A Deep Reinforcement Learning Approach to Supply Chain Inventory Management

Francesco Stranieri

Department of Informatics, Systems, and Communication,
University of Milano-Bicocca
Department of Control and Computer Engineering,
Polytechnic University of Turin

Introduction

Supply chain inventory management (SCIM) is a *sequential decision-making problem* consisting of determining the optimal quantity of products to produce at the factory and to ship to different distribution warehouses over a given time horizon. DRL algorithms are rarely applied to the SCIM field, although they can be used to develop near-optimal policies that are difficult, or impossible at worst, to achieve using traditional mathematical methods [1].

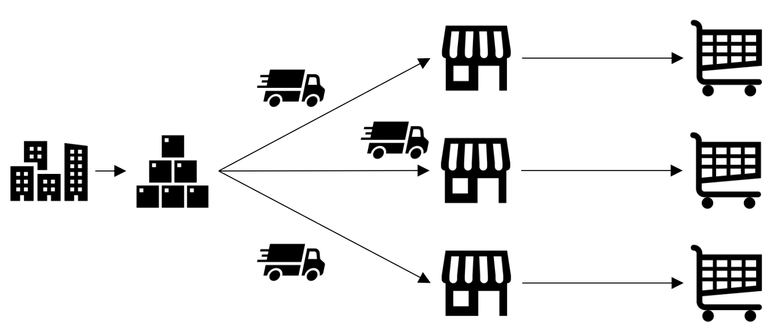

Fig. 1: A divergent two-echelon supply chain consisting of a factory and its warehouse (first echelon), plus three distribution warehouses (second echelon).

In our actual research, a MDP formulation of the SCIM environment is given, which includes a *factory* that can produce various *product types*, a *factory warehouse*, and a certain number of *distribution warehouses*.

MDP Formulation

For the *state vector*, we include all current stock levels for each warehouse and product type, plus the last τ demand values:

$$\mathrm{s}_t = (q_{0,0,t}, \ldots, q_{I,J,t}, d_{t-\tau}, \ldots, d_{t-1}),$$

where $d_{t-1} = (d_{0,1,t-1}, \ldots, d_{I,J,t-1})$.
For the *action vector*, we implement a *continuous action space* (i.e., the neural network generates the action value directly):

$$\mathrm{a}_t = (a_{0,0,t}, \ldots, a_{I,J,t}).$$

Our implementation provides an *independent upper bound* for each action value; for each distribution warehouse, it corresponds to its maximum capacity with respect to each product type ($0 \leq a_{i,j,t} \leq c_{i,j}$), while for the factory to the sum of all warehouses' capacities with regard to each product type ($0 \leq a_{i,0,t} \leq \sum_{j=0}^{J} c_{i,j}$).
To simulate a *seasonal behavior*, we represent the *demand* as a co-sinusoidal function with a stochastic component:

$$d_{i,j,t} = \left\lfloor \frac{d_{max_i}}{2} \left(1 + \cos\left(\frac{4\pi(2ij + t)}{T} \right) \right) + \mathcal{U}\left(0, d_{var_i}\right) \right\rfloor.$$

where d_{max_i} is the maximum demand value for each product type and $\mathcal{U}$ is a random variable uniformly distributed on the support $(0, d_{var_i})$ representing the *uncertainty*.

The DRL algorithms' goal is to *maximize the supply chain profit*. Accordingly, we design the *reward function* as:

$$\mathrm{r}_t = \sum_{j=1}^{J}\sum_{i=0}^{I} p_i \cdot d_{i,j,t} - \sum_{i=0}^{I} z_{i,0} \cdot a_{i,0,t} - \sum_{j=1}^{J}\sum_{i=0}^{I} z_{i,j}^{T} \cdot a_{i,j,t}$$
$$- \sum_{j=0}^{J}\sum_{i=0}^{I} z_{i,j}^{S} \cdot \max(q_{i,j,t}, 0) + \sum_{j=0}^{J}\sum_{i=0}^{I} z_{i}^{P} \cdot p_i \cdot \min(q_{i,j,t}, 0).$$

The first term represents revenues, the second one production costs, while the third one corresponds to transportation costs. The fourth term is the overall storage costs. The last term denotes the penalty costs (in case of *backordering*).
Finally, we define the *state's updating rule* as follows:

$$s_{t+1} = (\min[(q_{0,0,t} + a_{0,0,t} - \sum_{j=1}^{J} a_{0,j,t}), c_{0,0}], \cdots,$$
$$\min\left[\left(q_{I,J,t} + a_{I,J,t} - d_{I,J,t}\right), c_{I,J}\right], d_{t+1-\tau}, \cdots, d_t).$$

It is worth highlighting that the actual demand d_t will not be known until the next time step $t+1$.

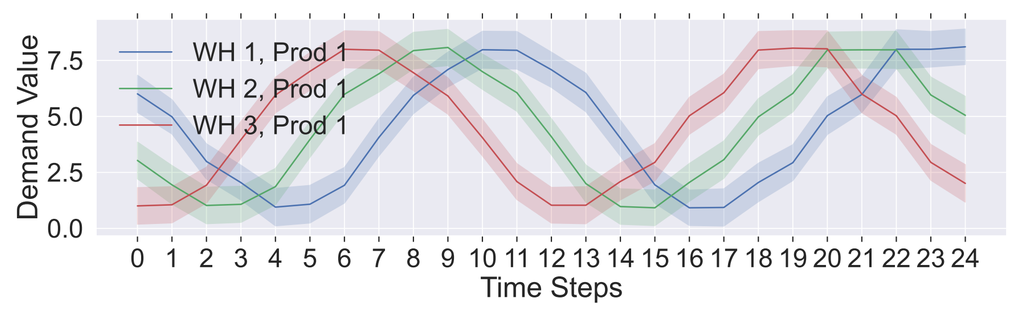

Fig. 2: An instance of the demand behavior considering one product type and three distribution warehouses with $d_{max} = 7$ and $d_{var} = 2$.

Results

Performances achieved by state-of-the-art DRL algorithms are compared with an (s, Q)-policy whose optimal parameters have been set through a *data-driven* Bayesian optimization (BO) approach, and with an *oracle*. Results of numerical experiments demonstrate that the SCIM environment we propose is *effective* in representing states, actions, and rewards; indeed, DRL algorithms have been able to learn nearly optimal policies in all the investigated scenarios. Naturally, our future research will be *extended and improved* in many directions [2].

References

[1] Robert N Boute et al. "Deep reinforcement learning for inventory control: A roadmap". In: *European Journal of Operational Research* (2021).

[2] Yimo Yan et al. "Reinforcement learning for logistics and supply chain management: Methodologies, state of the art, and future opportunities". In: *Transportation Research Part E: Logistics and Transportation Review* 162 (2022), p. 102712.