A Reinforcement Learning Approach to Dual-sourcing Inventory Problem

(ORIE 6590 Final Project)

Tonghua Tian and Xumei Xi
School of Operations Research and Information Engineering
Cornell University
{tt543,xx269}@cornell.edu

April 2021

1 Introduction

It is common practice that companies depend on multiple suppliers for product ordering. We focus on the model where we have a regular supplier and an express supplier. The regular supplier provides a cheaper price point but takes longer to deliver, while the express supplier has faster delivery but requires a higher expense. In our project, we examine the dual-sourcing inventory system and implement a reinforcement learning algorithm to make prudent ordering decisions. Please see the GitHub repo (`https://github.com/xixumei1226/dual-sourcing-rl`) for the implementation.

1.1 MDP Formulation

In this subsection, we formally introduce the Markov decision process (MDP) formulation of the problem. As in the work [XG18], suppose we have a regular supplier R, with a longer lead time L_r and a lower cost c_r, and an express supplier E, with a shorter lead time L_e and a higher cost c_e. We assume $L_r > L_e + 1$ and $c_r < c_e$. Demands are generated as an i.i.d. sequence $\{D_t, t \geq 1\}$, distributed as the nonnegative random variable D. Denote the unit holding and backorder costs by $h > 0$ and $b > 0$ respectively. Let I_t denote the on-hand inventory, and $\mathbf{q}_t^r = \{q_{t-i}^r, i \in [L_r]\}$, $\mathbf{q}_t^e = \{q_{t-i}^e, i \in [L_e]\}$ denote the pipeline vectors of orders placed but not yet delivered with R and E at the start of period t, where q_{t-i}^r, q_{t-i}^e are the orders placed in period $t - i$.

At period t, a sequence of events happen in the following order:

1. The on-hand inventory I_t is observed.
2. New orders q_t^r and q_t^e are placed with R and E.
3. New inventory $q_{t-L_r}^r + q_{t-L_e}^e$ is delivered and added to the on-hand inventory.
4. The demand D_t is realized; the inventory and pipeline vectors are updated.
5. Costs for period t are incurred.

Notice the on-hand inventory is updated according to

$$I_{t+1} = I_t + q_{t-L_r}^r + q_{t-L_e}^e - D_t.$$

The pipeline vectors are updated according to

$$\mathbf{q}_{t+1}^r = (q_{t-L_r+1}^r, \ldots, q_{t-1}^r, q_t^r),$$
$$\mathbf{q}_{t+1}^e = (q_{t-L_e+1}^e, \ldots, q_{t-1}^e, q_t^e).$$

Let C_t be the sum of the ordering cost and holding and backorder costs incurred in time period t:

$$C_t = c_r q_t^r + c_e q_t^e + hI_{t+1}^+ + bI_{t+1}^-.$$

Note that in [XG18], orders placed in period t are charged in period $t + L_e$. Here we use a different counting method to simplify the notations for the embedded MDP. This has no effect on the problem considered.

An admissible policy π consists of a sequence of deterministic measurable functions $\{f_t^\pi, t \geq 1\}$ from $\mathbb{R}^{L_r+L_e+1}$ to $\mathbb{R}_+^2$. Specifically, the new orders placed in period t are given by $(q_t^r, q_t^e) = f_t^\pi(\mathbf{q}_t^r, \mathbf{q}_t^e, I_t)$. Let Π denote the family of all admissible policies. The cost under a policy π is denoted by C_t^π. We aim to minimize the long-run average cost

$$C(\pi) = \limsup_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}[C_t^\pi].$$

Assume the demands follow Poisson distribution: $D \sim \text{Pois}(\lambda)$, where $\lambda > 0$. Furthermore, assume the orders can only take integer values. Then the above process can be formulated as a discrete MDP. At period t, let $s_t = (\mathbf{q}_t^r, \mathbf{q}_t^e, I_t)$ be the state of the system, and let $a_t = (q_t^r, q_t^e)$ be the action taken. The state space $\mathcal{S}$ and action space $\mathcal{A}$ are given by

$$\mathcal{S} = \mathbb{Z}_+^{L_r} \times \mathbb{Z}_+^{L_e} \times \mathbb{Z}, \quad \mathcal{A} = \mathbb{Z}_+^2.$$

