Dynamic Scheduling of Gantry Robots using Cooperative Multi-Agent Reinforcement Learning

Jannik Felix Hinrichs

Supervisors: Prof. Dr. Horst Zisgen, Prof. Dr. Arnim Malcherek

h_da
HOCHSCHULE DARMSTADT
UNIVERSITY OF APPLIED SCIENCES

Motivation

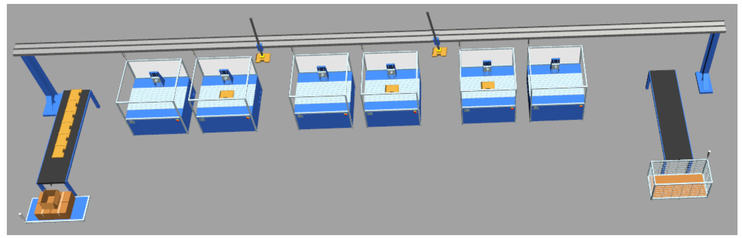

Figure 1. Digital twin of the employed production line.

The objective of the project *Dynamic Scheduling of Gantry Robots using Simulation and Reinforcement Learning* [3] is to examine the potential of using Reinforcement Learning (RL) [2] to control gantry robots in a production line. Figure 1 illustrates the digital twin of a production line comprising three work centers, each with two work stations, and two gantry robots, which are also referred to as loaders. The loaders are responsible for the transportation of the workpieces, which must be processed once at each work center. The previous research employed a Single-Agent Reinforcement Learning (SARL) approach, where one agent was responsible for controlling all employed loaders [3]. The agent utilized a Deep Q-Network (DQN) [1] to approximate the Q-function within the RL domain. The work successfully demonstrated that the RL agent can learn a strategy at least equivalent to conventional control methods.

Nevertheless, the practical application of the SARL approach is constrained by the increasing complexity of the simulated system, particularly in scenarios involving a greater number of work stations and multiple loaders. The exponential growth of the state space, which is a consequence of the necessity of mapping the entire environment, and the asynchronous environment are identified as the primary challenges. The asynchronous environment is a consequence of the varying execution times of the actions, as the digital twin simulates a continuous rather than a discrete process. It would be an inefficient use of time to await the completion of all robot actions. Consequently, at each time step t, the agent selects only the action designated for the requesting loader. This results in scenarios where the selected action a_t is not fully executed in the subsequent state s_{t+1}, when another loader requests the action in time step $t+1$.

Research Goal

This master's thesis examines how these challenges can be addressed through the application of a cooperative Multi-Agent Reinforcement Learning (MARL) approach. The following requirements must be satisfied:

- In order to guarantee that the observed experiences align with the principles of RL, it is essential to address the challenge of the asynchronous environment.
- In order to counteract the *curse of dimensionality*, it is necessary to reduce the cardinality of the state space.
- In order to achieve the highest possible throughput of workpieces, it is imperative that the agents adopt a unified strategy that guarantees the cooperative effort of all agents.

Methodology

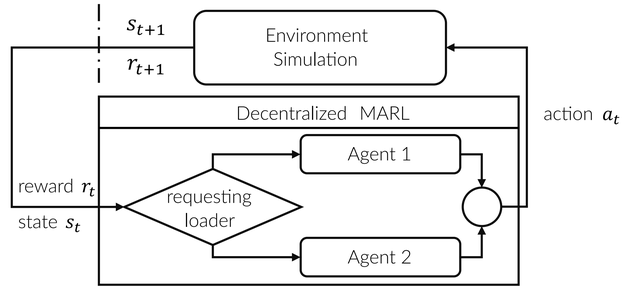

Figure 2. Structure of the RL problem with two independent agents.

The proposed MARL approach employs multiple independent DQN agents, with each agent responsible for controlling only a single loader. In order to address the asynchronous environment, the global state s_t and the observed reward r_t are distributed only to the agent, that is responsible for the requesting loader (see Figure 2). This guarantees that each agent exclusively observes environment states in which its own previous action is fully executed. All modifications to the environment resulting from the actions of other loaders can be considered "passive", i.e. modifications that are not actively caused by the executed action of the requesting loader.

