Available online at www.sciencedirect.com

ScienceDirect

Procedia CIRP 120 (2023) 986–991

www.elsevier.com/locate/procedia

56th CIRP Conference on Manufacturing Systems, CIRP CMS '23, South Africa

Optimal Manufacturing Configuration Selection: Sequential Decision Making and Optimization using Reinforcement Learning

Agajan Torayev[a,*], Jose Joaquin Peralta Abadia[b], Giovanna Martínez-Arellano[a], Mikel Cuesta[b], Jack C Chaplin[a], Felix Larrinaga[b], David Sanderson[a], Pedro-José Arrazola[b], Svetan Ratchev[a]

[a] *University of Nottingham, Nottingham, NG7 2RD, UK*
[b] *Mondragon Unibertsitatea, Arrasate 20500, Spain*

* Corresponding author. Tel.: +44 7799077892. *E-mail address:* agajan.torayev@nottingham.ac.uk

Abstract

In manufacturing, different costs must be considered when selecting the optimal manufacturing configuration. Costs include manufacturing costs, material costs, labor costs, and overhead costs. Optimal manufacturing configurations are those that minimize production criteria, such as costs, production speed, and flexibility, while still meeting the required production levels and quality standards. To find the optimal manufacturing configuration, manufacturers often use a combination of traditional techniques, e.g., mathematical modeling, simulation, and optimization, to evaluate the tradeoffs between different cost factors and identify configurations that provide the best balance between cost and performance. However, these techniques may require long development and simulation time, and/or may require expert knowledge. This paper presents a method for selecting the optimal manufacturing configuration, focusing on cost optimization, using a reinforcement learning (RL) approach for sequential decision-making. The proposed method involves developing a RL environment, requiring lower development and simulation times than traditional techniques, that captures the incurred costs, recurring costs, production rates, and setup times of manufacturing configurations. The problem is then solved using the Proximal Policy Optimization algorithm to identify the configuration that minimizes costs while still meeting the required production levels and quality standards. The effectiveness of the proposed method is validated through a machining process planning case study with multiple cost factors and production constraints. In particular, the machining process plan was developed for an industry-relevant product prototype. The results show that the proposed method can find solutions that are robust to stochastic noise, providing valuable insights for manufacturers looking to optimize manufacturing operations.

Keywords: manufacturing; optimization; decision making; artificial intelligence; machining

1. Introduction

With demand, requirements, and technologies changing constantly on a manufacturing shop floor [1], being able to select the optimal manufacturing configuration (MFG) for each demand period becomes essential to remain competitive in the global economy. To this end, on-demand self-adaptive manufacturing system approaches – such as reconfigurable manufacturing system (RMS) approaches – provide the tools to dynamically adapt to changing conditions. RMS approaches enable high throughput, flexibility, and quick and efficient reaction to changes, coping with the increasing industrial demand [2]. In particular, the capability of selecting the optimal MFG based on demand and costs is a critical research problem in RMS approaches.

Many researchers in the literature have addressed the selection of optimal manufacturing configurations. In [3], the authors propose an approach that considers important system-level evaluation criteria (e.g. cost and availability) and uses stochastic analysis to ensure the smoothness of the reconfiguration process. In [1], the authors go a step further by introducing a novel approach that utilizes module interactions and machine capability to measure the reconfigurability and operational capability of an reconfigurable machine tool, which can be used to optimize machine assignments for single-part flow lines. In [4], the authors present a two-phased method that addresses primary system configuration design and necessary system reconfigurations according to demand rate changes, leveraging the benefits of reconfigurable machine tools. Furthermore, the authors in [5] offer a simulation-based multi-objective optimization approach for system reconfiguration of multi-part flow lines, which ad-

10.1016/j.procir.2023.09.112

dresses task assignments and space allocation, and maximizes throughput while minimizing total space capacity.

While these approaches are effective, the lack of a crucial element is identified: sequential decision-making and optimization algorithms. Decision-making and optimization algorithms consider future consequences of current decisions. The immediate effects of a particular configuration can be considered and future demand and resource needs can be anticipated. This enables more efficient and effective decision-making, as configurations that meet current demand, while preparing for future demand, can be chosen. Additionally, sequential decision-making allows for the incorporation of constraints and uncertainty, further strengthening the robustness and effectiveness of the decision-making process. In particular, reinforcement learning (RL) algorithms prove useful for sequential decision-making and optimization problems with changing and uncertain environments [6].

