# 강화학습을 활용한 공급망 관리 최적화

김봉석† 황유정†† 심민규†††
서울과학기술대학교 데이터사이언스 학과†,†††
서울과학기술대학교 산업공학과††
22510105@seoultech.ac.kr†, 20102128@seoultech.ac.kr††, mksim@seoultech.ac.kr*†††

# Optimizing Supply Chain Management Using Reinforcement Learning

Bong-Seok Kim† Yoo-Jung Hwang†† Min Kyu Sim†††
Department of Data Science, Seoul National University of Science and Technology†,†††
Department of Industrial Engineering, Seoul National University of Science and Technology††

**요 약**

공급망 관리 능력은 기업의 경쟁력과 직결되는 중요 요소이다. 기존의 연구 방식은 고객 수요가 통계적 분포를 통해 미리 알려져 있으며 안정적인(stationary) 수요를 갖는다고 가정하며, 이를 다양한 최적화 기법을 통해 해를 근사한다. 하지만 실세계에서 고객 수요는 미리 알 수 없으며, 시간에 따라 변화하기 때문에 명확한 한계점이 존재한다. 따라서 본 연구는 강화학습을 통해 적응형 공급망 관리 정책을 찾는 것을 목적으로 하며, 이를 마코프 결정과정으로 정의한다. 또한 정의된 문제를 정책기반 강화학습 알고리즘을 사용하여 푼다. 본 연구에서 제안하는 최적 정책은 전통적인 재고 관리 모형인 Economic order quantity에 비해 약 9~24% 정도 증가된 수익을 기대할 수 있다.

## 1. 서 론

기업 운영의 전략으로 시계열 변동을 가진 수요의 변화를 감지하고 이에 대해 유연하게 반응할 수 있는 운영 전략의 수립은 매우 중요한 요소이다.[1]. 기존 공급망 관리와 관련한 대부분의 연구는 고객 수요의 비교적 안정적인(stationary) 정확한 분포 형태 등이 사전에 알려져 있다고 가정한다. 그 예시로 전통적인 재고 관리 모형인 Economic order quantity(EOQ) 등을 들 수 있다[2]. EOQ 모형은 주문 비용과 단위당 재고유지비용의 합계가 최소로 되는 최적의 주문 수량이 결정되는 모델을 말한다. 이 모형은 재고관리 의사결정에 필요한 측정값을 빠르고 쉽게 얻기 위한 계량적인 모형으로 널리 이용되었다. 또한 기계학습을 이용한 연구는 수요예측/자재조달과 같은 지도학습이 87%로 대부분을 차지한다[3]. 하지만 이를 이용하기 위해서는 정답 값이 필요하며, 이는 관리자의 경험에 의한 판단에 의존한다는 한계점이 존재한다. 이처럼 공급망 관리와 관련된 대부분의 연구는 고객 수요가 일정하며 사전에 이미 알려져 있다고 가정하며 다양한 제약조건이 추가되는 경우 현실에서 적용하기 어렵다. 따라서 본 연구에서는 이러한 기존의 한계점을 해결하고자 불확실하고 수요에 대한 정보를 모르는 상황에서 model-free 정책기반 강화학습 기법을 활용하여 최적의 전략을 효과적으로 찾도록 하는 방안을 제안. 본 논문의 구성은 다음과 같다. 2장에서는 강화학습과 관련된 기본 배경을 설명하고, 3장에서는 공급망 관리 문제를 마코프 결정과정으로 정의한다. 4장에서는 제안된 문제를 바탕으로 정책기반 강화학습 기법을 통해 해결하고 결과를 제시한다. 5장에서는 결론을 맺는다.

## 2. 강화학습

강화학습은 기계학습의 한 영역으로 순차적 의사결정 방법(sequential decision making)을 다루는 방법론이다[3]. 강화학습의 프레임워크는 그림 1과 같다. 어떤 환경 안에서 정의된 에이전트가 현재의 상태를 인식하여, 선택 가능한 행동들 중 누적 보상을 최대화하는 최적 행동 혹은 행동 순서를 학습하는 것이 목표이다.