The state space cardinality is reduced in two stages. The application of partial observability results in a local state space for each agent, which includes only a subset of the features present in the global state space. Furthermore, the local state spaces are reduced by limiting the operational area of each loader. This thesis proposes a symmetrical division of the production line. The first loader is capable of moving between the input (represented by the left conveyor belt in Figure 1) and the fourth station, while the second loader can travel between the third station and the output (represented by the right conveyor belt). These modifications require that each agent utilizes its own replay memory to store its observed experiences during the training process. This is essential, as the local states of the agents represent different subsets of the global state. Moreover, each agent employs its own DQN with distinct parameters, as the local state maps to the input layer and the local action state maps to the output layer of the neural network. Consequently, the training of the agent's DQNs must be conducted independently.

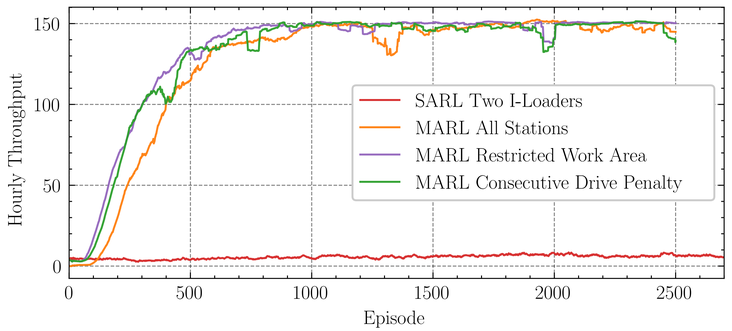

Figure 3. Moving average over 50 episodes of the achieved hourly workpiece throughput. For illustrative purposes, only the initial 2700 of the 40 000 episodes are presented for the SARL strategy.

Results

Figure 3 illustrates the training progress of three distinct MARL strategies in comparison to the existing SARL strategy. It is important to note that the SARL strategy requires training over 40 000 episodes to reach its maximum, which is not fully illustrated in Figure 3. The "MARL All Stations" strategy enables both loaders to reach all six work stations analogously to the SARL strategy. In contrast, the "MARL Restricted Work Area" strategy divides the production line symmetrically. In the "MARL Consecutive Drive Penalty" strategy, consecutive drive actions are penalized via the reward function, as they are deemed inefficient. It can be observed that all MARL strategies achieve a similar maximum throughput during the training process, outperforming the SARL strategy significantly.

The validation results of the trained strategies are presented in Table 1. The validation of the strategies is conducted over 100 episodes, during which the production process is simulated for two hours. As no random processes are employed within the system to enhance clarity, the values attained in each episode remain consistent. The results demonstrate that all three MARL strategies achieve a higher throughput than the SARL approach. By limiting the work area, throughput is further enhanced, albeit at the cost of a notable increase in the number of required actions. This issue can be addressed by implementing a penalty for consecutive drive actions.

Table 1. Validation results for existing SARL strategy and three different MARL strategies.

Strategy	Throughput	#Actions Agent 1	#Actions Agent 2
SARL Two I-Loaders	153	2600	-
MARL All Stations	158	1530	1350
MARL Restricted Work Area	159	1750	1450
MARL Consecutive Drive Penalty	159	1510	1290

Conclusion

The presented MARL approach demonstrates that the control of gantry robots can be achieved with independent agents. Notwithstanding the fact that the DQN agents act and are trained independently of one another, they have learned to cooperate in a manner that maximizes the achieved throughput. The number of training episodes required could be reduced from 40 000 to 2500 in comparison to the previous SARL approach. This is made possible due to the fact that the distribution of the environment state only to the responsible agents results in less variant training data. Moreover, the cardinality of the agent's state spaces can be reduced, due to the partial observation and the restriction of the work areas without any adverse effects. The global state space of the SARL approach has a cardinality of approximately 8×10^9. In contrast, the local state spaces of the MARL approach only have a cardinality of approximately 9×10^5 for the first agent and 4×10^5 for the second agent.

References

[1] Volodymyr Mnih et al. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–533, 2015.

[2] Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. The MIT Press, second edition, 2018.

[3] Horst Zisgen, Robert Miltenberger, Markus Hochhaus, and Niklas Stöhr. Dynamic scheduling of gantry robots using simulation and reinforcement learning. In *Proceedings of the Winter Simulation Conference*, WSC '23, page 3026–3034. IEEE Press, 2024.

Jannik.Hinrichs@stud.h-da.de