

# A Deep Q-Network for the Beer Game: Reinforcement Learning for Inventory Optimization

Afshin Oroojlooyjadid, MohammadReza Nazari, Lawrence V. Snyder, Martin Takáč

Department of Industrial and Systems Engineering, Lehigh University, Bethlehem, PA 18015, {oroojlooy, mon314, larry.snyder}@lehigh.edu, Takac.MT@gmail.com

*Problem definition:* The beer game is a widely used game that is played in supply chain management classes to demonstrate the bullwhip effect and the importance of supply chain coordination. The game is a decentralized, multi-agent, cooperative problem that can be modeled as a serial supply chain network in which agents choose order quantities while cooperatively attempting to minimize the network's total cost, even though each agent only observes its own local information. *Academic/practical relevance:* Under some conditions, a base-stock replenishment policy is optimal. However, in a decentralized supply chain in which some agents act irrationally, there is no known optimal policy for an agent wishing to act optimally.

*Methodology:* We propose a reinforcement learning (RL) algorithm, based on deep Q-networks, to play the beer game. Our algorithm has no limits on costs and other beer game settings. Like any deep RL algorithm, training can be computationally intensive, but this can be performed ahead of time; the algorithm executes in real time when the game is played. Moreover, we propose a transfer-learning approach so that the training performed for one agent can be adapted quickly for other agents and settings. *Results:* When playing with teammates who follow a base-stock policy, our algorithm obtains near-optimal order quantities. More importantly, it performs significantly better than a base-stock policy when other agents use a more realistic model of human ordering behavior. Finally, applying transfer-learning reduces the training time by one order of magnitude. *Managerial implications:* This paper shows how artificial intelligence can be applied to inventory optimization. Our approach can be extended to other supply chain optimization problems, especially those in which supply chain partners act in irrational or unpredictable ways.

*Key words:* Inventory Optimization, Reinforcement Learning, Beer Game

*History:*

## 1. Introduction

The beer game consists of a serial supply chain network with four agents—a retailer, a warehouse, a distributor, and a manufacturer—who must make independent replenishment decisions with limited information. The game is widely used in classroom settings to demonstrate the *bullwhip effect*, a phenomenon in which order variability increases as one moves upstream in the supply chain, as well as the importance of communication and coordination in the supply chain. The bullwhip effect occurs for a number of reasons, some rational (Lee et al. 1997) and some behavioral (Sterman 1989). It is an inadvertent outcome that emerges when the players try to achieve the stated purpose of the game, which is to minimize costs. In this paper, we are interested not in the bullwhip effect but in the stated purpose, i.e., the minimization of supply chain costs, which underlies the decision making in every real-world supply chain. For general discussions of the bullwhip effect, see, e.g., Lee et al. (2004), Geary et al. (2006), and Snyder and Shen (2019).

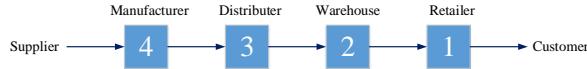
The agents in the beer game are arranged sequentially and numbered from 1 (retailer) to 4 (manufacturer), respectively. (See Figure 1.) The retailer node faces a stochastic demand from its customer, and the manufacturer node has an unlimited source of supply. There are deterministic transportation lead times ( $l^{tr}$ ) imposed on the flow of product from upstream to downstream, though the actual lead time is stochastic due to stockouts upstream; there are also deterministic information lead times ( $l^{in}$ ) on the flow of information from downstream to upstream (replenishment orders). Each agent may have nonzero shortage and holding costs.

In each period of the game, each agent chooses an order quantity  $q$  to submit to its predecessor (supplier) in an attempt to minimize the long-run system-wide costs,

$$\sum_{t=1}^T \sum_{i=1}^4 c_h^i (IL_t^i)^+ + c_p^i (IL_t^i)^-, \quad (1)$$

where  $i$  is the index of the agents;  $t = 1, \dots, T$  is the index of the time periods;  $T$  is the time horizon of the game (which is often unknown to the players);  $c_h^i$  and  $c_p^i$  are the holding and shortage cost coefficients, respectively, of agent  $i$ ; and  $IL_t^i$  is the inventory level of agent  $i$  in period  $t$ . If  $IL_t^i > 0$ ,

**Figure 1** Generic view of the beer game network.



then the agent has inventory on-hand, and if  $IL_t^i < 0$ , then it has backorders. The notation  $x^+$  and  $x^-$  denotes  $\max\{0, x\}$  and  $\max\{0, -x\}$ , respectively.

The standard rules of the beer game dictate that the agents may not communicate in any way, and that they do not share any local inventory statistics or cost information with other agents until the end of the game, at which time all agents are made aware of the system-wide cost. In other words, each agent makes decisions with only partial information about the environment while also cooperates with other agents to minimize the total cost of the system. According to the categorization by Claus and Boutilier (1998), the beer game is a decentralized, independent-learners (ILs), multi-agent, cooperative problem.

The beer game assumes the agents incur holding and stockout costs but not fixed ordering costs, and therefore the optimal inventory policy is a *base-stock policy* in which each stage orders a sufficient quantity to bring its inventory position (on-hand plus on-order inventory minus back-orders) equal to a fixed number, called its base-stock level (Clark and Scarf 1960). When there are no stockout costs at the non-retailer stages, i.e.,  $c_p^i = 0$ ,  $i \in \{2, 3, 4\}$ , the well known algorithm by Clark and Scarf (1960) provides the optimal base-stock levels. To the best of our knowledge, there is no algorithm to find the optimal base-stock levels for general stockout-cost structures. More significantly, when some agents do not follow a base-stock or other rational policy, the form and parameters of the optimal policy that a given agent should follow are unknown.

In this paper, we propose an extension of deep Q-networks (DQN) to solve this problem. Our algorithm is customized for the beer game, but we view it also as a proof-of-concept that DQN can be used to solve messier, more complicated supply chain problems than those typically analyzed in the literature. The remainder of this paper is as follows. Section 2 provides a brief summary of the relevant literature and our contributions to it. The details of the algorithm are introduced in Section 3. Section 4 provides numerical experiments, and Section 5 concludes the paper.

## 2. Literature Review

### 2.1. Current State of Art

The beer game consists of a serial supply chain network. Under the conditions dictated by the game (zero fixed ordering costs, no ordering capacities, linear holding and backorder costs, etc.), a base-stock policy is optimal at each stage (Lee et al. 1997). If the demand process and costs are stationary, then so are the optimal base-stock levels, which implies that in each period (except the first), each stage simply orders from its supplier exactly the amount that was demanded from it. If the customer demands are i.i.d. random and if backorder costs are incurred only at stage 1, then the optimal base-stock levels can be found using the exact algorithm by Clark and Scarf (1960).

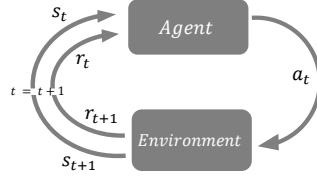
There is a substantial literature on the beer game and the bullwhip effect. We review some of that literature here, considering both independent learners (ILs) and joint action learners (JALs) (Claus and Boutilier 1998). (ILs have no information about the other agent's current states, whereas JALs may share such information.) For a more comprehensive review, see Devika et al. (2016). See Martinez-Moyano et al. (2014) for a thorough history of the beer game.

In the category of ILs, Mosekilde and Larsen (1988) develop a simulation and test different ordering policies, which are expressed using a formula that involves state variables such as the number of anticipated shipments and unfilled orders. They assume the customer demand is 4 in each of the first four periods, and then 7 per period for the remainder of the horizon. Sterman (1989) uses a similar version of the game in which the demand is 8 after the first four periods. (Hereinafter, we refer to this demand process as  $C(4,8)$  or the *classic* demand process.) Also, he do not allow the players to be aware of the demand process. He proposes a formula (which we call the *Sterman formula*) to determine the order quantity based on the current backlog of orders, on-hand inventory, incoming and outgoing shipments, incoming orders, and expected demand. His formula is based on the anchoring and adjustment method of Tversky and Kahneman (1979). In a nutshell, the Sterman formula attempts to model the way human players over- or under-react to situations they observe in the supply chain such as shortages or excess inventory. Note that Sterman's formula is not an attempt to optimize the order quantities in the beer game; rather, it

is intended to model typical human behavior. There are multiple extensions of Sterman's work. For example, Strozzi et al. (2007) considers the beer game when the customer demand increases constantly after four periods and proposes a genetic algorithm (GA) to obtain the coefficients of the Sterman model. Subsequent behavioral beer game studies include Croson and Donohue (2003) and Croson and Donohue (2006a).

Most of the optimization methods described in the first paragraph of this section assume that every agent follows a base-stock policy. The hallmark of the beer game, however, is that players do not tend to follow such a policy, or *any* policy. Often their behavior is quite irrational. There is comparatively little literature on how a given agent should optimize its inventory decisions when the other agents do not play rationally (Sterman 1989, Strozzi et al. 2007)—that is, how an individual player can best play the beer game when her teammates may not be making optimal decisions.

Some of the beer game literature assumes the agents are JALs, i.e., information about inventory positions is shared among all agents, a significant difference compared to classical IL models. For example, Kimbrough et al. (2002) propose a GA that receives a current snapshot of each agent and decides how much to order according to the  $d + x$  rule. In the  $d + x$  rule, agent  $i$  observes  $d_t^i$ , the received demand/order in period  $t$ , chooses  $x_t^i$ , and then places an order of size  $a_t^i = d_t^i + x_t^i$ . In other words,  $x_t^i$  is the (positive or negative) amount by which the agent's order quantity differs from his observed demand. Giannoccaro and Pontrandolfo (2002) consider a beer game with three agents with stochastic shipment lead times and stochastic demand. They propose a RL algorithm to make decisions, in which the state variable is defined as the three inventory positions, which each are discretized into 10 intervals. The agents may use any actions in the integers on  $[0, 30]$ . Chaharsooghi et al. (2008) consider the same game and solution approach except with four agents and a fixed length of 35 periods for each game. In their proposed RL, the state variable is the four inventory positions, which are each discretized into nine intervals. Moreover, their RL algorithm uses the  $d + x$  rule to determine the order quantity, with  $x$  restricted to be in  $\{0, 1, 2, 3\}$ . Note that these RL algorithms assume that real-time information is shared among agents, whereas ours adheres to the typical beer-game assumption that each agent only has local information.

**Figure 2** A generic procedure for RL.

## 2.2. Reinforcement Learning

Reinforcement learning (Sutton and Barto 1998) is an area of machine learning that has been successfully applied to solve complex sequential decision problems. RL is concerned with the question of how a software agent should choose an action to maximize a cumulative reward. RL is a popular tool in telecommunications, robot control, and game playing, to name a few (see Li (2017)).