Note that C_t is a function of s_{t+1} instead of s_t. So the reward of step t is actually received at step $t-1$:

$$r(s_t, a_t) = -C_{t-1} = -c_r q_{t-1}^r - c_e q_{t-1}^e - hI_t^+ - bI_t^-.$$

Define the function $g : \mathbb{R}^{L_r+L_e+1} \times \mathbb{R}^2 \to \mathbb{R}^{L_r+L_e+1}$ as

$$g(x_1, \ldots, x_{L_r}, y_1, \ldots, y_{L_e}, z, a_1, a_2) = (x_2, \ldots, x_{L_r}, a_1, y_2, \ldots, y_{L_e}, a_2, z + x_1 + y_1).$$

Then

$$s_{t+1} = g(s_t, a_t) - (0, \ldots, 0, D_t).$$

Hence the transition probabilities are given by

$$\mathbb{P}(s_{t+1} = g(s_t, a_t) - (0, \ldots, 0, k) \mid s_t, a_t) = \frac{\lambda^k e^{-\lambda}}{k!}, \quad k = 0, 1, \ldots.$$

1.2 Prior Works

In this subsection, we briefly review some prior works on the dual-sourcing inventory systems. Researchers have been investigating the dual-sourcing problem for many years due to its practicality as well as its intractability. Earlier works have already showed that when the lead time difference is exactly one, order-up-to policies are optimal. However, once we step into the regime where the lead time difference grows larger, such policies fall short. In our project, we focus on the general case where the lead time difference is relatively large. Over the years, people have developed various heuristic policies to better approximate the optimum. The work [VSW08] proposed the idea of dual-index (DI) policies which have two order-up-to levels for the two suppliers. Under such policies, we keep track of two inventory positions and make orders in an effort to hold the corresponding inventory positions up to certain levels. Another simple and intuitive policy is the tailored base-surge (TBS) policy proposed in the work [AM10]. With a TBS policy, a constant order is placed at the regular source in each period to meet a base level of demand, while the orders we place at the express source follow an order-up-to rule to manage demand surges. As showed in the work [KKM11], TBS policies are comparable to DI policies in practice, and outperform DI policies for some problem instances, especially with an increase in the lead time difference. It has also been prove theoretically in the paper [XG18] that when the lead time of the express source is fixed, a simple TBS policy is asymptotically optimal as the lead time of the regular source increases.

1.3 Project Goals

In this subsection, we state our goals for the project.

1. Develop practical reinforcement learning algorithm to approximate the optimal policy in the dual-sourcing inventory systems.

2. Compare our performance with the other team and heuristic policies like the TBS policy.

3. Gain insights by potentially analyzing the intrinsic structure within the output policy.

4. Acknowledge the challenges in solving the problem and the limitation of our approach. Indicate possible future directions.

References

[AM10] Gad Allon and Jan A. Van Mieghem. Global dual sourcing: Tailored base-surge allocation to near- and offshore production. *Management Science*, 56(1):110–124, 2010.

[KKM11] Steffen Klosterhalfen, Gudrun Kiesmüller, and Stefan Minner. A comparison of the constant-order and dual-index policy for dual sourcing. *International Journal of Production Economics*, 133(1):302–311, 2011. Leading Edge of Inventory Research.

[VSW08] Senthil Veeraraghavan and Alan Scheller-Wolf. Now or later: A simple policy for effective dual sourcing in capacitated systems. *Operations Research*, 56(4):850–864, 2008.

[XG18] Linwei Xin and David A. Goldberg. Asymptotic optimality of tailored base-surge policies in dual-sourcing inventory systems. *Management Science*, 64(1):437–452, 2018.