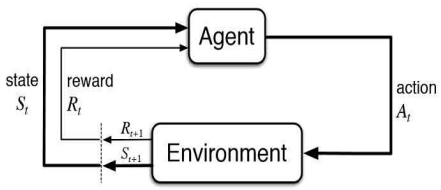


그림 1. 강화학습 프레임워크

### 2.1 마코프 결정과정

강화학습 분야에서 환경은 수학적으로 마코프 결정과정(Markov Decision Process, MDP)으로 표현된다. MDP의 근간을 이루는 마코프 특성(Markovian property)은, 미래의 결과는 과거의 상태에 관계없이, 현재 상태에 의해 결정된다는 것이다. 이러한 조건 속에서, 어떤 상태에서든지 향후 기대되는 누적 보상 값을 최대화하는 최적 행동을 찾는것이 목표이다. MDP는 다음과 같이 흔히, $<S, A, R, r, P>$로 5-tuple로 구성된다. $S$는 상태(state)의 유한 집합을 의미하며, 현재 환경의 상태를 묘사하는 정보로 생각할 수 있다. $A$는 현재 상태 s에서 의사결정자가 취할 수 있는 행동(action)의 유한 집합이다. $R$은 현재 상태 $s$에서 행동 $a$를 했을 때 얻는 보상의 기댓값의 집합으로 정의할 수 있다. $r$는 감가율(discount factor)을 나타내며, 즉각적으로 받는 보상과 미래에서 얻을 수 있는 보상의 중요도를 설정하는 모수이다. 마지막은 상태전이 확률로 $P[S_{t+1}=s|S_t=s, A_t=a]$로 나타낼 수 있으며, 어떠한 시점 $t$의 상태 $s$에서 행동 $a$를 취했을 때 다음 시점 $t+1$에 상태 $s$로

---

* 교신저자

* 이 성과는 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임 (NRF-2021R1A6A1A03039981 and NRF-2020R1F1A107387912)

전이할 확률을 나타낸다.

## 3. 문제정의

본 연구에서 모델링한 공급망의 구성도는 아래 그림 2와 같다. 공급망은 공장(factory), 공장창고(factory warehouse), 분배창고(distribution warehouse) 3개의 유형으로 구성되어 있으며, 1:1:N의 연결 관계를 형성한다[4]. 또한 물건의 방향은 일 방향으로 흐른다고 가정하였다. 에이전트가 결정해야 할 사항은 다음과 같다. 이윤(profit)을 극대화하기 위해서 첫 번째는 공장에서 얼마나 물품을 생산해야 하는가에 대한 것이며 두 번째는, 공장 창고에서 분배창고로 얼마만큼의 물품을 보내는가에 대한 것인가에 대한 것이다. 이후 장에서는 가정한 수요패턴 MDP구성에 대한 세부내용을 이어 나간다. 이후 사용할 용어들은 아래 표 1,2,3과 같다.

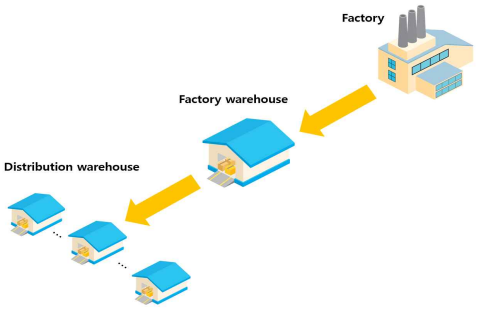


그림 2. 공급망 구성도

표 1. 공장 (factory) 기호정의

| | |
|---|---|
| $a_{f,t}$ | production level at time t |
| $pc_f$ | production cost per unit |

표 2 공장창고 (factory warehouse) 기호정의

| | |
|---|---|
| $f_w$ | factory warehouse |
| $sc_{fw}$ | storage cost at factory warehouse |
| $a_{w_{j,t}}$ | number of products shipped to warehouse $w_j$ at time t |
| $c_{fw}$ | maximum capacity of factory warehouse |
| $q_{f_{w,t}}$ | stock level at time t |