Consider an agent that interacts with an environment. In each time step  $t$ , the agent observes the current state of the system,  $s_t \in \mathbb{S}$  (where  $\mathbb{S}$  is the set of possible states), chooses an action  $a_t \in \mathbb{A}(s_t)$  (where  $\mathbb{A}(s_t)$  is the set of possible actions when the system is in state  $s_t$ ), and gets reward  $r_t \in \mathbb{R}$ ; and then the system transitions randomly into state  $s_{t+1} \in S$ . This procedure is known as a *Markov decision process* (MDP) (see Figure 2), and RL algorithms can be applied to solve this type of problem.

The matrix  $P_a(s, s')$ , which is called the *transition probability matrix*, provides the probability of transitioning to state  $s'$  when taking action  $a$  in state  $s$ , i.e.,  $P_a(s, s') = \Pr(s_{t+1} = s' | s_t = s, a_t = a)$ . Similarly,  $R_a(s, s')$  defines the corresponding reward matrix. In each period  $t$ , the decision maker takes action  $a_t = \pi_t(s)$  according to a given policy, denoted by  $\pi_t$ . The goal of RL is to maximize the expected discounted sum of the rewards  $r_t$ , when the system runs for an infinite horizon. In other words, the aim is to determine a policy  $\pi : \mathbb{S} \rightarrow \mathbb{A}$  to maximize  $\sum_{t=0}^{\infty} \gamma^t E[R_{a_t}(s_t, s_{t+1})]$ , where  $a_t = \pi_t(s_t)$  and  $0 \leq \gamma < 1$  is the discount factor. For given  $P_a(s, s')$  and  $R_a(s, s')$ , the optimal policy can be obtained through dynamic programming or linear programming (Sutton and Barto 1998).

Another approach for solving this problem is *Q-learning*, a type of RL algorithm that obtains the *Q-value* for any  $s \in S$  and  $a = \pi(s)$ , i.e.  $Q(s, a) = \mathbb{E}[r_t + \gamma r_{t+1} + \gamma^2 r_{t+2} + \dots | s_t = s, a_t = a; \pi]$ .

The Q-learning approach starts with an initial guess for  $Q(s, a)$  for all  $s$  and  $a$  and then proceeds to update them based on the iterative formula

$$Q(s_t, a_t) = (1 - \alpha_t)Q(s_t, a_t) + \alpha_t \left( r_{t+1} + \gamma \max_a Q(s_{t+1}, a) \right), \forall t = 1, 2, \dots, \quad (2)$$

where  $\alpha_t$  is the learning rate at time step  $t$ . In each observed state, the agent chooses an action through an  $\epsilon$ -greedy algorithm: with probability  $\epsilon_t$  in time  $t$ , the algorithm chooses an action randomly, and with probability  $1 - \epsilon_t$ , it chooses the action with the highest cumulative action value, i.e.,  $a_{t+1} = \operatorname{argmax}_a Q(s_{t+1}, a)$ . The random selection of actions, called exploration, allows the algorithm to explore the solution space and gives an optimality guarantee to the algorithm if  $\epsilon_t \rightarrow 0$  when  $t \rightarrow \infty$  (Sutton and Barto 1998). After finding optimal  $Q^*$ , one can recover the optimal policy as  $\pi^*(s) = \operatorname{arg max}_a Q^*(s, a)$ .

Both of the algorithms discussed so far (dynamic programming and Q-learning) guarantee that they will obtain the optimal policy. However, due to the curse of dimensionality, these approaches are not able to solve MDPs with large state or action spaces in reasonable amounts of time. Many problems of interest (including the beer game) have large state and/or action spaces. Moreover, in some settings (again, including the beer game), the decision maker cannot observe the full state variable. This case, which is known as a *partially observed MDP* (POMDP), makes the problem much harder to solve than MDPs.

In order to solve large POMDPs and avoid the curse of dimensionality, it is common to approximate the Q-values in the Q-learning algorithm (Sutton and Barto 1998). Linear regression is often used for this purpose (Melo and Ribeiro 2007); however, it is not powerful or accurate enough for our application. Non-linear functions and neural network approximators are able to provide more accurate approximations; on the other hand, they are known to provide unstable or even diverging Q-values due to issues related to non-stationarity and correlations in the sequence of observations (Mnih et al. 2013). The seminal work of Mnih et al. (2015) solved these issues by proposing *target networks* and utilizing *experience replay memory* (Lin 1992). They proposed a *deep Q-network* (DQN) algorithm, which uses a deep neural network to obtain an approximation of the Q-function

and trains it through the iterations of the Q-learning algorithm while updating another target network. This algorithm has been applied to many competitive games, which are reviewed by Li (2017). Our algorithm for the beer game is based on this approach.

The beer game exhibits one characteristic that differentiates it from most settings in which DQN is commonly applied, namely, that there are multiple agents that cooperate in a decentralized manner to achieve a common goal. Such a problem is called a decentralized POMDP, or Dec-POMDP. Due to the partial observability and the non-stationarity of the local observations of agents, Dec-POMDPs are hard to solve and are categorized as NEXP-complete problems (Bernstein et al. 2002).

The beer game exhibits all of the complicating characteristics described above—large state and action spaces, partial state observations, and decentralized cooperation. In the next section, we discuss the drawbacks of current approaches for solving the beer game, which our algorithm aims to overcome.

### 2.3. Drawbacks of Current Algorithms

In Section 2.1, we reviewed different approaches to solve the beer game. Although the model of Clark and Scarf (1960) can solve some types of serial systems, for more general serial systems neither the form nor the parameters of the optimal policy are known. Moreover, even in systems for which a base-stock policy is optimal, such a policy may no longer be optimal for a given agent if the other agents do not follow it. The formula-based beer-game models by Mosekilde and Larsen (1988), Sterman (1989), and Strozzi et al. (2007) attempt to model human decision-making; they do not attempt to model or determine optimal decisions.

A handful of models have attempted to optimize the inventory actions in serial supply chains with more general cost or demand structures than those used by Clark and Scarf (1960); these are essentially beer-game settings. However, these papers all assume full observation or a centralized decision maker, rather than the local observations and decentralized approach taken in the beer game. For example, Kimbrough et al. (2002) use a genetic algorithm (GA), while Chaharsooghi

et al. (2008), Giannoccaro and Pontrandolfo (2002) and Jiang and Sheng (2009) use RL. However, classical RL algorithms can handle only a small or reduced-size state space. Accordingly, these applications of RL in the beer game or even simpler supply chain networks also assume (implicitly or explicitly) that size of the state space is small. This is unrealistic in the beer game, since the state variable representing a given agent's inventory level can be any number in  $(-\infty, +\infty)$ . Solving such an RL problem would be nearly impossible, as the model would be extremely expensive to train. Moreover, Chaharsooghi et al. (2008) and Giannoccaro and Pontrandolfo (2002), which model beer-game-like settings, assume sharing of information. Also, to handle the curse of dimensionality, they propose mapping the state variable onto a small number of tiles, which leads to the loss of valuable state information and therefore of accuracy. Thus, although these papers are related to our work, their assumption of full observability differentiates their work from the classical beer game and from our paper.

Another possible approach to tackle this problem might be classical supervised machine learning algorithms. However, these algorithms also cannot be readily applied to the beer game, since there is no historical data in the form of “correct” input/output pairs. Thus, we cannot use a stand-alone support vector machine or deep neural network with a training data-set and train it to learn the best action (like the approach used by Oroojlooyjadid et al. (2017a,b) to solve some simpler supply chain problems). Based on our understanding of the literature, there is a large gap between solving the beer game problem effectively and what the current algorithms can handle. In order to fill this gap, we propose a variant of the DQN algorithm to choose the order quantities in the beer game.

## 2.4. Our Contribution

We propose a Q-learning algorithm for the beer game in which a DNN approximates the Q-function. Indeed, the general structure of our algorithm is based on the DQN algorithm (Mnih et al. 2015), although we modify it substantially, since DQN is designed for single-agent, competitive, zero-sum games and the beer game is a multi-agent, decentralized, cooperative, non-zero-sum game. In other words, DQN provides actions for one agent that interacts with an environment in a competitive

setting, and the beer game is a cooperative game in the sense that all of the players aim to minimize the total cost of the system in a random number of periods. Also, beer game agents are playing independently and do not have any information from other agents until the game ends and the total cost is revealed, whereas DQN usually assumes the agent fully observes the state of the environment at any time step  $t$  of the game. For example, DQN has been successfully applied to Atari games (Mnih et al. 2015), but in these games the agent is attempting to defeat an opponent and observes full information about the state of the systems at each time step  $t$ .

One naive approach to extend the DQN algorithm to solve the beer game is to use multiple DQNs, one for each agent. However, using DQN as the decision maker of each agent results in a competitive game in which each DQN agent plays independently to minimize its own cost. For example, consider a beer game in which players 2, 3, and 4 each have a stand-alone, well-trained DQN and the retailer (stage 1) uses a base-stock policy to make decisions. If the holding costs are positive for all players and the stockout cost is positive only for the retailer (as is common in the beer game), then the DQN at agents 2, 3, and 4 will return an optimal order quantity of 0 in every period, since on-hand inventory hurts the objective function for these players, but stockouts do not. This is a byproduct of the independent DQN agents minimizing their own costs without considering the total cost, which is obviously not an optimal solution for the system as a whole.

Instead, we propose a unified framework in which the agents still play independently from one another, but in the training phase, we use a feedback scheme so that the DQN agent learns the total cost for the whole network and can, over time, learn to minimize it. Thus, the agents in our model play smartly in all periods of the game to get a near-optimal cumulative cost for any random horizon length.

In principle, our framework can be applied to multiple DQN agents playing the beer game simultaneously on a team. However, to date we have designed and tested our approach only for a single DQN agent whose teammates are not DQNs, e.g., they are controlled by simple formulas or by human players. Enhancing the algorithm so that multiple DQNs can play simultaneously and cooperatively is a topic of ongoing research.

Another advantage of our approach is that it does not require knowledge of the demand distribution, unlike classical inventory management approaches (e.g., Clark and Scarf 1960). In practice, one can approximate the demand distribution based on historical data, but doing so is prone to error, and basing decisions on approximate distributions may result in loss of accuracy in the beer game. In contrast, our algorithm chooses actions directly based on the training data and does not need to know, or estimate, the probability distribution directly.

The proposed approach works very well when we tune and train the DQN for a given agent and a given set of game parameters (e.g., costs, lead times, action spaces, etc.). Once any of these parameters changes, or the agent changes, in principle we need to tune and train a new network. Although this approach works, it is time consuming since we need to tune hyper-parameters for each new set of game parameters. To avoid this, we propose using a *transfer learning* approach (Pan and Yang 2010) in which we transfer the acquired knowledge of one agent under one set of game parameters to another agent with another set of game parameters. In this way, we decrease the required time to train a new agent by roughly one order of magnitude.

