

# Efficient Reinforcement learning for Intelligent Transportation Systems

*Overview of PhD Thesis by:*  
**Bálint Kővári**



Budapest University of Technology and Economics  
Faculty of Transportation Engineering and Vehicle Engineering  
Department of Control Transportation and Vehicle Systems

*Supervisor:*  
**Tamás Bécsi, PhD**  
Department of Control for  
Transportation and Vehicle Systems  
Budapest University of Technology and Economics

Submitted in Partial Fulfillment of the Requirements for the Degree of  
*Doctor of Philosophy*

June 10, 2025

# Chapter 1

## Introduction

In my thesis, I focus on enhancing the efficiency of Reinforcement Learning (RL) in the context of Intelligent Transportation Systems (ITS). I approach this challenge hierarchically, addressing multiple levels of efficiency improvement. Reinforcement Learning has gained substantial attention due to its impressive achievements ([Silver et al. \[2017\]](#), [Fawzi et al. \[2022\]](#)), especially in applications involving Intelligent Transportation Systems [Zhu et al. \[2023\]](#). However, one of the significant challenges is the training efficiency, as RL often requires numerous iterations to achieve satisfactory performance ([Henderson et al. \[2018\]](#)). This is particularly problematic in ITS, where multiple objectives such as sustainability and control efficiency must be balanced. My primary goal is to make RL more efficient by developing methods that reduce the required training iterations while maintaining or enhancing performance. To systematically tackle RL efficiency in ITS, I structured my research into several levels:

- **Problem Formulation:** I start by addressing how the formulation of a control problem significantly impacts the agent’s learning efficiency. The primary challenge here is to define state, action, and reward abstractions that encapsulate the problem accurately and facilitate efficient learning.
- **Tool Development for Problem Formulation:** I developed methods to aid the formulation of RL problems, focusing on reward function comparison. Since rewards guide learning, choosing the most effective reward function without an iterative trial-and-error process significantly boosts efficiency.
- **Scalability through Multi-Agent Reinforcement Learning:** I explore scalability within Variable Speed Limit Control (VSLC), utilizing a multi-agent approach to handle varying problem sizes effectively. This involves training agents that can generalize their learning across different problem scales.
- **Sample Efficiency via Experience Prioritization:** I introduce a novel experience prioritization technique to enhance RL’s sample efficiency. This method is problem-independent and focuses on reducing the number of training iterations required to achieve convergence, thus making RL more practical for real-world applications.

---

As introduced above, the first research direction in my thesis is investigating the impact of problem formulation on the reachable performance. For this investigation, I chose the TSC problem since it has a broad literature, and after analysing the RL-related papers on TSC, it was evident that most of the papers are trying to create a new problem formulation, and a small portion of the papers focus on deeper methodology. Consequently, there were enough results to investigate the problem formulation in TSC.

In the first Chapter, I developed a novel approach to traffic signal control using model-free Reinforcement Learning (RL). The key innovation is introducing a new rewarding concept based on the standard deviation of queue lengths across lanes at a single intersection. Instead of directly minimizing traditional metrics like average waiting time or travel time, I focused on reducing the variance in queue lengths. (Haydari et al. [2021]) details the common approaches for TSC. This shift in perspective significantly improves the agent’s ability to balance traffic flow efficiently, which conventional methods often struggle with. Existing approaches to traffic signal control, such as fixed-cycle controllers and actuated systems like SUMO’s built-in algorithm, either lack adaptability or fail to account for uneven traffic distribution among lanes. Previous RL-based methods typically optimize metrics like queue length or average speed individually, but this can lead to situations where some lanes clear up while others remain congested. My method addresses this by directly targeting traffic imbalance, making it inherently more adaptive and robust. I trained RL agents to evaluate my method using the new rewarding concept. I compared their performance against fixed-cycle controllers, SUMO’s actuated controller, and other RL methods that utilize conventional reward functions. The results demonstrated that my approach significantly outperformed the baselines in both classical metrics (like waiting time, travel time, and queue length) and sustainability measures (such as CO2 emissions and fuel consumption). This improvement stems from the tight integration of the state representation, action space, and reward function, which allows the agent to make more context-aware decisions. The method also demonstrated better generalization and stability across diverse traffic scenarios, highlighting its practical viability. By focusing on balancing traffic loads rather than simply minimizing isolated metrics, my approach not only improves traffic efficiency but also reduces environmental impacts. This makes it a promising modern urban traffic management solution, offering operational and sustainable benefits.

