ICRA)*, 2022.

[4] Yuhan Huang, Elvin CY Ng, John L Zhou, Nic C Surawski, Edward FC Chan, and Guang Hong. Eco-driving technology for sustainable road transport: A review. *Renewable and Sustainable Energy Reviews*, 93:596–609, 2018.

[5] Konstantinos Katsaros, Ralf Kernchen, Mehrdad Dianati, and David Rieck. Performance study of a green light optimized speed advisory (glosa) application using an integrated cooperative its simulation platform. In *2011 7th International Wireless Communications and Mobile Computing Conference*, pages 918–923. IEEE, 2011.

[6] Seyed Amin Sajadi-Alamdari, Holger Voos, and Mohamed Darouach. Nonlinear model predictive control for ecological driver assistance systems in electric vehicles. *Robotics and Autonomous Systems*, 2019.

[7] Karl Cobbe, Oleg Klimov, Chris Hesse, Taehoon Kim, and John Schulman. Quantifying generalization in reinforcement learning. In *International Conference on Machine Learning*. PMLR, 2019.

[8] Vindula Jayawardana, Catherine Tang, Sirui Li, Dajiang Suo, and Cathy Wu. The impact of task underspecification in evaluating deep reinforcement learning. *Advances in Neural Information Processing Systems*, 35:23881–23893, 2022.

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016.

[10] Tom Silver, Kelsey Allen, Josh Tenenbaum, and Leslie Kaelbling. Residual policy learning. *arXiv preprint arXiv:1812.06298*, 2018.

[11] Tobias Johannink, Shikhar Bahl, Ashvin Nair, Jianlan Luo, Avinash Kumar, Matthias Loskyll, Juan Aparicio Ojea, Eugen Solowjow, and Sergey Levine. Residual reinforcement learning for robot control. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 6023–6029, 2019.

[12] Ruiqi Zhang, Jing Hou, Guang Chen, Zhijun Li, Jianxiao Chen, and Alois Knoll. Residual policy learning facilitates efficient model-free autonomous racing. *IEEE Robotics and Automation Letters*, 7(4):11625–11632, 2022.

[13] Peng Hao, Tao Lu, Shaowei Cui, Junhang Wei, Yinghao Cai, and Shuo Wang. Meta-residual policy learning: Zero-trial robot skill adaptation via knowledge fusion. *IEEE Robotics and Automation Letters*, 7(2):3656–3663, 2022.

[14] Lukas Hewing, Kim P Wabersich, Marcel Menner, and Melanie N Zeilinger. Learning-based model predictive control: Toward safe learning in control. *Annual Review of Control, Robotics, and Autonomous Systems*, 3:269–296, 2020.

[15] Igor Mordatch, Nikhil Mishra, Clemens Eppner, and Pieter Abbeel. Combining model-based policy search with online model learning for control of physical humanoids. In *2016 IEEE international conference on robotics and automation (ICRA)*, pages 242–248. IEEE, 2016.

[16] Torsten Koller, Felix Berkenkamp, Matteo Turchetta, and Andreas Krause. Learning-based model predictive control for safe exploration. In *2018 IEEE Conference on Decision and Control (CDC)*, pages 6059–6066, 2018.

[17] Marc Peter Deisenroth, Dieter Fox, and Carl Edward Rasmussen. Gaussian processes for data-efficient learning in robotics and control. *IEEE transactions on pattern analysis and machine intelligence*, 37(2):408–423, 2013.

[18] Nicolas Heess, Gregory Wayne, David Silver, Timothy Lillicrap, Tom Erez, and Yuval Tassa. Learning continuous control policies by stochastic value gradients. *Advances in neural information processing systems*, 28, 2015.

[19] Marcel Menner, Karl Berntorp, Melanie N Zeilinger, and Stefano Di Cairano. Inverse learning for human-adaptive motion planning. In *2019 IEEE 58th Conference on Decision and Control (CDC)*, pages 809–815. IEEE, 2019.

[20] Glen Chou, Necmiye Ozay, and Dmitry Berenson. Learning constraints from locally-optimal demonstrations under cost function uncertainty. *IEEE Robotics and Automation Letters*, 5(2), 2020.

[21] Anurag Ajay, Jiajun Wu, Nima Fazeli, Maria Bauza, Leslie P Kaelbling, Joshua B Tenenbaum, and Alberto Rodriguez. Augmenting physical simulators with stochastic neural networks: Case study of planar pushing and bouncing. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2018.

[22] Carolin Benjamins, Theresa Eimer, Frederik Schubert, Aditya Mohan, André Biedenkapp, Bodo Rosenhahn, Frank Hutter, and Marius Lindauer. Contextualize me–the case for context in reinforcement learning. *Transactions on Machine Learning Research*, 2022.

[23] Treiber, Hennecke, and Helbing. Congested traffic states in empirical observations and microscopic simulations. *Physical review. E*, 2000.

[24] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

[25] Edgar Ramirez Sanchez, Catherine Tang, Vindula Jayawardana, and Cathy Wu. Learning surrogates for diverse emission models. In *Tackling Climate Change with Machine Learning, NeurIPS*, 2022.

[26] Moves and other mobile source emissions models. Environmental Protection Agency.

[27] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.