To summarize, our algorithm is *a variant of the DQN algorithm for choosing actions in the beer game*. In order to attain near-optimal cooperative solutions, we develop *a feedback scheme as a communication framework*. Finally, to simplify training agents with new settings, we use *transfer learning* to efficiently make use of the learned knowledge of trained agents. In addition to playing the beer game well, we believe our algorithm serves as a proof-of-concept that DQN and other machine learning approaches can be used for real-time decision making in complex supply chain settings. Finally, we note that we have integrated our algorithm into a new online beer game developed by Opex Analytics (<http://beergame.opexanalytics.com/>); see Figure 3. The Opex beer game allows human players to compete with, or play on a team with, our DQN agent.

### 3. The DQN Algorithm

In this section, we first present the details of our DQN algorithm to solve the beer game, and then describe the transfer learning mechanism.

**Figure 3 Screenshot of Opex Analytics online beer game integrated with our DQN agent**



### 3.1. DQN for the Beer Game

In our algorithm, a DQN agent runs a Q-learning algorithm with DNN as the Q-function approximator to learn a semi-optimal policy with the aim of minimizing the total cost of the game. Each agent has access to its local information and considers the other agents as parts of its environment. That is, the DQN agent does not know any information about the other agents, including both static parameters such as costs and lead times, as well as dynamic state variables such as inventory levels. We propose a feedback scheme to teach the DQN agent to work toward minimizing the total system-wide cost, rather than its own local cost. The details of the scheme, Q-learning, state and action spaces, reward function, DNN approximator, and the DQN algorithm are discussed below.

**State variables:** Consider agent  $i$  in time step  $t$ . Let  $OO_t^i$  denote the on-order items at agent  $i$ , i.e., the items that have been ordered from agent  $i+1$  but not received yet; let  $AO_t^i$  denote the size of the arriving order (i.e., the demand) received from agent  $i-1$ ; let  $AS_t^i$  denote the size of the arriving shipment from agent  $i+1$ ; let  $a_t^i$  denote the action agent  $i$  takes; and let  $IL_t^i$  denote the inventory level as defined in Section 1. We interpret  $AO_t^1$  to represent the end-customer demand and  $AS_t^4$  to represent the shipment received by agent 4 from the external supplier. In each period  $t$  of the game, agent  $i$  observes  $IL_t^i$ ,  $OO_t^i$ ,  $AO_t^i$ , and  $AS_t^i$ . In other words, in period  $t$  agent  $i$  has historical observations  $o_t^i = [((IL_1^i)^+, IL_1^i)^-, OO_1^i, AO_1^i, RS_1^i), \dots, ((IL_t^i)^+, IL_t^i)^-, OO_t^i, AO_t^i, AS_t^i)]$ . In addition, any beer game will finish in a finite time horizon, so the problem can be modeled as a POMDP in which each historic sequence  $o_t^i$  is a distinct state and the size of the vector  $o_t^i$  grows over time, which is difficult for any RL or DNN algorithm to handle. To address this issue, we

capture only the last  $m$  periods (e.g.,  $m = 3$ ) and use them as the state variable; thus the state variable of agent  $i$  in time  $t$  is  $s_t^i = [((IL_j^i)^+, IL_j^i)^-, OO_j^i, AO_j^i, RS_j^i]_{j=t-m+1}^t$ .

**DNN architecture:** In our algorithm, DNN plays the role of the Q-function approximator, providing the Q-value as output for any pair of state  $s$  and action  $a$ . There are various possible approaches to build the DNN structure. The natural approach is to provide the state  $s$  and action  $a$  as the input of the DNN and then get the corresponding  $Q(s, a)$  from the output. Thus, we provide as input the  $m$  previous state variables into the DNN and get as output  $Q(s, a)$  for every possible action  $a \in \mathbb{A}$  (since in beer game  $\mathbb{A}(s)$  is fixed for any  $s$ , we use  $\mathbb{A}$  hereinafter).

**Action space:** In each period of the game, each agent can order any amount in  $[0, \infty)$ . Since our DNN architecture provides the Q-value of all possible actions in the output, having an infinite action space is not practical. Therefore, to limit the cardinality of the action space, we use the  $d + x$  rule for selecting the order quantity: The agent determines how much more or less to order than its received order; that is, the order quantity is  $d + x$ , where  $x$  is in some bounded set. Thus, the output of the DNN is  $x \in [a_l, a_u]$  ( $a_l, a_u \in \mathbb{Z}$ ), so that the action space is of size  $a_u - a_l + 1$ .

**Experience replay:** The DNN algorithm requires a mini-batch of input and a corresponding set of output values to learn the Q-values. Since we use DQN algorithm as our RL engine, we have the new state  $s_{t+1}$ , the current state  $s_t$ , the action  $a_t$  taken, and the observed reward  $r_t$ , in each period  $t$ . This information can provide the required set of input and output for the DNN; however, the resulting sequence of observations from the RL results in a non-stationary data-set in which there is a strong correlation among consecutive records. This makes the DNN and, as a result, the RL prone to over-fitting the previously observed records and may even result in a diverging approximator (Mnih et al. 2015). To avoid this problem, we follow the suggestion of Mnih et al. (2015) and use *experience replay* (Lin 1992). In this way, agent  $i$  has experience memory  $E^i$  that in iteration  $t$  of the algorithm, agent  $i$ 's observation  $e_t^i = (s_t^i, a_t^i, r_t^i, s_{t+1}^i)$  is added in, so that  $E^i$  includes  $\{e_1^i, e_2^i, \dots, e_t^i\}$  in period  $t$ . Then, in order to avoid having correlated observations, we select a random mini-batch of the agent's experience replay to train the corresponding DNN (if applicable). This approach

breaks the correlations among the training data and reduces the variance of the output (Mnih et al. 2013). Moreover, as a byproduct of experience replay, we also get a tool to keep every piece of the valuable information, which allows greater efficiency in a setting in which the state and action spaces are huge and any observed experience is valuable.

**Reward function:** In iteration  $t$  of the game, agent  $i$  observes state variable  $s_t^i$  and takes action  $a_t^i$ ; we need to know the corresponding reward value  $r_t^i$  to measure the quality of action  $a_t^i$ . The state variable,  $s_{t+1}^i$ , allows us to calculate  $IL_{t+1}^i$  and thus the corresponding shortage or holding costs, and we consider the summation of these costs for  $r_t^i$ . However, since there are information and transportation lead times, there is a delay between taking action  $a_t^i$  and observing its effect on the reward. Moreover, the reward  $r_t^i$  reflects not only the action taken in period  $t$ , but also those taken in previous periods, and it is not possible to decompose  $r_t^i$  to isolate the effects of each of these actions. However, defining the state variable to include information from the last  $m$  periods resolves this issue to some degree; the reward  $r_t^i$  represents the reward of state  $s_t^i$ , which includes the observations of the previous  $m$  steps.

On the other hand, the reward values  $r_t^i$  are the intermediate rewards of each agent, and the objective of the beer game is to minimize the total reward of the game,  $\sum_{i=1}^4 \sum_{t=1}^T r_t^i$ , which the agents only learn after finishing the game. In order to add this information into the agents' experience, we use reward shaping through a feedback scheme.

**Feedback scheme:** When any episode of the beer game is finished, all agents are made aware of the total reward. In order to share this information among the agents, we propose a penalization procedure in the training phase to provide feedback to the DQN agent about the way that it has played. Let  $\omega = \sum_{i=1}^4 \sum_{t=1}^T \frac{r_t^i}{T}$  and  $\tau^i = \sum_{t=1}^T \frac{r_t^i}{T}$ , i.e., the average reward per period and the average reward of agent  $i$  per period, respectively. After the end of each episode of the game (i.e., after period  $T$ ), for each DQN agent  $i$  we update its observed reward in all  $T$  time steps in the experience replay memory using  $r_t^i = r_t^i + \frac{\beta_i}{3}(\omega - \tau^i)$ ,  $\forall t \in \{1, \dots, T\}$ , where  $\beta_i$  is a regularization coefficient for agent  $i$ . With this procedure, agent  $i$  gets appropriate feedback about its actions and learns to take actions that result in minimum total cost, not locally optimal solutions.

**Determining the value of  $m$ :** As noted above, the DNN maintains information from the most recent  $m$  periods in order to keep the size of the state variable fixed and to address the issue with the delayed observation of the reward. In order to select an appropriate value for  $m$ , one has to consider the value of the lead times throughout the game. First, when agent  $i$  takes action  $a_t^i$  at time  $t$ , it does not observe its effect until at least  $l_i^{tr} + l_i^{in}$  periods later, when the order may be received. Moreover, node  $i+1$  may not have enough stock to satisfy the order immediately, in which case the shipment is delayed and in the worst case agent  $i$  will not observe the corresponding reward  $r_t^i$  until  $\sum_{j=i}^4(l_j^{tr} + l_j^{in})$  periods later. However, one needs the reward  $r_t^i$  to evaluate the action  $a_t^i$  taken. Thus, ideally  $m$  should be chosen at least as large as  $\sum_{j=1}^4(l_j^{tr} + l_j^{in})$ . On the other hand, this value can be large and selecting a large value for  $m$  results in a large input size for the DNN, which increases the training time. Therefore, selecting  $m$  is a trade-off between accuracy and computation time, and  $m$  should be selected according to the required level of accuracy and the available computation power. In our numerical experiment,  $\sum_{j=1}^4(l_j^{tr} + l_j^{in}) = 15$  or  $16$ , and we test  $m \in \{5, 10\}$ .

**The algorithm:** Our algorithm to get the policy  $\pi$  to solve the beer game is provided in Algorithm 1. The algorithm, which is based on that of Mnih et al. (2015), finds weights  $\theta$  of the DNN network to minimize the Euclidean distance between  $Q(s, a; \theta)$  and  $y_j$ , where  $y_j$  is the prediction of the Q-value that is obtained from target network  $Q^-$  with weights  $\theta^-$ . Every  $C$  iterations, the weights  $\theta^-$  are updated by  $\theta$ . Moreover, the actions in each training step of the algorithm are obtained by an  $\epsilon$ -greedy algorithm, which is explained in Section 2.2.

In the algorithm, in period  $t$  agent  $i$  takes action  $a_t^i$ , satisfies the arriving demand/order  $AO_{t-1}^i$ , observes the new demand  $AO_t^i$ , and then receives the shipments  $AS_t^i$ . This sequence of events, which is explained in detail in online supplement E, results in the new state  $s_{t+1}$ . Feeding  $s_{t+1}$  into the DNN network with weights  $\theta$  provides the corresponding Q-value for state  $s_{t+1}$  and all possible actions. The action with the smallest Q-value is our choice. Finally, at the end of each episode, the feedback scheme runs and distributes the total cost among all agents.