In this paper, an optimal MFG selection approach based on sequential decision-making and optimization algorithms is proposed. The main contributions of this paper are two-fold:

- **RL environment:** A RL environment is proposed to represent real-life industrial manufacturing environments. An open-source customizable implementation is provided as a basis for future research [1].
- **RL-based methodology:** Building on the previous contribution, a sequential decision-making methodology for optimal MFG selection problem is presented and solved using the Proximal Policy Optimization (PPO)[7] RL algorithm.

The remainder of this paper is structured as follows. Section 2 presents the RL-based methodology, comprised of problem formulation and concepts of RL, as well as the RL environment definition. Section 3 presents the experiment conducted, describing the implementation and validation of the methodology. In addition, results and discussion are presented. Finally, Section 4 provides conclusions and an outlook on future work.

2. Methodology

The problem formulation of optimal configuration selection for demand satisfaction in manufacturing is presented in this section. Thereafter, RL for sequential decision making and optimization is presented as approach. Finally, this section ends with the description of the proposed manufacturing environment.

2.1. Problem formulation

The optimal manufacturing configuration selection for the demand satisfaction problem must be first mathematically formulated. To this end, the problem is formulated in this paper as follows.

- **Given:**
  - **Demand,** D: Number of products required.
  - **Demand time,** T_D: Maximum time allowed to produce the demanded products.
  - **Set of manufacturing configurations,** $\mathbb{M}$: An MFG is a group of resources that have the capability to produce the demanded product. The total number of unique manufacturing configurations is $M = |\mathbb{M}|$. Each MFG has the following attributes:
    - **Incurring cost**: Cost of purchasing the MFG.
    - **Recurring cost**: Cost of running a MFG for 1 unit of time.
    - **Production rate**: Number of products produced by the MFG per 1 unit of time.
    - **Setup time**: Time required to set up the MFG.
  - **Space size,** B: Maximum number of allowed manufacturing configurations to purchase.
- **Problem**: Find the multiset of manufacturing configurations that can meet the given demand in the given demand time with minimum cost, where the cost is the sum of incurred and recurred costs.

2.2. Reinforcement learning

In the context of sequential decision making and optimization, RL is a technique that allows an agent to learn how to make optimal decisions. Agents are key elements in RL, perceiving the environment, acting autonomously, and improving its performance with reward-based learning. Optimal decisions are made by interacting with an environment and receiving feedback in the form of rewards or penalties. The goal of RL is to train the agent to maximize the cumulative reward over time based on the feedback received from the environment, improving the decision-making process and adapting to changing conditions [6]. RL approaches have been successfully applied to various sequential decision-making and optimization problems such as smart grids [8] and marketing [9].

An environment in RL comprises a set of all possible actions, known as action space. The environment also includes a complete description of itself, with nothing hidden, known as the state space. An agent observes certain parts of the state space, known as the observation space, takes actions, and receives feedback as reward for each action it takes. The goal of the agent is defined by learning principles that aim to maximize the expected sum of all future rewards. Moreover, changes in the environment due to the actions of the agent are known as environment dynamics. The environment-agent interaction is depicted in Figure 1.

2.3. RL environment for manufacturing

In this sub-section, an RL environment for manufacturing is defined for the problem formulated in Subsection 2.1. The environment is defined based on necessary RL elements, such as state space, action space, environment dynamics, and learning principles for agents:

[1] https://github.com/torayeff/mfgrl

2 METHODOLOGY

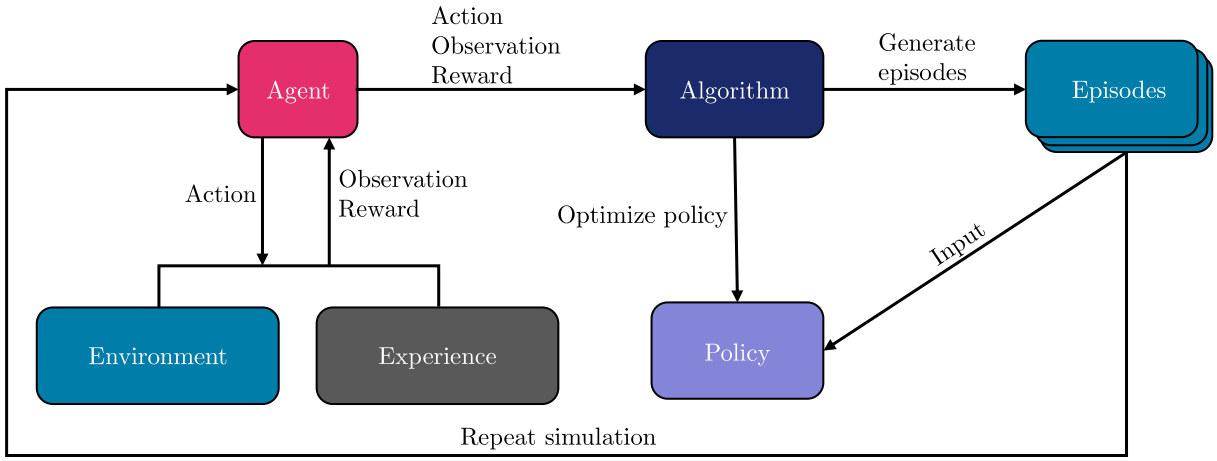

Fig. 1. Environment-Agent interaction in reinforcement learning. Adapted from [10].

State and action space. The **state space** of the environment is an $(2 + 6B + 4M)$-dimensional vector, where B is the space size and M is the set of manufacturing configurations defined in Subsection 2.1. Specifically, the vector consists of the following information:

- Remaining demand: $D_r \in \mathbb{Z}^+$, at initialization $D_r = D$.
- Remaining demand time: $T_r \in \mathbb{Z}^+$, at initialization $T_r = T_D$.
- Incurring costs of purchased manufacturing configurations: $I \in \mathbb{R}^B_{>0}$.
- Recurring costs of purchased manufacturing configurations: $R \in \mathbb{R}^B_{>0}$.
- Production rates of purchased manufacturing configurations: $P \in \mathbb{R}^B_{>0}$.
- Setup times, i.e., the time required to set up the newly purchased MFG: $U \in \mathbb{R}^B_{>0}$.
- Statuses of purchased manufacturing configurations, i.e., whether the configuration is producing or in the setup (or maintenance) phase: $S \in \mathbb{R}^B_{\geq 0, \leq 1}$.
- Produced products, i.e., the number of products produced by each of the purchased manufacturing configurations: $O \in \mathbb{R}^B_{\geq 0}$.
- Market incurring costs, i.e., stochastically changing the purchase prices of manufacturing configurations in the market: $\mathcal{I} \in \mathbb{R}^M_{>0}$.
- Market recurring costs, i.e., stochastically changing the recurring costs of manufacturing configurations in the market: $\mathcal{R} \in \mathbb{R}^M_{>0}$.
- Market production rates, i.e., stochastically changing the production rates of manufacturing configurations in the market: $\mathcal{P} \in \mathbb{R}^M_{>0}$.
- Market setup times, i.e., stochastically changing the setup times of manufacturing configurations in the market: $\mathcal{U} \in \mathbb{R}^M_{>0}$.

The **action space** in the environment is represented as an integer between 0 and M inclusive, i.e., $a \in [0, M]$ and is formally defined as

$$Step(a) = \begin{cases} \text{“buy configuration } a\text{”, if } 0 \leq a < M \\ \text{“continue production”, otherwise} \end{cases} \quad (1)$$

where $Step(a)$ makes one episode step in the environment.

Environment dynamics. Selected actions by an agent affect the dynamics of the environment as follows:

- Action “buy configuration a” adds a configuration a into the production space. It also pauses the remaining demand time, T_r, in the environment. Counter-intuitively, stopping the remaining demand time resembles a decision-making process in the real world where purchasing decisions can be made while production is still running.
- Action “continue production” decreases the remaining demand time and updates the produced products.
- An agent can make purchase decisions until the space is full. As soon as the space is full, an agent exceeds all its action choices, and the environment advances independently until the termination criteria are reached.
- The environment terminates when the condition “$D_r \leq 0$ OR $T_r \leq 0$” is met.

Learning principles. Two main **learning principles** are defined for agents:

- P1: Demand must be met at any cost.
- P2: Total cost must be minimized.