표 3. 분배창고 (distribution warehouse) 기호정의

| | |
|---|---|
| $w_j$ | distribution warehouse j |
| $pc_{w_j}$ | penalty cost at warehouse j |
| $tc_{w_j}$ | transportation cost per unit for warehouse j |
| $c_{w_j}$ | maximum capacity of distribution warehouse |
| $q_{w_{j,t}}$ | current stock level of distribution warehouse at time t |
| $d_{w_{j,t}}$ | demand at warehouse j at time t |
| $p$ | product price for retailer |
| $sc_{w_j}$ | storage cost at warehouse j |

### 3.1 수요 패턴 정의

본 연구에서의 수요가정은 다양하게 구성하여 실험을 진행한 후 기존 고정적 주문(EOQ) 방식과 비교한다. 수요는 그림 3과 같이 크게 3가지 Constant, Linear drop, Seasonality 유형으로 나뉘며 각각의 수요에서 uniform random noise를 추가하여 총 6가지로 실험을 진행한다. 에이전트는 현재 시점의 demand를 관찰할 수 없으며, 오직 과거의 불확실한 수요 정보만을 활용하여 현재 공장에서 얼마만큼 생산하고, 각 창고로 보낼 것인지 결정하여야 한다.

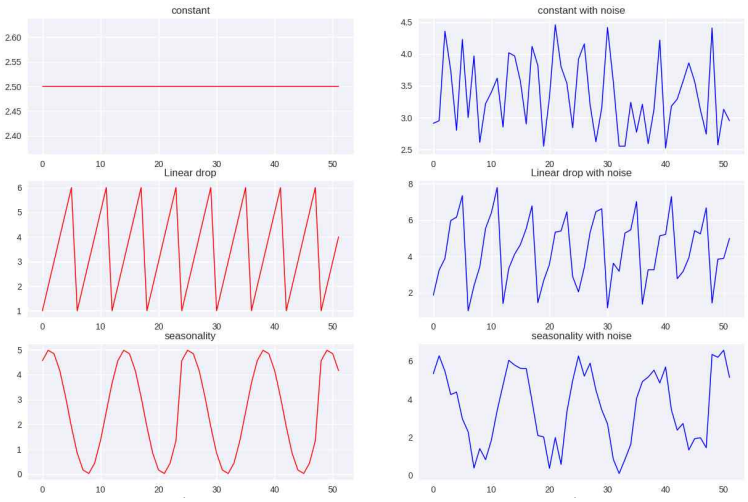


그림 3. 수요패턴

### 3.2 MDP 정의

에이전트가 관찰하는 상태 $S_t$는 $t$시점의 모든 창고 재고 수준과, 과거에 발생했던 수요의 정보를 포함한 벡터이다. 상태는 크게 두 가지 종류의 정보로 구성된다. 첫 번째는 재고에 관한 정보로, 현재 시점 공장의 창고와, 분배창고의 재고 수준을 포함한다. 두 번째는 과거 수요에 대한 패턴 정보를 관찰한다. 이는 과거의 수요의 기록들과 현재 재고 수준만을 보고 행동(action)을 결정함으로써, 일종의 수요 예측능력을 갖게 하기 위함이다.

$$S_t = [q_{fw_t}, q_{w_{1\,.\,t}}, \cdots, q_{w_{f\,.\,t}}, d_{t-1}, d_{t-T}], \text{ where } d_t = [d_{w_{1\,.\,t}}, d_{w_{2\,.\,t}}, \cdots, d_{w,t}]$$

에이전트가 결정할 행동(action)은 크게 두 가지 종류이며, 상태를 입력으로 받아 마찬가지로 벡터 형태로 행동을 내뱉는다. 첫 번째는 $a_{f,t}$로 $t$시점에서 공장에서 얼마만큼 물품을 생산하는지에 대한 것이며, 두 번째는 $a_{w_{j,t}}$로 $t$시점에서 공장 창고에서 분배창고 $j$로 얼마만큼 물품을 보낼 것인가에 대한 것이다.