---

**Algorithm 1** DQN for Beer Game

---

```

1: procedure DQN
2:   Initialize Experience Replay Memory  $E_i = [ ]$ ,  $\forall i$ 
3:   for Episode = 1 : n do
4:     Reset  $IL$ ,  $OO$ ,  $d$ ,  $AO$ , and  $AS$  for each agent
5:     for  $t = 1 : T$  do
6:       for  $i = 1 : 4$  do
7:         With probability  $\epsilon$  take random action  $a_t$ ,
8:         otherwise set  $a_t = \operatorname{argmin}_a Q(s_t, a; \theta)$ 
9:         Execute action  $a_t$ , observe reward  $r_t$  and state  $s_{t+1}$ 
10:        Add  $(s_t^i, a_t^i, r_t^i, s_{t+1}^i)$  into the  $E_i$ 
11:        Get a mini-batch of experiences  $(s_j, a_j, r_j, s_{j+1})$  from  $E_i$ 
12:        Set  $y_j = \begin{cases} r_j & \text{if it is the terminal state} \\ r_j + \min_a Q(s_j, a; \theta^-) & \text{otherwise} \end{cases}$ 
13:        Run forward and backward step on the DNN with loss function  $(y_j - Q(s_j, a_j; \theta))^2$ 
14:        Every  $C$  iterations, set  $\theta^- = \theta$ 
15:      end for
16:    end for
17:    Run feedback scheme, update experience replay of each agent
18:  end for
19: end procedure

```

---

**Evaluation procedure:** In order to validate our algorithm, we compare the results of our algorithm to those obtained using the optimal base-stock levels (when possible) in serial systems by Clark and Scarf (1960), as well as models of human beer-game behavior by Sterman (1989). (Note that none of these methods attempts to do exactly the same thing as our method. The methods by Clark and Scarf (1960) optimizes the base-stock levels assuming all players follow a base-stock policy—which beer game players do not tend to do—and the formula by Sterman (1989) models human beer-game play, but they do not attempt to optimize.) The details of the training procedure and benchmarks are described in Section 4.

### 3.2. Transfer Learning

Transfer learning (Pan and Yang 2010) has been an active and successful field of research in machine learning and especially in image processing. In transfer learning, there is a *source* dataset  $S$  and a trained neural network to perform a given task, e.g. classification, regression, or decisioning through RL. Training such networks may take a few days or even weeks. So, for similar or even slightly different *target* datasets  $T$ , one can avoid training a new network from scratch and instead use the

same trained network with a few customizations. The idea is that most of the learned knowledge on dataset  $S$  can be used in the target dataset with a small amount of additional training. This idea works well in image processing (e.g. Rajpurkar et al. (2017)) and considerably reduces the training time.

In order to use transfer learning in the beer game, assume there exists a source agent  $i \in \{1, 2, 3, 4\}$  with trained network  $S_i$  (with a fixed size on all agents), parameters  $P_1^i = \{|\mathbb{A}_1^j|, c_{p_1}^j, c_{h_1}^j\}$ , observed demand distribution  $D_1$ , and co-player policy  $\pi_1$ . The weight matrix  $W_i$  contains the learned weights such that  $W_i^q$  denotes the weight between layers  $q$  and  $q+1$  of the neural network, where  $q \in \{0, \dots, nh\}$ , and  $nh$  is the number of hidden layers. The aim is to train a neural network  $S_j$  for target agent  $j \in \{1, 2, 3, 4\}$ ,  $j \neq i$ . We set the structure of the network  $S_j$  the same as that of  $S_i$ , and initialize  $W_j$  with  $W_i$ , making the first  $k$  layers not trainable. Then, we train neural network  $S_j$  with a small learning rate. Note that, as we get closer to the final layer, which provides the Q-values, the weights become less similar to agent  $i$ 's and more specific to each agent. Thus, the acquired knowledge in the first  $k$  hidden layer(s) of the neural network belonging to agent  $i$  is transferred to agent  $j$ , in which  $k$  is a tunable parameter. Following this procedure, in Section 4.3, we test the use of transfer learning in six cases to transfer the learned knowledge of source agent  $i$  to:

1. Target agent  $j \neq i$  in the same game.
2. Target agent  $j$  with  $\{|\mathbb{A}_1^j|, c_{p_2}^j, c_{h_2}^j\}$ , i.e., the same action space but different cost coefficients.
3. Target agent  $j$  with  $\{|\mathbb{A}_2^j|, c_{p_1}^j, c_{h_1}^j\}$ , i.e., the same cost coefficients but different action space.
4. Target agent  $j$  with  $\{|\mathbb{A}_2^j|, c_{p_2}^j, c_{h_2}^j\}$ , i.e., different action space and cost coefficients.
5. Target agent  $j$  with  $\{|\mathbb{A}_2^j|, c_{p_2}^j, c_{h_2}^j\}$ , i.e., different action space and cost coefficients, as well as a different demand distribution  $D_2$ .
6. Target agent  $j$  with  $\{|\mathbb{A}_2^j|, c_{p_2}^j, c_{h_2}^j\}$ , i.e., different action space and cost coefficients, as well as a different demand distribution  $D_2$  and co-player policy  $\pi_2$ .

Unless stated otherwise, the demand distribution and co-player policy are the same for the source and target agents. Transfer learning could also be used when other aspects of the problem change,

e.g., lead times, state representation, and so on. This avoids having to tune the parameters of the neural network for each new problem, which considerably reduces the training time. However, we still need to decide how many layers should be trainable, as well as to determine which agent can be a base agent for transferring the learned knowledge. Still, this is computationally much cheaper than finding each network and its hyper-parameters from scratch.

#### 4. Numerical Experiments

In Section 4.1, we discuss a set of numerical experiments that uses a simple demand distribution and a relatively small action space:

- $d_0^t \in \mathbb{U}[0, 2]$ ,  $\mathbb{A} = \{-2, -1, 0, 1, 2\}$ .

After exploring the behavior of our algorithm under different co-player policies, in Section 4.2 we test the algorithm using three well-known cases from the literature, which have larger possible demand values and action spaces:

- $d_0^t \in \mathbb{U}[0, 8]$ ,  $\mathbb{A} = \{-8, \dots, 8\}$  (Croson and Donohue 2006b)
- $d_0^t \in \mathbb{N}(10, 2^2)$ ,  $\mathbb{A} = \{-5, \dots, 5\}$  (adapted from Chen and Samroengraja 2000, , who assume  $\mathbb{N}(50, 20^2)$ )
- $d_0^t \in C(4, 8)$ ,  $\mathbb{A} = \{-8, \dots, 8\}$  (Sterman 1989).

As noted above, we only consider cases in which a single DQN plays with non-DQN agents, e.g., simulated human co-players. In each of the cases listed above, we consider three types of policies that the non-DQN co-players follow: (i) base-stock policy, (ii) Sterman formula, (iii) random policy. In the random policy, agent  $i$  also follows a  $d + x$  rule, in which  $a_i^t \in \mathbb{A}$  is selected randomly and with equal probability, for each  $t$ . After analyzing these cases, in Section 4.3 we provide the results obtained using transfer learning for each of the six proposed cases.

We test values of  $m$  in  $\{5, 10\}$  and  $C \in \{5000, 10000\}$ . Our DNN network is a fully connected network, in which each node has a ReLU activation function. The input is of size  $5m$ , and there are three hidden layers in the neural network. There is one output node for each possible value of the action, and each of these nodes takes a value in  $\mathbb{R}$  indicating the Q-value for that action. Thus, there are  $a_u - a_l + 1$  output nodes, and the neural network has shape  $[5m, 180, 130, 61, a_u - a_l + 1]$ .

In order to optimize the network, we used the Adam optimizer (Kingma and Ba 2014) with a batch size of 64. Although the Adam optimizer has its own weight decaying procedure, we used exponential decay with a stair of 10000 iterations with rate 0.98 to decay the learning rate further. This helps to stabilize the training trajectory. We trained each agent on at most 60000 episodes and used a replay memory  $E$  equal to the one million most recently observed experiences. Also, the training of the DNN starts after observing at least 500 episodes of the game. The  $\epsilon$ -greedy algorithm starts with  $\epsilon = 0.9$  and linearly reduces it to 0.1 in the first 80% of iterations.

In the feedback mechanism, the appropriate value of the feedback coefficient  $\beta_i$  heavily depends on  $\tau_j$ , the average reward for agent  $j$ , for each  $j \neq i$ . For example, when  $\tau_i$  is one order of magnitude larger than  $\tau_j$ , for all  $j \neq i$ , agent  $i$  needs a large coefficient to get more feedback from the other agents. Indeed, the feedback coefficient has a similar role as the regularization parameter  $\lambda$  has in the lasso loss function; the value of that parameter depends on the  $\ell$ -norm of the variables, but there is no universal rule to determine the best value for  $\lambda$ . Similarly, proposing a simple rule or value for each  $\beta_i$  is not possible, as it depends on  $\tau_i$ ,  $\forall i$ . For example, we found that a very large  $\beta_i$  does not work well, since the agent tries to decrease other agents' costs rather than its own. Similarly, with a very small  $\beta_i$ , the agent learns how to minimize its own cost instead of the total cost. Therefore, we used a similar cross validation approach to find good values for each  $\beta_i$ .

#### 4.1. Basic Cases

In this section, we test our approach using a beer game setup with the following characteristics. Information and shipment lead times,  $l_j^{tr}$  and  $l_j^{in}$ , equal 2 periods at every agent. Holding and stockout costs are given by  $c_h = [2, 2, 2, 2]$  and  $c_p = [2, 0, 0, 0]$ , respectively, where the vectors specify the values for agents 1, ..., 4. The demand is an integer uniformly drawn from  $\{0, 1, 2\}$ . Additionally, we assume that agent  $i$  observes the arriving shipment  $AS_t^i$  when it chooses its action for period  $t$ . We relax this assumption later. We use  $a_l = -2$  and  $a_u = 2$ ; so that there are 5 outputs in the neural network. i.e., each agent chooses an order quantity that is at most 2 units greater or less than the observed demand. (Later, we expand these to larger action spaces.)