After the exhaustive training process in the above Chapter, I focused on the mitigation of the iterative training process that is attached to the evaluation and comparison of every problem formulation.

Identifying an optimal reward function requires iteratively training the RL agent with multiple candidate functions, which is both time-consuming and resource-intensive. To overcome this issue, I introduced a novel method using the Monte Carlo Tree Search (MCTS) algorithm to evaluate reward functions before training (Kocsis and Szepesvári [2006]). This approach eliminates the need for exhaustive training, saving significant computational resources. The core idea of my method is to leverage MCTS to predict the performance ranking of various reward functions for a given problem. By integrating the reward functions directly into the MCTS structure, the algorithm can simulate

---

potential outcomes without needing complete RL training. This way, MCTS not only evaluates the performance of each reward function but also identifies potential behavioral patterns that might emerge during training. The main advantage of this approach is that it allows one to select the most promising reward function in advance, significantly reducing the trial-and-error process traditionally used in RL. To demonstrate the effectiveness of this method, I applied it to two distinct control tasks: Traffic Signal Control (TSC) and Lane Keeping. In the TSC problem, I compared different reward strategies, such as minimizing queue length and maximizing average speed, using both MCTS and traditional RL methods like Deep Q-Network (DQN). The results showed that MCTS could accurately predict the best-performing reward function, matching the outcomes obtained through complete RL training. Similarly, in the lane-keeping task, where the goal is to maintain the vehicle's trajectory within a designated lane, MCTS effectively ranked the reward functions without requiring full agent training. The key finding was that the ranking of reward functions obtained through MCTS aligned consistently with the rankings from actual training, proving that MCTS can act as a reliable pre-evaluation tool. For instance, in the TSC scenario, the reward function that minimized the standard deviation of queue lengths consistently outperformed others in terms of CO2 emissions and fuel consumption when trained, as predicted by MCTS. Similarly, in the lane-keeping problem, the reward function emphasizing stability and minimal deviation from the centerline was correctly identified as the most effective. My approach addresses a critical bottleneck in applying RL to real-world problems by reducing the number of necessary RL training runs. Instead of experimenting with multiple reward functions through time-intensive training, MCTS provides an efficient pre-screening process, guiding researchers toward the most promising configurations. This advancement accelerates the development of RL-based solutions and cuts down on computational costs, making RL more practical for complex applications. My research shows that MCTS can significantly enhance RL efficiency by pre-evaluating reward functions. This method bridges the gap between theoretical reward formulation and practical application, allowing for faster, more reliable deployment of RL in various control tasks, including traffic management and autonomous driving. This approach will pave the way for more resource-efficient RL applications in intelligent transportation systems.

After analysing the MCTS algorithm from several aspects, I decided to continue the investigation of problem formulation, but in this case, it was from a scalability point of view. Scalability is a crucial question in RL, especially in Multi-Agent systems, since the complexity of the problems is greater. There are several multi-agent problems in ITS; hence, I formalized my goal as creating an agent that can handle different problem sizes while maintaining great performance. For the application, I chose the Variable Speed Limit Control problem because it has huge potential, and RL is relatively new in this domain in the literature.

I developed a novel approach for Variable Speed Limit Control (VSLC) using Multi-Agent Reinforcement Learning (MARL). The key contribution lies in the innovative sliding-window-based state representation, which makes the method scalable and invariant to the length of the controlled highway segment. This formulation allows the

---

MARL agent to perform efficiently regardless of the highway size, solving the challenge of scalability faced by previous methods (Zheng et al. [2023], Kušić et al. [2020]). The problem I addressed is the inefficiency of existing VSLC methods, especially when applied to highways of varying lengths. Traditional approaches either use fixed discrete speed limits or continuous speed limits, and typically require separate training for each highway length, making them impractical and computationally intensive (Zheng et al. [2023], Zhang et al. [2023]). My method, however, only needs to be trained once and remains effective even when applied to longer or more complex highway segments. The core innovation is the state representation, which leverages a sliding window over adjacent highway zones, allowing each agent to perceive local traffic conditions without requiring information from the entire highway. This method enables consistent performance regardless of the highway size and reduces the computational load by avoiding retraining for different segment lengths. Additionally, the action space is designed to incrementally adjust speed limits, which helps maintain smooth traffic flow and reduces shockwave effects. I demonstrated the effectiveness of this method by training the MARL agent on a 1 km highway segment and evaluating it on 1 km, 3 km, and 10 km segments. The results consistently showed that my approach significantly outperforms the baseline methods, including the Motorway Control System (MCS) and control-free scenarios. Moreover, the agent maintained stable performance even when the highway length increased, proving the method’s scalability. In conclusion, my method offers a scalable, efficient solution for VSLC by leveraging MARL with a sliding-window-based state representation. This improves traffic flow and reduces environmental impact, making it a promising approach for real-world traffic management systems.