Learning principle P1 gives a high penalty if D is not met within the T_D. The penalty is defined as a function of the remaining demand and a K penalty coefficient as

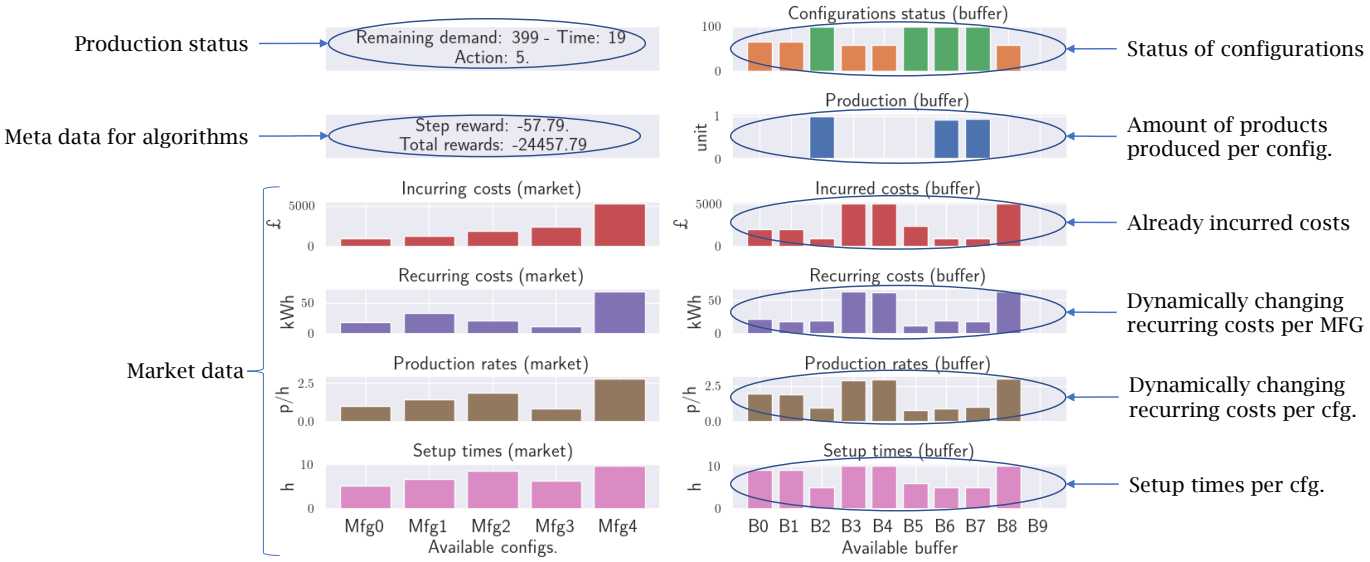

Fig. 2. Graphical user interface of the RL manufacturing environment.

$$J(D_r) = \begin{cases} -D_r K, & \text{if } D_r > 0 \\ 0, & \text{otherwise} \end{cases} \tag{2}$$

where K is a penalty coefficient, $K = R_{max} + 1$, and $R_{max} = \sum_{i=1}^{B} (\max_{1 \le j \le M} \mathcal{I}_j + T_D \max_{1 \le j \le M} \mathcal{R}_j)$. The equation (2) is derived from inequality $DK > R_{\max} + (D-1)K$.

Learning principle P2 gives two negative rewards, i.e. rewards are multiplied by negative one. The "buy configuration a" action rewards with incurring cost $\mathcal{I}_a$ of MFG a. The "continue production" action rewards the sum of recurring costs of running manufacturing configurations $\sum_a \mathcal{R}_a$.

3. Experiment

This section presents the experiment conducted in this paper. First, the implementation of the methodology is detailed, followed by a description of a machining process-planning case study with multiple cost factors and production constraints as validation of the RL-based methodology. Finally, results obtained from the validation and a discussion on the results are presented.

3.1. Implementation

The proposed environment is implemented using the Gymnasium library by Farama Foundation[2]. This library provides suitable templates for defining a custom environment compatible with reinforcement learning algorithms. The main goal of using the Gymnasium library to implement a manufacturing reinforcement learning environment is to provide better accessibility for manufacturing researchers to experiment with state-of-the-art reinforcement learning algorithms. The visual representation of the proposed manufacturing reinforcement learning environment is shown in Figure 2.