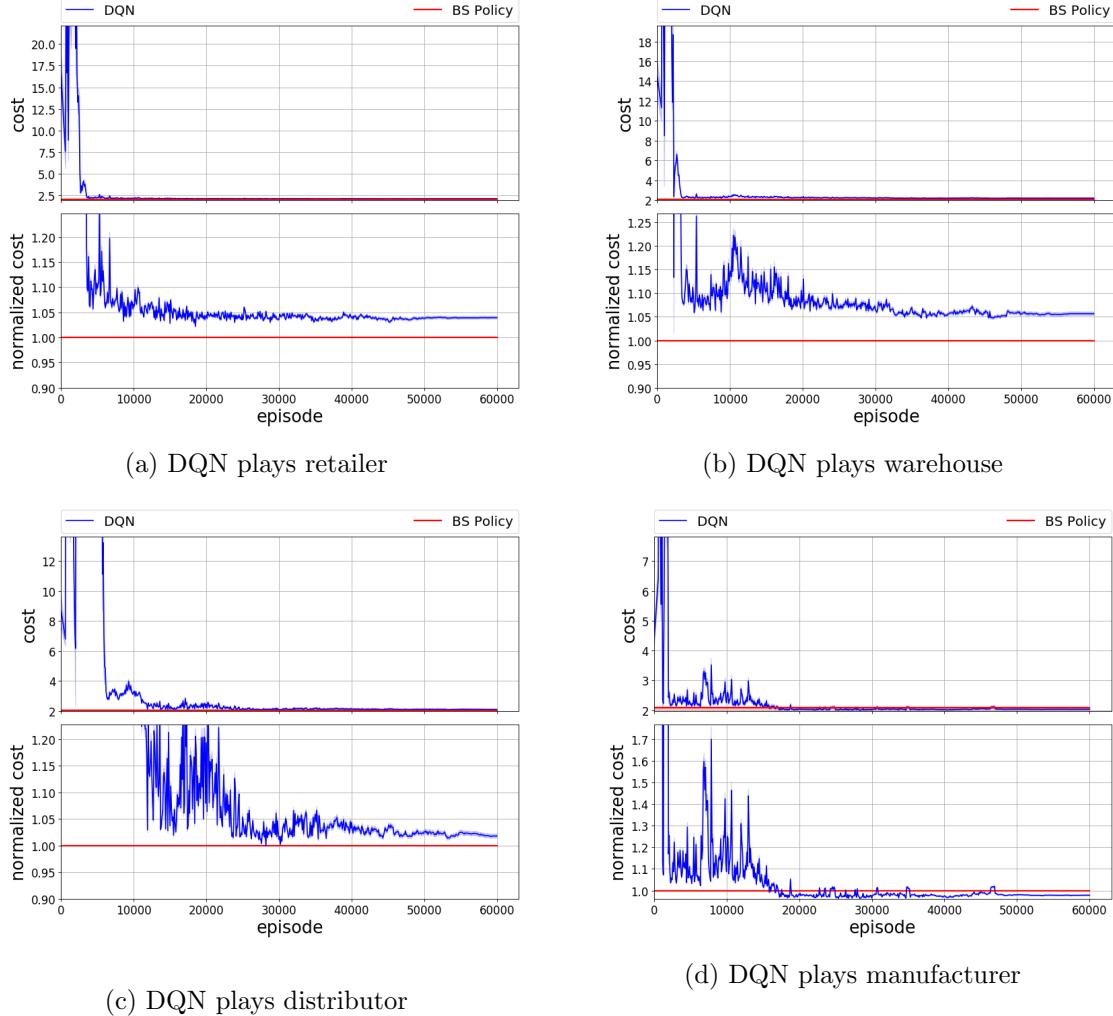
We consider two types of simulated human players. In Section 4.1.1, we discuss results for the case in which one DQN agent plays on a team in which the other three players use a base-stock policy to choose their actions, i.e., the non-DQN agents behave rationally. See <https://youtu.be/gQa6iWGcGWY> for a video animation of the policy that the DQN learns in this case. Then, in Section 4.1.2, we assume that the other three agents use the Sterman formula (i.e., the formula by Sterman (1989)), which models irrational play.

For the cost coefficients and other settings specified for this beer game, it is optimal for all players to follow a base-stock policy, and we use this policy as a benchmark and a lower bound on the base stock cost. The vector of optimal base-stock levels is [8, 8, 0, 0], and the resulting average cost per period is 2.0705, though these levels may be slightly suboptimal due to rounding. This cost is allocated to stages 1–4 as [2.0073, 0.0632, 0.03, 0.00]. In the experiments in which one of the four agents is played by DQN, the other three agents continue to use their optimal base-stock levels.

**4.1.1. DQN Plus Base-Stock Policy** We consider four cases, with the DQN playing the role of each of the four players and the co-players using a base-stock policy. We then compare the results of our algorithm with the results of the case in which all players follow a base-stock policy, which we call **BS** hereinafter.

The results of all four cases are shown in Figure 4. Each plot shows the training curve, i.e., the evolution of the average cost per game as the training progresses. In particular, the horizontal axis indicates the number of training episodes, while the vertical axis indicates the total cost per game. After every 100 episodes of the game and the corresponding training, the cost of 50 validation points (i.e., 50 new games) each with 100 periods, are obtained and their average plus a 95% confidence interval are plotted. (The confidence intervals, which are light blue in the figure, are quite narrow, so they are difficult to see.) The red line indicates the cost of the case in which all players follow a base-stock policy. In each of the sub-figures, there are two plots; the upper plot shows the cost, while the lower plot shows the normalized cost, in which each cost is divided by the corresponding **BS** cost; essentially this is a “zoomed-in” version of the upper plot. We trained

**Figure 4 Total cost (upper figure) and normalized cost (lower figure) with one DQN agent and three agents that follow base-stock policy**

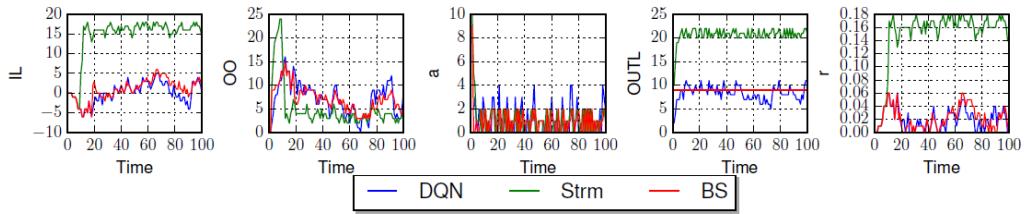


the network using values of  $\beta \in \{5, 10, 20, 50, 100, 200\}$ , each for at most 60000 episodes. Figure 4 plots the results from the best  $\beta_i$  value for each agent; we present the full results using different  $\beta_i, m$  and  $C$  values in Section C of the online supplement.

The figure indicates that DQN performs well in all cases and finds policies whose costs are close to those of the BS policy. After the network is fully trained (i.e., after 60000 training episodes), the average gap between the DQN cost and the BS cost, over all four agents, is 2.31%.

Figure 5 shows the trajectories of the retailer's inventory level ( $IL$ ), on-order quantity ( $OO$ ), order quantity ( $a$ ), reward ( $r$ ), and order up to level (OUTL) for a single game, when the retailer is played by the DQN with  $\beta_1 = 50$ , as well as when it is played by a base-stock policy (BS), and

**Figure 5**  $IL_t$ ,  $OO_t$ ,  $a_t$ ,  $r_t$ , and  $OUTL$  when DQN plays retailer and other agents follow base-stock policy



the Sterman formula (**Strm**). The base-stock policy and DQN have similar  $IL$  and  $OO$  trends, and as a result their rewards are also very close: **BS** has a cost of  $[1.42, 0.00, 0.02, 0.05]$  (total 1.49) and DQN has  $[1.43, 0.01, 0.02, 0.08]$  (total 1.54, or 3.4% larger). (Note that **BS** has a slightly different cost here than reported on page 20 because those costs are the average costs of 50 samples while this cost is from a single sample.) Similar trends are observed when the DQN plays the other three roles; see Section B of the online supplement. This suggests that the DQN can successfully learn to achieve costs close to **BS** when the other agents also play **BS**. (The  $OUTL$  plot shows that the DQN does not quite *follow* a base-stock policy, even though its costs are similar.)

**4.1.2. DQN Plus Sterman Formula** Figure 6 shows the results of the case in which the three non-DQN agents use the formula proposed by Sterman (1989) instead of a base-stock policy. (See Section A of online supplement for the formula and its parameters.) For comparison, the red line represents the case in which the single agent is played using a base-stock policy and the other three agents continue to use the Sterman formula, a case we call **Strm-BS**.

From the figure, it is evident that the DQN plays much *better* than **Strm-BS**. This is because if the other three agents do not follow a base-stock policy, it is no longer optimal for the fourth agent to follow a base-stock policy, or to use the same base-stock level. In general, the optimal inventory policy when other agents do not follow a base-stock policy is an open question. This figure suggests that our DQN is able to learn to play effectively in this setting.

Table 1 gives the cost of all four agents when a given agent plays using either DQN or a base-stock policy and the other agents play using the Sterman formula. From the table, we can see that DQN learns how to play to decrease the costs of the other agents, and not just its own costs—for

**Table 1** Average cost under different choices of which agent uses DQN or Sterm-BS.

DQN Agent	Cost (DQN, Sterm-BS)				
	Retailer	Warehouse	Distributer	Manufacturer	Total
Retailer	(0.89, 1.89)	(10.87, 10.83)	(10.96, 10.98)	(12.42, 12.82)	(35.14, 36.52)
Warehouse	(1.74, 9.99)	(0.00, 0.13)	(11.12, 10.80)	(12.86, 12.34)	(25.72, 33.27)
Distributer	(5.60, 10.72)	(0.11, 9.84)	(0.00, 0.14)	(12.53, 12.35)	(18.25, 33.04)
Manufacturer	(4.68, 10.72)	(1.72, 10.60)	(0.24, 10.13)	(0.00, 0.07)	(6.64, 31.52)

