Delivery assignment to respect CO_2 constraint/quota using Reinforcement Learning

Farid Najar, Yann Strozecki, Dominique Barth

DAVID Lab, UVSQ, Université Paris-Saclay
{first_name.last_name}@uvsq.fr

Keywords : *machine learning, reinforcement learning, operations research, optimization, logistics.*

1 Introduction

Amidst the urgent backdrop of climate change, nations are compelled to take decisive measures to reduce their emissions, thereby mitigating the exacerbation of this critical issue and its far-reaching consequences. Specifically, local authorities are actively pursuing strategies to curtail their environmental footprint, which includes implementing restrictions on emissions of greenhouse gases like CO_2 and other pollutants. Our particular focus centers on the effective management of supply chain emissions, with a specific emphasis on emissions associated with the delivery process.

In this work, we begin by formalizing the problem as a combinatorial optimization model. Subsequently, we employ well-established classical methods in operations research to obtain baseline results. Building upon this foundation, we embark on the development of innovative methods based on machine learning to address and resolve these complex issues in our future research.

2 The assignment

2.1 The model and existing methods

We consider a transporter that has K deliveries to make in different locations with M vehicles that have different characteristics like energy type and capacity. This is a classical VRP model. However, a pollution quota is imposed to restrict the emission. Therefore, as it is a strong constraint, the transporter has to omit some deliveries to respect it because the emission depends on the distance traveled, and the energy used by the vehicles. So, our objective is to find an *assignment* $a \in \{0, 1\}^K$, with $a_k = 0$ if the k-th delivery is omitted and 1 otherwise, that respects the quota and yields the smallest cost possible for the transporter. Note that the omission of deliveries are penalized. This is a combinatorial optimization problem that can be solved with methods such as *simulated annealing*. We also applied our variant of A^* to find solutions, and it has comparable performances with the latter method.

2.2 Our approach

We aim for our algorithm to leverage past experiences, enabling it to swiftly produce valid and preferably near-optimal solutions, thereby enhancing overall performance.

Our problem can be conceptualized as a tree, where each action involves the removal of a delivery. Employing the reinforcement learning paradigm, we utilize the PPO algorithm [4]. Specifically, we adopt its maskable variant [2] to prevent the selection of deliveries already omitted. Furthermore, our approach incorporates diverse observation and reward functions.

The observations may be the cost matrix, the routes given by the VRP solver, etc. And the reward is the negative total cost penalized by omission that has been transformed to be in range $[0, 1]$.

2.3 Preliminary results

We tried our model for a model with 100 deliveries and 4 vehicles (1 EV, 1 hybrid, 2 diesel) on a synthetic graph city with different observation functions on one instance, i.e. keeping same deliveries for all episodes, and assignment observation for different instances (see Figure 1). It achieves performances like simulated annealing for one instance, but it is unable to learn a global strategy for all instances yet.

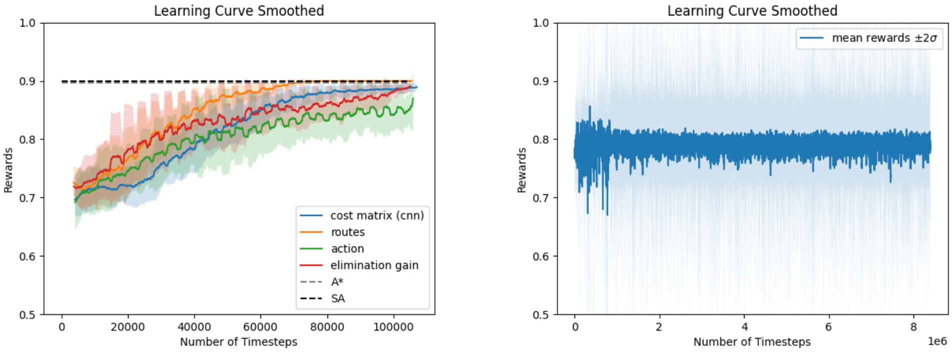

FIG. 1: On the left the learning curve of different types of models (with different observation) that converges to the values obtained by simulated annealing and A^* on one instance in fewer than 100K iterations. On the right, we can see that the RL agent is not able to generalize and learn a strategy for all instances yet, even after 8 millions iterations.

3 Conclusions et perspectives

Our model proved to be able to find a suitable solution when we give him the same destinations in every episode, but it is, yet, unable to generalize its success. The problem remain open, and we are interested in pushing our research further by designing new methods using the graph structure of the routes with GNNs for RL[3], and other methods based on kernels [1]

References

[1] Omar Darwiche Domingues, Pierre Ménard, Matteo Pirotta, Emilie Kaufmann, and Michal Valko. Kernel-Based Reinforcement Learning: A Finite-Time Analysis, March 2022. arXiv:2004.05599 [cs, stat].

[2] Shengyi Huang and Santiago Ontañón. A Closer Look at Invalid Action Masking in Policy Gradient Algorithms. *The International FLAIRS Conference Proceedings*, 35, May 2022. arXiv:2006.14171 [cs, stat].

[3] Mingshuo Nie, Dongming Chen, and Dongqi Wang. Reinforcement learning on graphs: A survey, January 2023. arXiv:2204.06127 [cs].

[4] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal Policy Optimization Algorithms, August 2017. arXiv:1707.06347 [cs].