The sequential decision-making process for optimal MFG selection for the demand satisfaction problem is solved using Proximal Policy Optimization, a reinforcement learning algorithm [7]. Compared to other reinforcement learning algorithms, PPO improves the performance of the agent by making small adjustments to decision making policies rather than making large changes that may cause instability or poor performance. Therefore, the algorithm converges faster and is more efficient and stable, simplifying the implementation and achieving good results.

The training and hyper-parameter tuning is implemented and executed using the industry-grade reinforcement learning library RLlib[10]. One of the main benefits of using RLlib is quick experimentation and comparison of different algorithms, as well as possibilities of integration with existing systems. RLib provides built-in features for logging and monitoring the training process and for distributed training across multiple machines. In addition, pre-built implementations of popular reinforcement learning algorithms that can be used as-is or with minor modifications are also provided, thus reducing development time.

3.2. Validation Scenario

Validation tests are performed to showcase the ability of the RL-based methodology to identify configurations that minimize costs while meeting required production levels and quality standards. Machining process are extensively used in manufacturing, providing high-precision parts with good surface finishing [11]. However, industrial machining workshops may

[2] https://github.com/Farama-Foundation/Gymnasium

have changing demand and requirements, and technology may change constantly. Consequently, a machining process planning case study, involving turning, drilling and milling operations, is used as validation of the approach. Figure 3 presents the desired shape of the product. Three phases are required to achieve the desired shape: (i) right-side turning operation, (ii) left-side turning operation, and (iii) center-side milling operation. The case study has a demand D of 2000 products, a demand time T_D of 100 hours, and a space size B of 10 machine configurations.

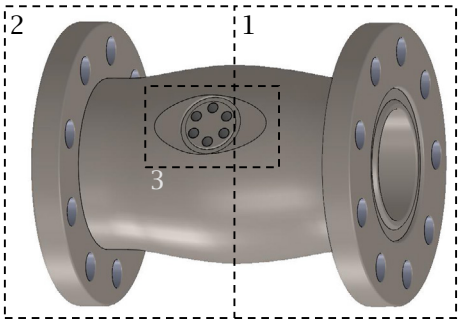

Fig. 3. Desired shape of the product.

The first phase requires a turning lathe capable of performing drilling, boring, turning, facing, and slotting operations. First, the center hole is drilled, bored and finished. Second, slotting and rough and finish facing operations are performed, followed by drilling of the nine front holes. At the end of the phase, rough turning is performed from the border to the center of the workpiece. The second phase is to be performed similarly in the turning lathe. The phase starts with slotting and rough and finish facing operations, followed by drilling of the other nine front holes. Thereafter, rough turning, external finish profiling and threading operations are executed. Finally, the third phase requires a milling center capable of performing face and slot milling, as well as drilling and threading. First, the cilindrical part is face milled and the six holes are drilled. Then, the six holes are threaded and the last slot milling operation is performed.

Given the characteristics of the desired shape, the minimum requirements for D satisfaction would include a turning lathe and a milling center, or a combination of both. Nevertheless, only one machine configuration may not suffice to satisfy T_D. As a solution, combinations of more than one MFG may satisfy T_D. As such, as a simulated example, five assets are proposed in the case study: (A0) a CNC lathe, (A1) a CNC milling center, (A2) a multitask CNC lathe, (A3) a dual-spindle CNC turning center, and (A4) a twin-spindle twin-turret turning center. Capabilities and normalized manufacturing data of the assets are presented in Table 1.

Drawing from the assets, five machine configurations are presented as being able to satisfy D. The machine configurations are comprised by one or more assets and the routes they follow are presented in Figure 4.

The data in Table 2 presents the normalized manufacturing data of the machine configurations. The data is provided to the RL environment for manufacturing, and optimization is performed using the PPO algorithm. Even though the data in Table 2 is fixed, the RL environment is stochastic, i.e., the incurring cost, recurring cost, production rate, and setup times change by $+/-10\%$ at every simulation step, resembling real industrial fluctuations. Moreover, the environment provides the possibility to define and experiment with different fluctuations.

Table 1. Normalized manufacturing data of the assets.