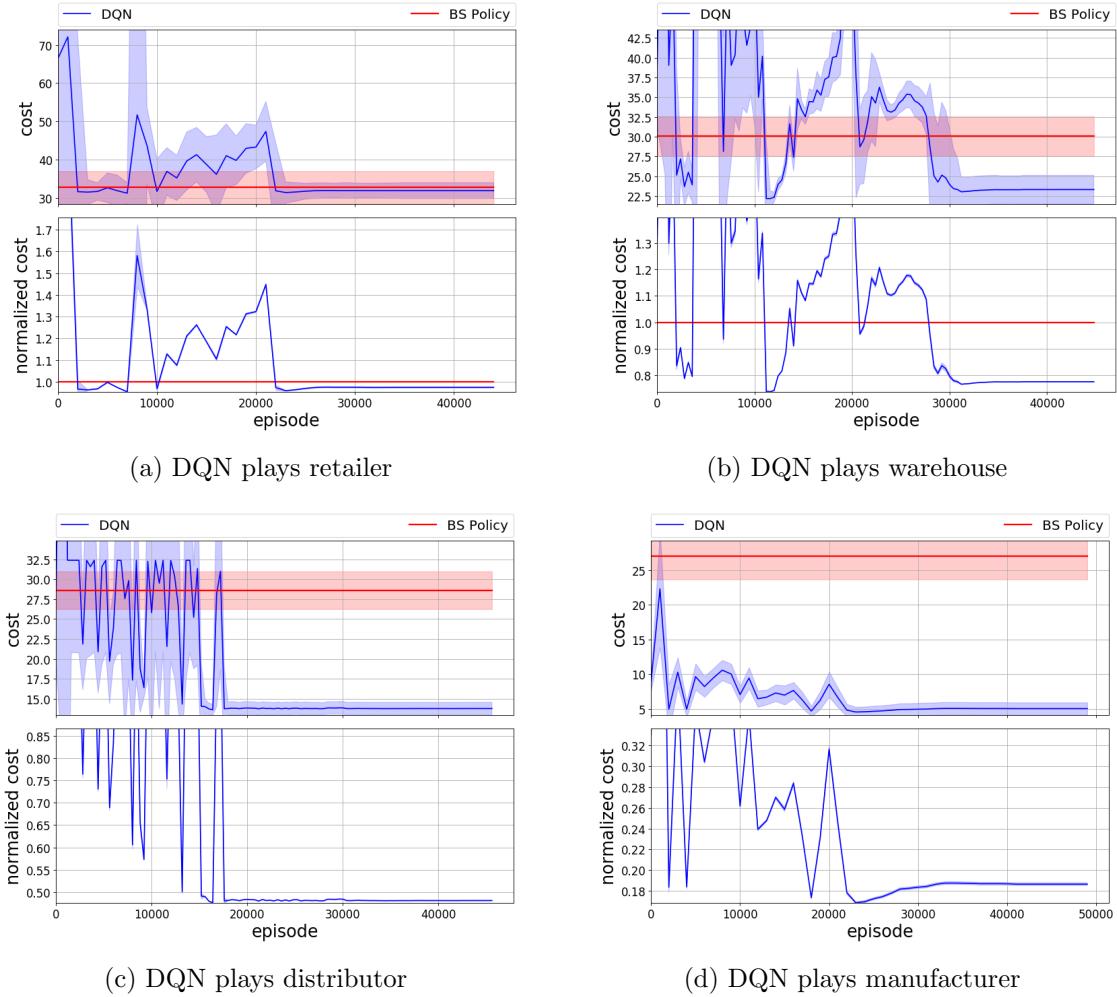
example, the retailer's and warehouse's costs are significantly lower when the distributor uses DQN than they are when the distributor uses a base-stock policy. Similar conclusions can be drawn from Figure 6. This shows the power of DQN when it plays with co-player agents that do not play rationally, i.e., do not follow a base-stock policy, which is common in real-world supply chains. Also, we note that when all agents follow the Sterman formula, the average cost of the agents is [10.81, 10.76, 10.96, 12.6], for a total of 45.13, much higher than when any one agent uses DQN. Finally, for details on  $IL$ ,  $OO$ ,  $a$ ,  $r$ , and  $OUTL$  on this case see Section B of the online supplement.

#### 4.2. Literature Benchmarks

We next test our approach on beer game settings from the literature. These have larger demand-distribution domains, and therefore larger plausible action spaces, and thus represent harder instances to train the DQN for. In all instances in this section,  $l^{in} = [2, 2, 2, 2]$  and  $l^{tr} = [2, 2, 2, 1]$ . Shortage and holding cost coefficients and the base-stock levels for each instance are presented in Table 2.

Note that the Clark–Scarf algorithm assumes that stage 1 is the only stage with non-zero stockout costs, whereas the  $\mathbb{U}[0, 8]$  instance has non-zero costs at every stage. Therefore, we used a heuristic approach based on a two-moment approximation, similar to that proposed by Graves (1985), to choose the base-stock levels; see Snyder (2018). In addition, the  $C(4, 8)$  demand process is non-stationary—4, then 8—but we allow only stationary base-stock levels. Therefore, we chose to set the base-stock levels equal to the values that would be optimal if the demand were 8 in every period. Finally, in the experiments in this section, we assume that agent  $i$  observes  $AS_t^i$  after choosing  $a_t^i$ , whereas in Section 4.1 we assumed the opposite. Therefore, the agents in these experiments have one fewer piece of information when choosing actions, and are therefore more difficult to train.

**Figure 6 Total cost (upper figure) and normalized cost (lower figure) with one DQN agent and three agents that follow the Sterman formula**



**Table 2 Cost parameters and base-stock levels for instances with uniform, normal, and classic demand distributions.**

demand	$c_p$	$c_h$	BS level
$\mathbb{U}[0, 8]$	[1.0, 1.0, 1.0, 1.0]	[0.50, 0.50, 0.50, 0.50]	[19, 20, 20, 14]
$N(10, 2^2)$	[10.0, 0.0, 0.0, 0.0]	[1.00, 0.75, 0.50, 0.25]	[48, 43, 41, 30]
$C(4, 8)$	[1.0, 1.0, 1.0, 1.0]	[0.50, 0.50, 0.50, 0.50]	[32, 32, 32, 24]

Tables 3, 4, and 5 show the results of the cases in which the DQN agent plays with co-players who follow base-stock, Sterman, and random policies, respectively. In each group of columns, the first column (“DQN”) gives the average cost (over 50 instances) when one agent (indicated by the first column in the table) is played by the DQN and co-players are played by base-stock (Table 3), Sterman (Table 4), or random (Table 5) agents. The second column in each group (“BS”, “Strm-BS”, “Rand-BS”) gives the corresponding cost when the DQN agent is replaced by a base-stock agent

**Table 3 Results of DQN playing with co-players who follow base-stock policy.**

	Uniform			Normal			Classic		
	DQN	BS	Gap (%)	DQN	BS	Gap (%)	DQN	BS	Gap (%)
R	904.88	799.20	13.22	881.66	838.14	5.19	0.50	0.34	45.86
W	960.44	799.20	20.18	932.65	838.14	11.28	0.47	0.34	36.92
D	903.49	799.20	13.05	880.40	838.14	5.04	0.67	0.34	96.36
M	830.16	799.20	3.87	852.33	838.14	1.69	0.30	0.34	-13.13
Average			12.58			5.80			41.50

(using the base-stock levels given in Table 2) and the co-players remain as in the previous column.

The third column (“Gap”) gives the percentage difference between these two costs.

As Table 3 shows, when the DQN plays with base-stock co-players under uniform or normal demand distributions, it obtains costs that are reasonably close to the case when all players use a base-stock policy, with average gaps of 12.58% and 5.80%, respectively. These gaps are not quite as small as those in Section 4.1, due to the larger action spaces in the instances in this section. Since a base-stock policy is optimal at every stage, the small gaps demonstrate that the DQN can learn to play the game well for these demand distributions. For the classic demand process, the percentage gaps are larger. To see why, note that if the demand were to equal 8 in every period, the base-stock levels for the classic demand process will result in ending inventory levels of 0 at every stage. The four initial periods of demand equal to 4 disrupt this effect slightly, but the cost of the optimal base-stock policy for the classic demand process is asymptotically 0 as the time horizon goes to infinity. The absolute gap attained by the DQN is quite small—an average of 0.49 vs. 0.34 for the base-stock cost—but the percentage difference is large simply because the optimal cost is close to 0. Indeed, if we allow the game to run longer, the cost of both algorithms decreases, and so does the absolute gap. For example, when the DQN plays the retailer, after 500 periods the discounted costs are 0.0090 and 0.0062 for DQN and BS, respectively, and after 1000 periods, the costs are 0.0001 and 0.0000 (to four-digit precision).

Similar to the results of Section 4.1.2, when the DQN plays with co-players who follow the Sterman formula, it performs far better than **Strm-BS**. As Table 4 shows, DQN performs 34% better than **Strm-BS** on average. Finally, when DQN plays with co-players who use the random

**Table 4 Results of DQN playing with co-players who follow Sterman policy.**

	Uniform			Normal			Classic		
	DQN	Strm-BS	Gap (%)	DQN	Strm-BS	Gap (%)	DQN	Strm-BS	Gap (%)
R	6.88	8.99	-23.45	9.98	10.67	-6.44	3.80	13.28	-71.41
W	5.90	9.53	-38.10	7.11	10.03	-29.06	2.85	8.17	-65.08
D	8.35	10.99	-23.98	8.49	13.83	-38.65	3.82	20.07	-80.96
M	12.36	13.90	-11.07	13.86	15.37	-9.82	15.80	19.96	-20.82
Average			-24.15			-20.99			-59.57

**Table 5 Results of DQN playing with co-players who follow random policy.**

	Uniform			Normal			Classic		
	DQN	Rand-BS	Gap (%)	DQN	Rand-BS	Gap (%)	DQN	Rand-BS	Gap (%)
R	31.39	28.24	11.12	13.03	28.39	-54.10	19.99	25.88	-22.77
W	29.62	28.62	3.49	27.87	35.80	-22.15	23.05	23.44	-1.65
D	30.72	28.64	7.25	34.85	38.79	-10.15	22.81	23.53	-3.04
M	29.03	28.13	3.18	37.68	40.53	-7.02	22.36	22.45	-0.42
Average			6.26			-23.36			-6.97

policy, for all demand distributions DQN learns very well to play so as to minimize the total cost of the system, and on average obtains 8% better solutions than Rand-BS.

To summarize, DQN does well regardless of the way the other agents play, and regardless of the demand distribution. The DQN agent learns to attain near-BS costs when its co-players follow a BS policy, and when playing with irrational co-players, it achieves a much smaller cost than a base-stock policy would. Thus, when the other agents play irrationally, DQN should be used.

#### 4.3. Faster Training through Transfer Learning

We trained a DQN network with shape  $[50, 180, 130, 61, 5]$ ,  $m = 10$ ,  $\beta = 20$ , and  $C = 10000$  for each agent, with the same holding and stockout costs and action spaces as in section 4.1, using 60000 training episodes, and used these as the base networks for our transfer learning experiment. (In transfer learning, all agents should have the same network structure to share the learned network among different agents.) The remaining agents use a BS policy.

