Offline Reinforcement Learning for Learning to Dispatch for Job Shop Scheduling

Jesse van Remmerden$^{1[0009-0005-1966-6907]}$, Zaharah Bukhsh$^{1[0000-0003-3037-8998]}$, and Yingqian Zhang$^{1[0000-0002-5073-0787]}$

Technical University of Eindhoven, Groene Loper 3, 5612 AE Eindhoven, The Netherlands

Abstract. The Job Shop Scheduling Problem (JSSP) is a complex NP-hard combinatorial optimization problem where jobs must be scheduled on machines to minimize the makespan, which is the total processing time. Traditional methods, such as Constraint Programming (CP), produce high-quality solutions but struggle with scalability. In contrast, heuristic approaches like Priority Dispatching Rules (PDRs) offer faster, suboptimal solutions. Recent reinforcement learning (RL) methods for JSSP attempt to learn effective PDRs but suffer from inefficiencies when trained online due to their inability to leverage existing high-quality solutions.

We introduce Offline-LD, a novel offline reinforcement learning approach to learn dispatching rules for JSSP using pre-existing solution data generated from CP. Offline-LD utilizes Conservative Q-learning (CQL) with two advanced Q-learning methods (mQRDQN and discrete mSAC) adapted for maskable action spaces. Our experiments show that Offline-LD outperforms traditional online RL methods, generalizing well across various problem sizes. Additionally, training on noisy datasets, instead of expert datasets, significantly enhances the quality of dispatching policies by providing counterfactual information, leading to more robust solutions.

Keywords: Job Shop Scheduling Problem · Reinforcement Learning · Offline Reinforcement Learning.

1 Introduction

The Job Shop Scheduling Problem (JSSP) involves optimally scheduling jobs on machines to minimize completion times, a computationally complex task. Classical approaches like CP produce high-quality schedules but fail to scale with larger instances. Heuristic methods such as Priority Dispatching Rules (PDR) [6] are significantly faster and can scale to larger instance sizes; however, the found solutions are of lower quality.

Recent developments in reinforcement learning (RL) for JSSP have focussed on learning dispatching rules [7]. However, existing online RL approaches for JSSP must be trained from scratch using environments where they train through trial and error. This is because online RL methods cannot use pre-existing solutions generated by methods such as CP. Our work introduces Offline-LD, an offline RL approach for JSSP that can leverage pre-existing data to train dispatching policies without requiring any further online training.

2 Method

Offline-LD applies Conservative Q-learning (CQL) [4] to train dispatching policies for JSSP using offline datasets. We furthermore introduce two Q-learning methods for maskable action spaces, namely **d-mSAC**, Discrete Maskable Soft Actor-Critic, a maskable variant of discrete SAC [1], and **mQRDQN**, Maskable Quantile Regression DQN, a maskable variant of QRDQN [2].

We generated two types of datasets using Constraint Programming (CP) [5], an **expert** dataset and a **noisy** dataset. The expert dataset contains the original trajectory generated by the CP solutions and are either optimal or near-optimal. The noisy dataset is generated using the same solutions as the expert dataset; however, we introduce noise into the solutions. Therefore, the noisy dataset is more diverse since it contains high-reward and low-reward trajectories. These low-reward trajectories teach Offline-LD what actions not to take.

A key contribution of Offline-LD is reward normalization. The optimal makespan, the total processing time, can vary significantly between instances in JSSP, even if they are the same size. Therefore, the reward structure is also harder to learn. By normalizing based on the optimal makespan found in the expert dataset, we ensure that the rewards across different instances are comparable. This normalizing improves the results significantly.

3 Results

We evaluated Offline-LD on both generated and benchmark JSSP instances of varying sizes. Offline-LD was compared with traditional methods (PDR [6], L2D [7], and Behavioral Cloning [3]) across 100 instances for each problem size. Each instance was solved using both mQRDQN and d-mSAC, with training conducted on datasets generated with CP.

Offline-LD outperformed L2D and baseline methods in four out of five problem sizes. Notably, it achieved comparable or better results on benchmark instances, highlighting its generalization capabilities. Including noisy datasets improved the policy's robustness and stability, further demonstrating the importance of diverse training datasets for offline RL.

4 Conclusion

Our proposed Offline-LD method bridges the gap between the scalability of heuristic methods and the quality of solutions from CP by leveraging offline RL. The results demonstrate the efficiency and adaptability of Offline-LD for JSSP. Future work will focus on extending our work to other combinatorial optimization problems and improving the performance for JSSP.

References

1. Christodoulou, P.: Soft actor-critic for discrete action settings. CoRR **abs/1910.07207** (2019), http://arxiv.org/abs/1910.07207
2. Dabney, W., Rowland, M., Bellemare, M.G., Munos, R.: Distributional reinforcement learning with quantile regression. In: Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence. AAAI'18/IAAI'18/EAAI'18, AAAI Press (2018)
3. Kumar, A., Hong, J., Singh, A., Levine, S.: When should we prefer offline reinforcement learning over behavioral cloning? (2022), https://arxiv.org/abs/2204.05618
4. Kumar, A., Zhou, A., Tucker, G., Levine, S.: Conservative q-learning for offline reinforcement learning. In: Proceedings of the 34th International Conference on Neural Information Processing Systems. NIPS '20, Curran Associates Inc., Red Hook, NY, USA (2020)
5. Reijnen, R., van Straaten, K., Bukhsh, Z., Zhang, Y.: Job shop scheduling benchmark: Environments and instances for learning and non-learning methods (2023)
6. Veronique Sels, N.G., Vanhoucke, M.: A comparison of priority rules for the job shop scheduling problem under different flow time- and tardiness-related objective functions. International Journal of Production Research **50**(15), 4255–4270 (2012). https://doi.org/10.1080/00207543.2011.611539, https://doi.org/10.1080/00207543.2011.611539
7. Zhang, C., Song, W., Cao, Z., Zhang, J., Tan, P.S., Xu, C.: Learning to dispatch for job shop scheduling via deep reinforcement learning. In: Proceedings of the 34th International Conference on Neural Information Processing Systems. NIPS '20, Curran Associates Inc., Red Hook, NY, USA (2020)