All the above-mentioned results and methods that I proposed are somewhat domain-specific or do not improve RL itself. Consequently, my goal in the last thesis is to develop a methodology that can improve RL sample inefficiency. Sample inefficiency can be tackled in many ways, but I choose training sample prioritization since it is a lightweight methodology that can have a tremendous influence on the final performance and the convergence speed of the agent.

I designed a novel experience prioritization method for reinforcement learning (RL) that significantly improves training efficiency. The key contribution is an innovative approach that combines Upper Confidence Bound (UCB) principles with existing prioritization techniques, specifically enhancing the exploration-exploitation trade-off during training. This method optimizes sampling efficiency and accelerates convergence compared to the state-of-the-art Prioritized Experience Replay (PER) (Schulman et al. [2015]). The problem I addressed is the inefficiency of the sampling strategies used during the training of RL agents. Traditional methods, particularly PER, often over-prioritize high-error experiences, leading to excessive exploitation while neglecting exploration. This imbalance results in slower convergence and increased computational costs. My proposed method integrates UCB to dynamically balance exploration and exploitation by considering the frequency of experience usage alongside the temporal difference (TD) error. This dual focus lets the agent prioritize informative and under-explored experiences, thus maintaining a more efficient learning trajectory. To demonstrate the

---

effectiveness of my approach, I conducted experiments using four benchmark environments from the OpenAI Gym: CartPole, Acrobot, Taxi, and CliffWalking. I trained RL agents using my UCB-based method and the baseline PER under identical conditions. The results showed that my method consistently outperformed PER regarding faster convergence and higher cumulative rewards. For example, in the The baseline for comparison, Prioritized Experience Replay (PER), is RL’s most commonly used method for experience prioritization. PER ranks experiences based on their TD error, often leading to repetitive learning from a small set of highly prioritized samples. In contrast, my UCB-based approach introduces a more balanced sampling strategy by combining priority value with exploration incentives, allowing for a more diverse and informative learning process. In summary, my proposed experience prioritization method addresses the inefficiency of sampling strategies in RL by offering a more balanced exploration-exploitation approach. The improvements in convergence speed and training stability show that my process can significantly reduce computational requirements, making RL more practical for complex real-world applications.

# Chapter 2

## New Results

### 2.1 Thesis I

I developed a new problem formulation for model-free Reinforcement Learning agents in single intersection traffic signal control. The new problem formulation introduces a reward paradigm using the standard deviation of the lanes' queue lengths and a queue length-based state representation. The Reinforcement Learning agent trained with the new problem formulation performs better regarding waiting time, travel time, queue length, fuel consumption, CO<sub>2</sub> emission, and NO<sub>x</sub> emission than other problem formulations from the literature.

Related publications:

- (KPAB22) Kővári B, Pelenczei B, Aradi S and Bécsi T (2022), "Reward Design for Intelligent Intersection Control to Reduce Emission", IEEE Access. Vol. 10, pp. 39691 - 39699.
- (KTB21) Kővári B, Tettamanti T and Bécsi T (2021), "Deep Reinforcement Learning based approach for Traffic Signal Control", In Proceedings of The 24th Euro Working Group on Transportation Meeting (EWGT2021).
- (KSzBAG21) Kővári B, Szőke L, Bécsi T, Aradi S and Gáspár P (2021), "Traffic Signal Control via Reinforcement Learning for Reducing Global Vehicle Emission", Sustainability, MDPI., October, 2021. Vol. 13(11254), pp. 18.

---

## 2.2 Thesis II

I demonstrated that the Monte-Carlo Tree Search algorithm can evaluate the different reward functions of a Markov Decision Process. This virtue of the Monte-Carlo Tree Search algorithm can replace the iterative reward design or comparison of an MDP that requires retraining of an agent, since the MCTS can yield the ranking between the agents' performance trained with the different reward functions in advance. I utilized a single-intersection traffic signal control and a lateral-control task of an autonomous vehicle to show the reward function evaluation capability of the MCTS algorithm. In both cases, the ranking between the performances reached with different reward functions was the same with training the agent and evaluating with MCTS.