Table 6 shows a summary of the results of the six cases discussed in Section 3.2. The first set of columns indicates the holding and shortage cost coefficients, the size of the action space, as well as the demand distribution and the co-players' policy for the base agent (first row) and the target agent (remaining rows). The “Gap” column indicates the average gap between the cost of the resulting DQN and the cost of a BS policy; in the first row, it is analogous to the 2.31% average gap reported

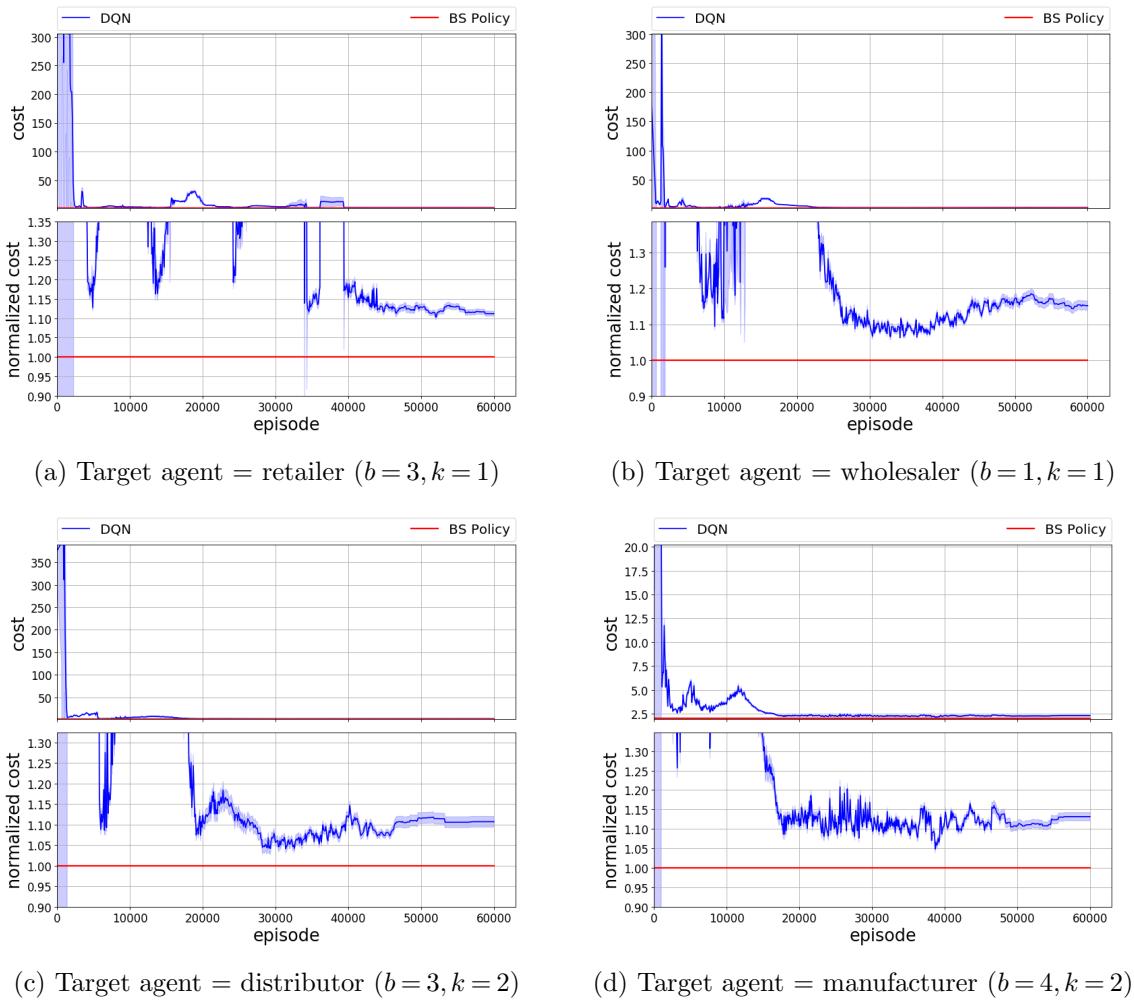
**Table 6 Results of transfer learning when  $\pi_1$  is BS and  $D_1$  is  $\mathbb{U}[0,2]$**

	(Holding, Shortage) Cost Coefficients				A	$D_2$	$\pi_2$	Gap (%)	CPU Time (sec)
	R	W	D	M					
Base agent	(2,2)	(2,0)	(2,0)	(2,0)	5	$\mathbb{U}[0,2]$	BS	2.31	28,390,987
Case 1	(2,2)	(2,0)	(2,0)	(2,0)	5	$\mathbb{U}[0,2]$	BS	6.06	1,593,455
Case 2	(5,1)	(5,0)	(5,0)	(5,0)	5	$\mathbb{U}[0,2]$	BS	6.16	1,757,103
Case 3	(2,2)	(2,0)	(2,0)	(2,0)	11	$\mathbb{U}[0,2]$	BS	10.66	1,663,857
Case 4	(10,1)	(10,0)	(10,0)	(10,0)	11	$\mathbb{U}[0,2]$	BS	12.58	1,593,455
Case 5	(1,10)	(0.75,0)	(0.5,0)	(0.25,0)	11	$N(10, 2^2)$	BS	17.41	1,234,461
Case 6	(1,10)	(0.75,0)	(0.5,0)	(0.25,0)	11	$N(10, 2^2)$	Strm	-38.20	1,153,571
Case 6	(1,10)	(0.75,0)	(0.5,0)	(0.25,0)	11	$N(10, 2^2)$	Rand	-0.25	1,292,295

in Section 4.1.1. The average gap is relatively small in all cases, which shows the effectiveness of the transfer learning approach. Moreover, this approach is efficient, as demonstrated in the last column, which reports the average CPU times for all agent. In order to get the base agents, we did hyper-parameter tuning and trained 140 instances to get the best possible set of hyper-parameters, which resulted in a total of 28,390,987 seconds of training. However, using the transfer learning approach, we do not need any hyper-parameter tuning; we only need to check which source agent and which  $k$  provides the best results. This requires only 12 instances to train and resulted in an average training time (across case 1-4) of 1,613,711 seconds—17.6 times faster than training the base agent. Additionally, in case 5, in which a normal distribution is used, full hyper-parameter tuning took 20,396,459 seconds, with an average gap of 4.76%, which means transfer learning was 16.6 times faster on average. We did not run the full hyper-parameter tuning for the instances of case-6, but it is similar to that of case-5 and should take similar training time, and as a result a similar improvement from transfer learning. Thus, once we have a trained agent  $i$  with a given set  $P_1^i$  of parameters, demand  $D_1$  and co-players' policy  $\pi_1$ , we can efficiently train a new agent  $j$  with parameters  $P_2^j$ , demand  $D_2$  and co-players' policy  $\pi_2$ .

In order to get more insights about the transfer learning process, Figure 7 shows the results of case 4, which is a quite complex transfer learning case that we test for the beer game. The target agents have holding and shortage costs (10,1), (10,0), (10,0), and (10,0) for agents 1 to 4, respectively; and each agent can select any action in  $\{-5, \dots, 5\}$ . Each caption reports the base agent (shown by b) and the value of  $k$  used. Compared to the original procedure (see Figure 4),

**Figure 7 Results of transfer learning for case 4 (different agent, cost coefficients, and action space)**



i.e.,  $k = 0$ , the training is less noisy and after a few thousand non-fluctuating training episodes, it converges into the final solution. The resulting agents obtain costs that are close to those of BS, with a 12.58% average gap compared to the BS cost. (The details of the other cases are provided in Sections D.1—D.5 of the online supplement.)

Finally, Table 7 explores the effect of  $k$  on the tradeoff between training speed and solution accuracy. As  $k$  increases, the number of trainable variables decreases and, not surprisingly, the CPU times are smaller but the costs are larger. For example, when  $k = 3$ , the training time is 46.89% smaller than the training time when  $k = 0$ , but the solution cost is 17.66% and 0.34% greater than the BS policy, compared to 4.22% and -11.65% for  $k = 2$ .

**Table 7 Savings in computation time due to transfer learning.** First row provides average training time among all instances. Third row provides average of the best obtained gap in cases for which an optimal solution exists. Fourth row provides average gap among all transfer learning instances, i.e., cases 1–6.

	$k = 0$	$k = 1$	$k = 2$	$k = 3$
Training time	185,679	126,524	118,308	107,711
Decrease in time compared to $k = 0$	—	37.61%	41.66%	46.89%
Average gap in cases 1–4	2.31%	4.39%	4.22%	17.66%
Average gap in cases 1–6	—	-15.95%	-11.65%	0.34%

To summarize, transferring the acquired knowledge between the agents is very efficient. The target agents achieve costs that are close to those of the BS policy (when co-players follow BS) and they achieve smaller costs than Strm-BS and Rand-BS, regardless of the dissimilarities between the source and the target agents. The training of the target agents start from relatively small cost values, the training trajectories are stable and fairly non-noisy, and they quickly converge to a cost value close to that of the BS policy or smaller than Strm-BS and Rand-BS. Even when the action space and costs for the source and target agents are different, transfer learning is still quite effective, resulting in a 12.58% gap compared to the BS policy. This is an important result, since it means that if the settings change—either within the beer game or in real supply chain settings—we can train new DQN agents much more quickly than we could if we had to begin each training from scratch.

## 5. Conclusion and Future Work

In this paper, we consider the beer game, a decentralized, multi-agent, cooperative supply chain problem. A base-stock inventory policy is known to be optimal for special cases, but once some of the agents do not follow a base-stock policy (as is common in real-world supply chains), the optimal policy of the remaining players is unknown. To address this issue, we propose an algorithm based on deep Q-networks. It obtains near-optimal solutions when playing alongside agents who follow a base-stock policy and performs much better than a base-stock policy when the other agents use a more realistic model of ordering behavior. Furthermore, the algorithm does not require knowledge of the demand probability distribution and uses only historical data.

To reduce the computation time required to train new agents with different cost coefficients or action spaces, we propose a transfer learning method. Training new agents with this approach takes

less time since it avoids the need to tune hyper-parameters and has a smaller number of trainable variables. Moreover, it is quite powerful, resulting in beer game costs that are close to those of fully-trained agents while reducing the training time by an order of magnitude.

A natural extension of this paper is to apply our algorithm to supply chain networks with other topologies, e.g., distribution networks. Another important extension is having multiple learnable agents. Finally, developing algorithms capable of handling continuous action spaces will improve the accuracy of our algorithm.

## References

- D. S. Bernstein, R. Givan, N. Immerman, and S. Zilberstein. The complexity of decentralized control of markov decision processes. *Mathematics of operations research*, 27(4):819–840, 2002.
- S. K. Chaharsooghi, J. Heydari, and S. H. Zegordi. A reinforcement learning model for supply chain ordering management: An application to the beer game. *Decision Support Systems*, 45(4):949–959, 2008.
- F. Chen and R. Samroengraja. The stationary beer game. *Production and Operations Management*, 9(1):19, 2000.
- A. J. Clark and H. Scarf. Optimal policies for a multi-echelon inventory problem. *Management science*, 6(4):475–490, 1960.
- C. Claus and C. Boutilier. The dynamics of reinforcement learning in cooperative multiagent systems. *AAAI/IAAI*, 1998:746–752, 1998.
- R. Croson and K. Donohue. Impact of POS data sharing on supply chain management: An experimental study. *Production and Operations Management*, 12(1):1–11, 2003.
- R. Croson and K. Donohue. Behavioral causes of the bullwhip effect and the observed value of inventory information. *Management Science*, 52(3):323–336, 2006a.
- R. Croson and K. Donohue. Behavioral causes of the bullwhip effect and the observed value of inventory information. *Management science*, 52(3):323–336, 2006b.
- K. Devika, A. Jafarian, A. Hassanzadeh, and R. Khodaverdi. Optimizing of bullwhip effect and net stock amplification in three-echelon supply chains using evolutionary multi-objective metaheuristics. *Annals of Operations Research*, 242(2):457–487, 2016.