$$a_t = [a_{f,t}, a_{w_{1,t}}, a_{w_{2,t}}, \cdots, a_{w_{j,t}}]$$

상태전이는 다음과 같이 현재 재고 수준과 수요, 행동 사이에서 발생하는 다음 시점의 상태로 정의한다. 수요 정보의 경우 과거의 수요 정보만 관찰할 수 있으며, sliding window 방식과 같이 과거의 수요패턴 정보의 타임 스텝이 한 칸씩 이동하는 것과 같다.

$$S_{t+1} = [\min(q_{fw,t} + a_{f,t} - \sum_{i}^{W} a_{w_{j,t}}, c_{fw}),$$
$$\min(q_{f1,t} + a_{w_{1,t}} - d_{w1,t,c_{w_j}}, \cdots, \min(q_{w_{j,t}} + a_{w_{j,t-d_{w_{j,t}}}}, c_{w_j})$$
$$d_t, d_{t-1}, \ldots, d_{t-\tau}]$$

에이전트가 매 시점 행동을 결정하고 받을 보상은 다음과 같이 공급망의 profit이며, revenue, production cost, storage cost, transportation cost, penalty cost 5개의 항으로 구성된다.

$$r_t = p\sum_{j=1}^{n} d_{w_{j,t}} + \sum_{j=1}^{n} tc_{w,j} \cdot a_{w_{j,t}} + sc_{fw} \cdot \max(q_{fw,t},0) \cdot \sum_{j=1}^{n} pc_f \cdot a_{f,t} + \sum_{j=1}^{n} pc_{w,j} \cdot \min(q_{w_j},0)$$

각 창고의 재고 수준은 수치적으로 음수 값(-)를 가질 수 있다고 가정하였다. 이는 현재 재고량이 수요 수준을 충족시키지 못한 $\min(q_{w_j},0)$값으로, 현재 재고가 없다는 뜻으로 생각한다. 따라서 storage cost 경우 만약 음수 값(-)인 경우 0으로 처리하기 위해 $\max(q_{fw,t},0)$제약조건을 넣어 준다. 마찬가지로 penalty cost의 경우 양수(+)인 경우 수요를 전부 충족했으므로, 제약조건을 삽입해 처리를 하지 않도록 제약조건을 삽입하였다.

## 4. 결과

실험 결과는 아래의 표 1과 같으며, 6개의 수요패턴에 대해서 각 10번씩 실험을 진행하여 평균 결과를 종합한 것이다. 1 episode 당 52의 time step을 가지며, 학습은 10000의 episode를 진행하였다. 알고리즘은 정책기반 강화학습 알고리즘의 일종인 Twined delayed DDPG(TD3)를 사용하였다[5]. 공급망 네트워크는 간단한 공급망인 1:1:1 네트워크를 사용하여 비교 실험한다.

에이전트는 학습 진행됨에 따라 EOQ 모델에 비해 높은 이윤(profit)을 달성하였다. revenue 관점에서는 EOQ 모델보다 떨어지는 결과를 보였다. 이는 수요가 발생하더라도 상황에 따라 이 수요를 모두 충족하지 않는 것이 이익을 극대화하는 것에 도움이 될 때가 있다는 뜻으로 생각할 수 있다. 비용관점에서 분석해보면, 그림 4와 그림 5와 같이 에이전트는 비용 관점에서 크게 절약을 하여 높은 profit을 달성한 것을 확인할 수 있다. 그림 3의 storage cost 그래프를 보았을 때, 비용 중에서도 특히 저장비용 관점에서 크게 차이가 나는 것을 확인할 수 있다. 이는 필요한 만큼만 생산하여 불필요한 재고는 최소화했기 때문이다. 결론적으로 에이전트는 profit 관점에서 전통적인 EOQ 모델과 대비해 약 9~24%의 증가된 결과를 보인다.

