Proceedings of the 2017 Winter Simulation Conference
W. K. V. Chan, A. D'Ambrogio, G. Zacharewicz, N. Mustafee, G. Wainer, and E. Page, eds.

THE APPLICATION OF ACTOR-CRITIC REINFORCEMENT LEARNING FOR FAB DISPATCHING SCHEDULING

Namyong Kim
Hayong Shin

KAIST
Department of Industrial and Systems Engineering
291, Daehak-ro, Yuseong-gu, Daejeon 34141, KOREA

ABSTRACT

This paper applies Actor-Critic reinforcement learning to control lot dispatching scheduling in reentrant line manufacture model. To minimize the Work-In-Process(WIP) and Cycle Time(CT), the lot dispatching policy is directly optimized through Actor-Critic algorithm. The results show that the optimized dispatching policy yields smaller average WIP and CT than traditional dispatching policy such as Shortest Processing Time, Latest-Step-First-Served, and Least-Work-Next-Queue.

1 INTRODUCTION

Most previous research such as Ramírez-Hernández (2009) about applying reinforcement learning to fab scheduling focuses on the Critic only method or Critic based heuristic method. These papers show the good possibility of application of Critic based reinforcement learning for fab scheduling. However, the Critic based methods are indirect in the sense of making only approximation of the critic value function, not optimizing directly policy function.

Actor-Critic algorithm, proposed by Konda(2000), is combination of Actor only and Critic only method. In this algorithm, Critic learns a critic value function from the Bellman equation, and then Actor uses the critic value function to update the actor policy function through stochastic gradient methods. That is, Actor Critic algorithm optimizes directly policy function over policy space.

In this paper, we apply Actor Critic algorithm to optimize the lot dispatching policy, which is to minimize the Work-In-Process(WIP) and Cycle Time(CT) in reentrant line manufacturing model.

2 ACTOR-CRITIC ALOGORITHM IN DISCOUNTED REWARD SETTING

A Markov Decision Process is specified by the tuple M = <S, A, P, C, r>, where S is a finite set of states, A is a finite set of actions, P is a state transition probability, C is a cost function, r is a discount rate.

The state S is defined as follows:

$$S(t) = \left(s_1(t), s_2(t), s_3(t), s_4(t), s_5(t), s_6(t)\right), \quad \forall t \in T \tag{1}$$

where s_i represents the WIP of each processing step i. T represents the decision time when action is made

A policy π is defined as a distribution over actions given states as follows:

$$\pi_\theta(a_i|s) = P[A(t) = lot_i|S(t) = s, \theta = \theta(t)] = \frac{e^{\phi^T(lot_i)\theta(t)}}{\sum_j e^{\phi^T(lot_j)\theta(t)}}, \quad i,j \in Input\ Queue \tag{2}$$

where a_i represent that lot i in input queue is selected to next job.

A Actor basis $\phi(lot_i)$ is defined as follow:

$$\phi^T(lot_i) = (\phi_1(lot_i), \phi_2(lot_i), \phi_3(lot_i)) \tag{3}$$

where $\phi_1(lot_i), \phi_2(lot_i), \phi_3(lot_i)$ indicate the processing time, processing step, and next machine queue level of lot i

A cost C is defined as follows:

$$C(t+1) = WIP_{S(t+1)} - WIP_{S(t)} \quad (4)$$

where C(t+1) represents the difference between total WIP of state S(t+1) and state S(t).

In each decision time, Critic parameters(W) and Actor parameters(θ) are iteratively updated as follows:

$$\delta(t+1) = C(t+1) + \gamma S^T(t+1)W(t) - S^T(t)W(t) \quad (5\text{-}1)$$

$$E(t) = \lambda\gamma E(t-1) + S(t) \quad (5\text{-}2)$$

$$W(t+1) = W(t) + \alpha_{w,t}\delta_t E(t) \quad (5\text{-}3)$$

$$\theta(t+1) = \theta(t) + \alpha_{\theta,t}\delta_t \nabla_\theta \log \pi_\theta(t) \quad (5\text{-}4)$$

where $\lambda \in [0,1)$ is the eligibility rate, $\alpha_{w,t}$ is the critic learning rate, $\alpha_{\theta,t}$ is the actor learning rate. In real experiment, ADAM, proposed by Kingma(2014), was used to update the actor parameters instead of (5-4).

3 EXPERIMENT RESULTS

The reentrant line manufacturing model used in this experiment composes of three machine groups: G1, G2, G3. Each machine group has two machines, which share the queue. The setup depending on the change of processing step occurs in G1. The preventive maintenance occurs in G1 and G2. There are three lot type: A, B, C. All lot has same arrival rate and same processing steps: G1-G2-G3-G2-G1-G3, but each lot type has different deterministic processing time. Queue limit and transport time are not considered in this model.

The dispatching policies over each machine group were optimized through 100,000 hours simulation. To evaluate Actor-Critic based Dispatching Policy(ACDP), 50 scenarios of which run time is 100,000 hours were ran. Shortest Processing Time(SPT), Latest-Step-First-Served(LSFS), and Least-Work-Next-Queue(LWNQ), are used benchmark dispatching policy. The statics of average WIP of lot and CT are listed with the corresponding 95% confidence interval.

Dispatching Policy	Work-In-Process(lot number)		Cycle Time(hours)	
	Average	95% confidence interval	Average	95% confidence interval
ACDP	22.32	[22.07, 22.57]	15.63	[15.45, 15.80]
LSFS	25.02	[24.76, 25.29]	17.52	[17.33, 17.70]
SPT	25.55	[25.26, 25.84]	17.89	[17.68, 18.09]
LWNQ	28.47	[28.15, 28.79]	19.93	[19.71, 20.15]

Table 1. The experiment results of 50 scenarios

The results show that the ACDP yields minimum average WIP and CT. The ACDP is statistically better than benchmark dispatching policies. The average WIP and CT of ACDP are smaller 10.8% than that of LSFS, which is second best dispatching policy.

4 ACKNOWLEDGEMENT

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Science, ICT & Future Planning (2017R1A2B4006290)

REFERENCES

Kinga, D., and J. B. Adam. 2015. A Method for Stochastic Optimization. International Conference on Learning Representations (ICLR).

Konda, V. R., and J. N. Tsitsiklis. 2000. Actor-Critic Algorithms. Advances in neural information processing systems.

Ramírez-Hernández, J. A., and E. Fernandez. 2009. A Simulation-Based Approximate Dynamic Programming Approach for the Control of the Intel Mini-Fab Benchmark Model. Simulation Conference (WSC), Proceedings of the 2009 Winter.