Asset	Turning lathe	Milling center	Incur. cost	Recur. Cost	Setup time
A0	X		300	12	2
A1		X	700	8	3
A2	X	X	2000	20	9
A3	X		1700	5	3
A4	X	X	5000	65	10

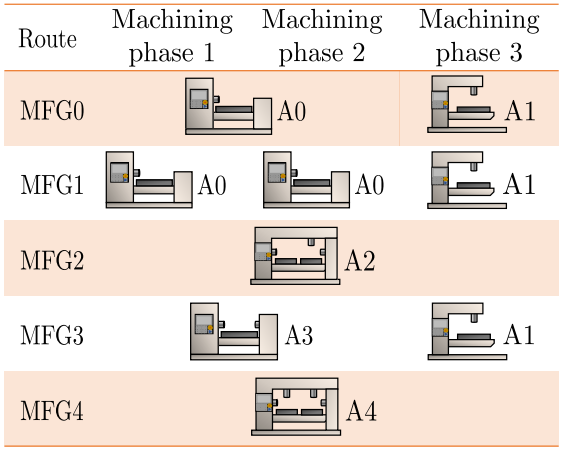

Fig. 4. Machine configuration routes.

Table 2. Normalized manufacturing data of the machine configurations. Demand $D = 2000$, demand time $T_D = 100$, space size $B = 10$

Cfg.	Incur. cost	Recur. cost	Prod. rate	Setup time
MFG0	1000.0	20.0	1.0	5
MFG1	1300.0	32.0	1.5	7
MFG2	2000.0	20.0	2.0	9
MFG3	2400.0	13.0	0.75	6
MFG4	5000.0	65.0	3.0	10

Six experiments have been defined to validate the robustness of the methodology to changing demands. Demand D and demand time T_D defined for each experiment are presented in Table 3, as well as the results obtained. Development and simulations time were faster than traditional techniques. Training (development) times lasted on average 10 min per experiment. Simulation times lasted less than one second without visual rendering, varying slightly based on the time horizon of the experiments. On the one hand, experiments E1 and E3 have enough T_D for accomplishing D (D is at least 10 times T_D). On the other hand, experiments E2, E4, and E5 have few T_D for accomplishing D (D is at least 16 times T_D). Experiment E6 represents a scenario where T_D is greater than D (D is almost one fourth of T_D). All experiments are capable of satisfying D, using the full

space size B except E6 which uses only one, with remaining time T_r and usually with excess production D_r. Additionally for E1 to E5, the unit cost (Mean reward/D) has a decreasing behavior as demand rises, due to economies of scale. E6 has a particular cost behavior, as only one configuration is needed and the agent does not need to purchase more than one configuration.

Table 3. Experiments performed to validate the RL-based methodology.

Exp.	D	T_D	Cost	D_r	T_r	Purchased MFGs	Cost per part
E1	2000	150	46432	-4	44	9	23.17
E2	2000	80	93806	-25	3	10	46.32
E3	1000	100	28573	0	7	10	28.57
E4	1000	48	63545	-5	3	10	63.22
E5	400	24	55438	-7	1	10	136.21
E6	24	100	2589	-1	79	1	103.56

Given the stochastic nature of the proposed environment, the decision-making procedure for E3, using RL, is shown in Figure 5. E3 is chosen as it shows purchasing decisions throughout T_D. As seen from the figure, the decisions to buy manufacturing configurations are not done at once. The trained agent learns to buy the necessary manufacturing configurations only when it decides that demand will not be satisfied with current options demonstrating that sequential decision-making algorithms are necessary for robust optimal MFG selection problems.

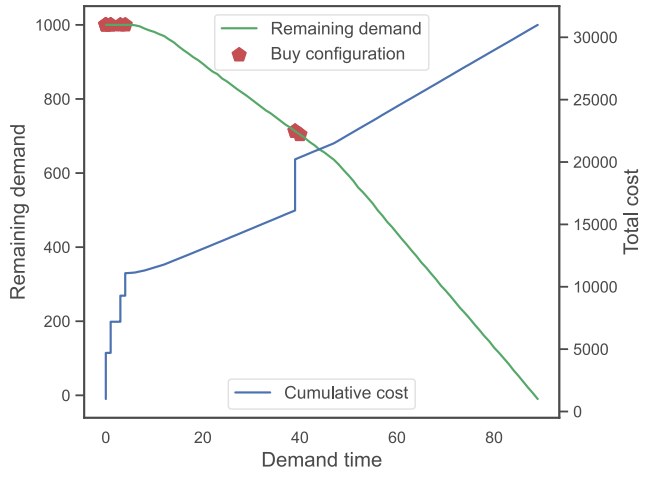

Fig. 5. The sequential decision making process by trained agent for $D = 1000$ and $T_D = 100$