표 1. 실험결과

| Demand | Agent profit | EOQ profit | 성능 차이(%) |
|---|---|---|---|
| constant | 5842.234 | 5036.97 | 9.2 % |
| constant uniform | 8537.6 | 7520.56 | 12 % |
| linear drop | 7896.26 | 6844.00 | 13.4 % |
| linear drop uniform | 10962.42 | 9170.75 | 16.4 % |
| seasonality | 5715.991 | 4511.04 | 22 % |
| seasonality uniform | 8423.516 | 6416.30 | 24 % |

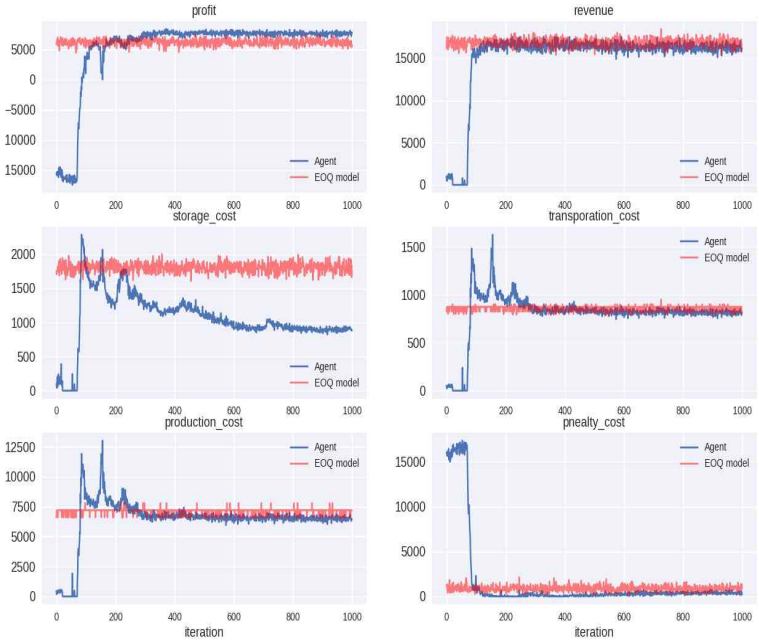


그림 4. Performance curve

## 5. 향후 연구 및 결론

강화학습 에이전트는 스스로 수요패턴을 추론하고 최적 전략을 효과적으로 찾아내었다. 특히 고정적인 경우보다 불확실성이 내재된 수요패턴의 경우 더 효과적인 방법을 제안했다. 이에 따라 실제 산업에서 신제품과 같이 수요패턴을 예측하기 힘들거나 기존 고객의 수요와 관련한 데이터가 없는 경우에도 효과적으로 적응형 재고정책을 제안하며 대응할 수 있을 것이라 기대한다.

## 참고문헌

[1] Chen, F., Drezner, Z., Ryan, J. K., & Simchi-Levi, D. (2000). "Quantifying the bullwhip effect in a simple supply chain: The impact of forecasting, lead times, and information." *Management science*, 46(3), 436-443.

[2] Agarwal, Sachin. "Economic order quantity model: a review." *VSRD International Journal of Mechanical, Civil, Automobile and Production Engineering* 4(12) (2014): 233-236.

[3] Ni, D., Xiao, Z., & Lim, M. K. (2020). A systematic review of the research trends of machine learning in supply chain management. International Journal of Machine Learning and Cybernetics, 11(7), 1463-1482.

[4] LR. Sutton and A. Barto, "Introduction to reinforcement learning," Cambridge, MA 1998

[5] Cakravastia, A., Toha, I. S., & Nakamura, N. (2002). "two-stage model for the design of supply chain networks." *International journal of production economics*, 80(3), 231-248.

[6] Fujimoto, Scott, Herke Hoof, and David Meger. "Addressing function approximation error in actor-critic methods." *International conference on machine learning. PMLR*, 2018.