Related publications:

- (KPAB24) Kővári B, Pelenczei B, Knáb IG and Bécsi T (2024), "Beyond Trial and Error: Lane Keeping with Monte Carlo Tree Search-Driven Optimization of Reinforcement Learning", *Electronics*. Vol. 13(11)
- (KPB22) Kővári B, Pelenczei B, and Bécsi T (2022) "Monte Carlo Tree Search to Compare Reward Functions for Reinforcement Learning." 2022 IEEE 16th International Symposium on Applied Computational Intelligence and Informatics (SACI). IEEE

---

## 2.3 Thesis III

I have designed a novel problem formulation for model-free Multi-Agent Reinforcement Learning agents in the Highway section's Variable Speed Limit Control. The novel problem formulation introduces a state representation used as a sliding window. Compared to other state representations from the literature, the proposed state representation does not use information from the entire controlled Highway section; this way, the proposed state representation is invariant to the size of the controlled Highway section. I demonstrated the Highway section size invariance of the state representation by training an agent for a 1km long Highway section, then evaluated on 1km, 5km, and 10km long Highway sections, and the agent consistently outperformed the baseline solutions regarding waiting time, CO2 emission, and NOx emission.

- (KKEBA24) Kővári B, Knáb IG, Esztergár-Kiss D, Bécsi T and Aradi S (2024), "Distributed highway control: a cooperative reinforcement learning-based approach", IEEE ACCESS, vol. 12, pp. 104463-104472
- (KKB23) Kővári, B., Knáb, I.G., Bécsi, T. (2025). Variable Speed Limit Control for Highway Scenarios a Multi-agent Reinforcement Learning Based Approach. In: Proceedings of the 2nd Cognitive Mobility Conference. COGMOB 23 2023. Lecture Notes in Networks and Systems, vol 1345. Springer, Cham.

---

## 2.4 Thesis IV

I developed a novel, problem-independent training sample prioritization method for model-free value-based Reinforcement Learning agents. It calculates the choice probability of each training sample by combining its temporal-difference error and the number of updates it undergoes. This approach addresses the exploration–exploitation trade-off in sample prioritization. The proposed method demonstrates faster convergence and higher cumulative rewards than the widely used Prioritized Experience Replay, based on extensive evaluations in the CartPole, Acrobot, Taxi, and CliffWalking Gymnasium environments.

Related publications:

- (KPB23) Kóvári B, Pelenczei B and Bécsi T (2023), "Enhanced Experience Prioritization: A Novel Upper Confidence Bound Approach", IEEE Access. Vol. 11, pp. 138488-138501.

## Chapter 3

# Future Research directions

The future direction of my research will focus on enhancing the sample efficiency of Reinforcement Learning (RL) by tackling the issue of redundant training samples. In my previous work, I introduced a method to estimate the information value of each training sample, but there remains the challenge of determining which samples are even worth using. In RL, every interaction is typically stored in the memory buffer and used for training, but many of these samples are redundant, hindering the agent from fully exploiting the most valuable data.

To address this, I am considering two main approaches. The first involves combining Reinforcement Learning with Active Learning (AL). Traditionally used in Computer Vision, AL helps identify unlabeled training samples that are likely to enhance model performance when labeled. Applying this concept to RL would involve selecting the most informative experiences from the memory buffer rather than including every interaction.

The second approach focuses on sufficient exploration during the sample collection phase. In RL, the agent's trial-and-error strategy often results in uneven coverage of the state space, leading to gaps in learning. To mitigate this, I aim to develop a method that actively considers the distribution of training samples within the state space, ensuring a more balanced sample density. This would enable the agent to learn more efficiently by focusing on under-explored regions.

These methods also have potential applications beyond RL, particularly in motion planning, where guided exploration can optimize state space coverage. Additionally, in other Deep Learning domains like Computer Vision, achieving balanced data representation can improve performance while reducing training time by eliminating redundant samples.

By pursuing these directions, I aim to not only improve RL's efficiency but also provide a framework that could benefit other areas of Deep Learning, offering more efficient training processes across various applications.