4. Summary and conclusions

Given the uncertainties and changes in modern manufacturing environment, the RL methodology is designed to solve demand satisfaction problems in manufacturing, building upon the concepts of sequential decision making and optimization. As a second contribution, the demand satisfaction problem in manufacturing has been formulated and an RL environment for selecting the optimal MFG in manufacturing has been defined based on the problem formulation.

The RL-based methodology has been validated using a machining process planning case study, involving turning, drilling and milling operations. The machines required to provide the desired shape of the product of the case study have been used to define five simulated assets, as well as the normalized manufacturing data required for the RL agent training. Drawing from the assets, five machine configurations have been presented as being capable of satisfying demand D. Thereafter, the RL agent has been trained to solve the problem using six experiments. From the preliminary results of the validation tests, it has been proven that the methodology is able to obtain solutions that are robust to stochastic noise. Future work will involve validating the approach using real-world asset manufacturing data, as well as considering future demand periods.

Acknowledgements

This project has received funding from the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement No 814078 and by the Department of Education, Universities and Research of the Basque Government under the projects Ikerketa Taldeak (Grupo de Ingeniería de Software y Sistemas IT1519-22 and Grupo de investigación de Mecanizado de Alto Rendimiento IT1443-22).

References

[1] K. K. Goyal, P. Jain, M. Jain, Optimal configuration selection for reconfigurable manufacturing system using NSGA II and TOPSIS, Int. J. Prod. Res. 50 (15) (2012) 4175–4191.
[2] Y. Koren, M. Shpitalni, Design of reconfigurable manufacturing systems, J. Manuf. Syst. 29 (4) (2010) 130–141.
[3] A. M. Youssef, H. A. ElMaraghy, Optimal configuration selection for reconfigurable manufacturing systems, Int. J. Flex. Manuf. Syst. 19 (2007) 67–106.
[4] S. K. Moghaddam, M. Houshmand, O. Fatahi Valilai, Configuration design in scalable reconfigurable manufacturing systems (RMS); a case of single-product flow line (SPFL), Int. J. Prod. Res. 56 (11) (2018) 3932–3954.
[5] C. A. B. Diaz, T. Aslam, A. H. Ng, Optimizing Reconfigurable Manufacturing Systems for Fluctuating Production Volumes: A Simulation-Based Multi-Objective Approach, IEEE Access 9 (2021) 144195–144210.
[6] P. Dayan, Y. Niv, Reinforcement learning: the good, the bad and the ugly, Curr. Opin. Neurobiol. 18 (2) (2008) 185–196.
[7] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, O. Klimov, Proximal policy optimization algorithms, arXiv preprint arXiv:1707.06347 (2017).
[8] Z. Ni, S. Paul, X. Zhong, Q. Wei, A reinforcement learning approach for sequential decision-making process of attacks in smart grid, in: 2017 IEEE SSCI, 2017, pp. 1–8.
[9] E. Pednault, N. Abe, B. Zadrozny, Sequential cost-sensitive decision making with reinforcement learning, in: Proc. 8th ACM SIGKDD Int. Conf. KDD, 2002, pp. 259–268.
[10] E. Liang, R. Liaw, R. Nishihara, P. Moritz, R. Fox, K. Goldberg, J. Gonzalez, M. Jordan, I. Stoica, RLlib: Abstractions for Distributed Reinforcement Learning, in: J. Dy, A. Krause (Eds.), Proc. 35th ICML, Vol. 80 of Proceedings of Machine Learning Research, PMLR, 2018, pp. 3053–3062.
[11] C. H. Lauro, L. C. Brandão, D. Baldo, R. A. Reis, J. P. Davim, Monitoring and processing signal applied in machining processes–A review, Measurement 58 (2014) 73–86.