## Chapter 4

# Publication of the Author related to the thesis

- (KPAB22) Kővári B, Pelenczei B, Aradi S and Bécsi T (2022), "Reward Design for Intelligent Intersection Control to Reduce Emission", IEEE Access. Vol. 10, pp. 39691 - 39699.
- (KTB21) Kővári B, Tettamanti T and Bécsi T (2021), "Deep Reinforcement Learning based approach for Traffic Signal Control", In Proceedings of The 24th Euro Working Group on Transportation Meeting (EWGT2021).
- (KSzBAG21) Kővári B, Szőke L, Bécsi T, Aradi S and Gáspár P (2021), "Traffic Signal Control via Reinforcement Learning for Reducing Global Vehicle Emission", Sustainability, MDPI., October, 2021. Vol. 13(11254), pp. 18.
- (KPAB24) Kővári B, Pelenczei B, Knáb IG and Bécsi T (2024), "Beyond Trial and Error: Lane Keeping with Monte Carlo Tree Search-Driven Optimization of Reinforcement Learning", Electronics. Vol. 13(11)
- (KPB22) Kővári B, Pelenczei B, and Bécsi T (2022) "Monte Carlo Tree Search to Compare Reward Functions for Reinforcement Learning." 2022 IEEE 16th International Symposium on Applied Computational Intelligence and Informatics (SACI). IEEE
- (KKEBA24) Kővári B, Knáb IG, Esztergár-Kiss D, Bécsi T and Aradi S (2024), "Distributed highway control: a cooperative reinforcement learning-based approach", IEEE ACCESS, vol. 12, pp. 104463-104472
- (KKB23) Kővári, B., Knáb, I.G., Bécsi, T. (2025). Variable Speed Limit Control for Highway Scenarios a Multi-agent Reinforcement Learning Based Approach. In: Proceedings of the 2nd Cognitive Mobility Conference. COGMOB 23 2023. Lecture Notes in Networks and Systems, vol 1345. Springer, Cham.

---

(KPB23) Kővári B, Pelenczei B and Bécsi T (2023), "Enhanced Experience Prioritization: A Novel Upper Confidence Bound Approach", IEEE Access. Vol. 11, pp. 138488-138501.

# References

- Fawzi, A., Balog, M., Huang, A., Hubert, T., Romera-Paredes, B., Barekatin, M., Novikov, A., R Ruiz, F. J., Schrittwieser, J., Swirszcz, G., et al. (2022). Discovering faster matrix multiplication algorithms with reinforcement learning. *Nature*, 610(7930):47–53. [1](#)
- Haydari, A., Zhang, M., Chuah, C.-N., and Ghosal, D. (2021). Impact of deep rl-based traffic signal control on air quality. In *2021 IEEE 93rd Vehicular Technology Conference (VTC2021-Spring)*, pages 1–6. IEEE. [2](#)
- Henderson, P., Islam, R., Bachman, P., Pineau, J., Precup, D., and Meger, D. (2018). Deep reinforcement learning that matters. In *Thirty-Second AAAI Conference on Artificial Intelligence*. [1](#)
- Kocsis, L. and Szepesvári, C. (2006). Bandit based monte-carlo planning. In *European conference on machine learning*, pages 282–293. Springer. [2](#)
- Kušić, K., Dusparic, I., Guériau, M., Gregurić, M., and Ivanjko, E. (2020). Extended variable speed limit control using multi-agent reinforcement learning. In *2020 IEEE 23rd International Conference on Intelligent Transportation Systems (ITSC)*, pages 1–8. IEEE. [4](#)
- Schulman, J., Levine, S., Abbeel, P., Jordan, M., and Moritz, P. (2015). Trust region policy optimization. In *International conference on machine learning*, pages 1889–1897. [4](#)
- Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., Hubert, T., Baker, L., Lai, M., Bolton, A., et al. (2017). Mastering the game of go without human knowledge. *Nature*, 550(7676):354–359. [1](#)
- Zhang, Y., Quinones-Gruero, M., Barbour, W., Zhang, Z., Scherer, J., Biswas, G., and Work, D. (2023). Cooperative multi-agent reinforcement learning for large scale variable speed limit control. In *2023 IEEE International Conference on Smart Computing (SMARTCOMP)*, pages 149–156. IEEE. [4](#)
- Zheng, S., Li, M., Ke, Z., and Li, Z. (2023). Coordinated variable speed limit control for consecutive bottlenecks on freeways using multiagent reinforcement learning. *Journal of advanced transportation*, 2023(1):4419907. [4](#)

Zhu, Z., Lin, K., Jain, A. K., and Zhou, J. (2023). Transfer learning in deep reinforcement learning: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

[1](#)