- S. Geary, S. M. Disney, and D. R. Towill. On bullwhip in supply chains—historical review, present practice and expected future impact. *International Journal of Production Economics*, 101(1):2–18, 2006.
- I. Giannoccaro and P. Pontrandolfo. Inventory management in supply chains: A reinforcement learning approach. *International Journal of Production Economics*, 78(2):153 – 161, 2002. ISSN 0925-5273. doi: [http://dx.doi.org/10.1016/S0925-5273\(00\)00156-0](http://dx.doi.org/10.1016/S0925-5273(00)00156-0).
- S. C. Graves. A multi-echelon inventory model for a repairable item with one-for-one replenishment. *Management Science*, 31(10):1247–1256, 1985.
- C. Jiang and Z. Sheng. Case-based reinforcement learning for dynamic inventory control in a multi-agent supply-chain system. *Expert Systems with Applications*, 36(3):6520–6526, 2009.
- S. O. Kimbrough, D.-J. Wu, and F. Zhong. Computers play the beer game: Can artificial agents manage supply chains? *Decision support systems*, 33(3):323–333, 2002.
- D. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.
- H. L. Lee, V. Padmanabhan, and S. Whang. Information distortion in a supply chain: The bullwhip effect. *Management Science*, 43(4):546–558, 1997.
- H. L. Lee, V. Padmanabhan, and S. Whang. Comments on “Information distortion in a supply chain: The bullwhip effect”. *Management Science*, 50(12S):1887–1893, 2004.
- Y. Li. Deep reinforcement learning: An overview. *arXiv preprint arXiv:1701.07274*, 2017.
- L.-J. Lin. Self-improving reactive agents based on reinforcement learning, planning and teaching. *Machine Learning*, 8(3-4):293–321, 1992.
- I. J. Martinez-Moyano, J. Rahn, and R. Spencer. The Beer Game: Its History and Rule Changes. Technical report, University at Albany, 2014.
- F. S. Melo and M. I. Ribeiro. Q-learning with linear function approximation. In *International Conference on Computational Learning Theory*, pages 308–322. Springer, 2007.
- V. Mnih, K. Kavukcuoglu, D. Silver, A. Graves, I. Antonoglou, D. Wierstra, and M. Riedmiller. Playing Atari with deep reinforcement learning. *arXiv preprint arXiv:1312.5602*, 2013.

- V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, et al. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–533, 2015.
- E. Mosekilde and E. R. Larsen. Deterministic chaos in the beer production-distribution model. *System Dynamics Review*, 4(1-2):131–147, 1988.
- A. Oroojlooyjadid, L. Snyder, and M. Takáč. Applying deep learning to the newsvendor problem. <http://arxiv.org/abs/1607.02177>, 2017a.
- A. Oroojlooyjadid, L. Snyder, and M. Takáč. Stock-out prediction in multi-echelon networks. *arXiv preprint arXiv:1709.06922*, 2017b.
- S. J. Pan and Q. Yang. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359, 2010.
- P. Rajpurkar, J. Irvin, K. Zhu, B. Yang, H. Mehta, T. Duan, D. Ding, A. Bagul, C. Langlotz, K. Shpanskaya, et al. Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning. *arXiv preprint arXiv:1711.05225*, 2017.
- L. V. Snyder. Multi-echelon base-stock optimization with upstream stockout costs. Technical report, Lehigh University, 2018.
- L. V. Snyder and Z.-J. M. Shen. *Fundamentals of Supply Chain Theory*. John Wiley & Sons, 2nd edition, 2019.
- J. D. Sterman. Modeling managerial behavior: Misperceptions of feedback in a dynamic decision making experiment. *Management Science*, 35(3):321–339, 1989.
- F. Strozzi, J. Bosch, and J. Zaldivar. Beer game order policy optimization under changing customer demand. *Decision Support Systems*, 42(4):2153–2163, 2007.
- R. S. Sutton and A. G. Barto. *Reinforcement learning: An introduction*. MIT Press, Cambridge, 1998.
- A. Tversky and D. Kahneman. Judgment under uncertainty: Heuristics and biases. *Science*, 185(4157):1124–1131, 1979.

# Online Supplements for A Deep Q-Network for the Beer Game: Reinforcement Learning for Inventory Optimization

## Appendix A: Sterman Formula Parameters

The computational experiments that use **Strm** agents calculate the order quantity using formula (3), adapted from Sterman (1989).

$$q_t^i = \max\{0, AO_{t+1}^{i-1} + \alpha^i(IL_t^i - a^i) + \beta^i(OO_t^i - b^i)\} \quad (3)$$

where  $\alpha^i$ ,  $a^i$ ,  $\beta^i$ , and  $b^i$  are the parameters corresponding to the inventory level and on-order quantity. The idea is that the agent sets the order quantity equal to the demand forecast plus two terms that represent adjustments that the agent makes based on the deviations between its current inventory level (resp., on-order quantity) and a target value  $a^i$  (resp.,  $b^i$ ). We set  $a^i = \mu_d$ , where  $\mu_d$  is the average demand;  $b^i = \mu_d(l_i^{fi} + l_i^{tr})$ ;  $\alpha^i = -0.5$ ; and  $\beta^i = -0.2$  for all agents  $i = 1, 2, 3, 4$ . The negative  $\alpha$  and  $\beta$  mean that the player over-orders when the inventory level or on-order quantity fall below the target value  $a_i$  or  $b_i$ .

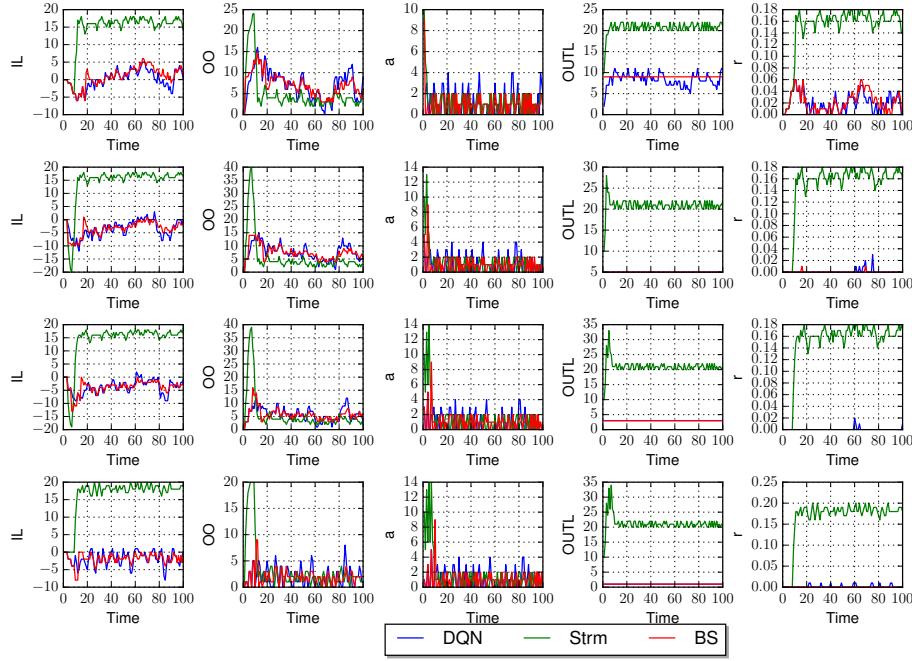
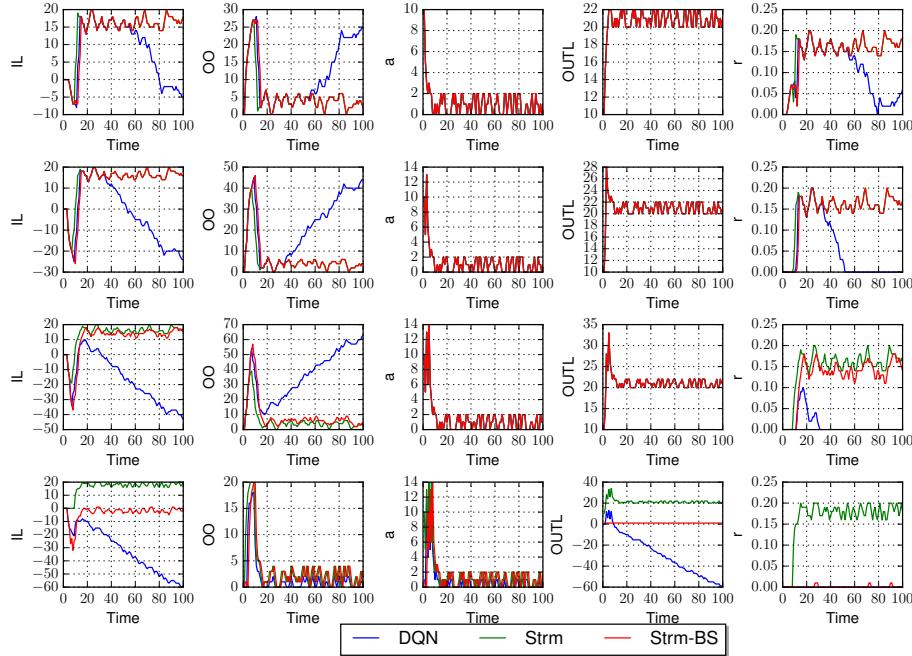
## Appendix B: Extended Numerical Results

This appendix shows additional results on the details of play of each agent. Figure 8 provides the details of  $IL$ ,  $OO$ ,  $a$ ,  $r$ , and OUTL for each agent when the DQN retailer plays with co-players who use the **BS** policy. Clearly, DQN attains a similar IL, OO, action, and reward to those of **BS**. Figure 9 provides analogous results for the case in which the DQN manufacturer plays with three **Strm** agents. The DQN agent learns that the shortage costs of the non-retailer agents are zero and exploits that fact to reduce the total cost. In each of the figures, the top set of charts provides the results of the retailer, followed by the warehouse, distributor, and manufacturer.

## Appendix C: The Effect of $\beta$ on the Performance of Each Agent

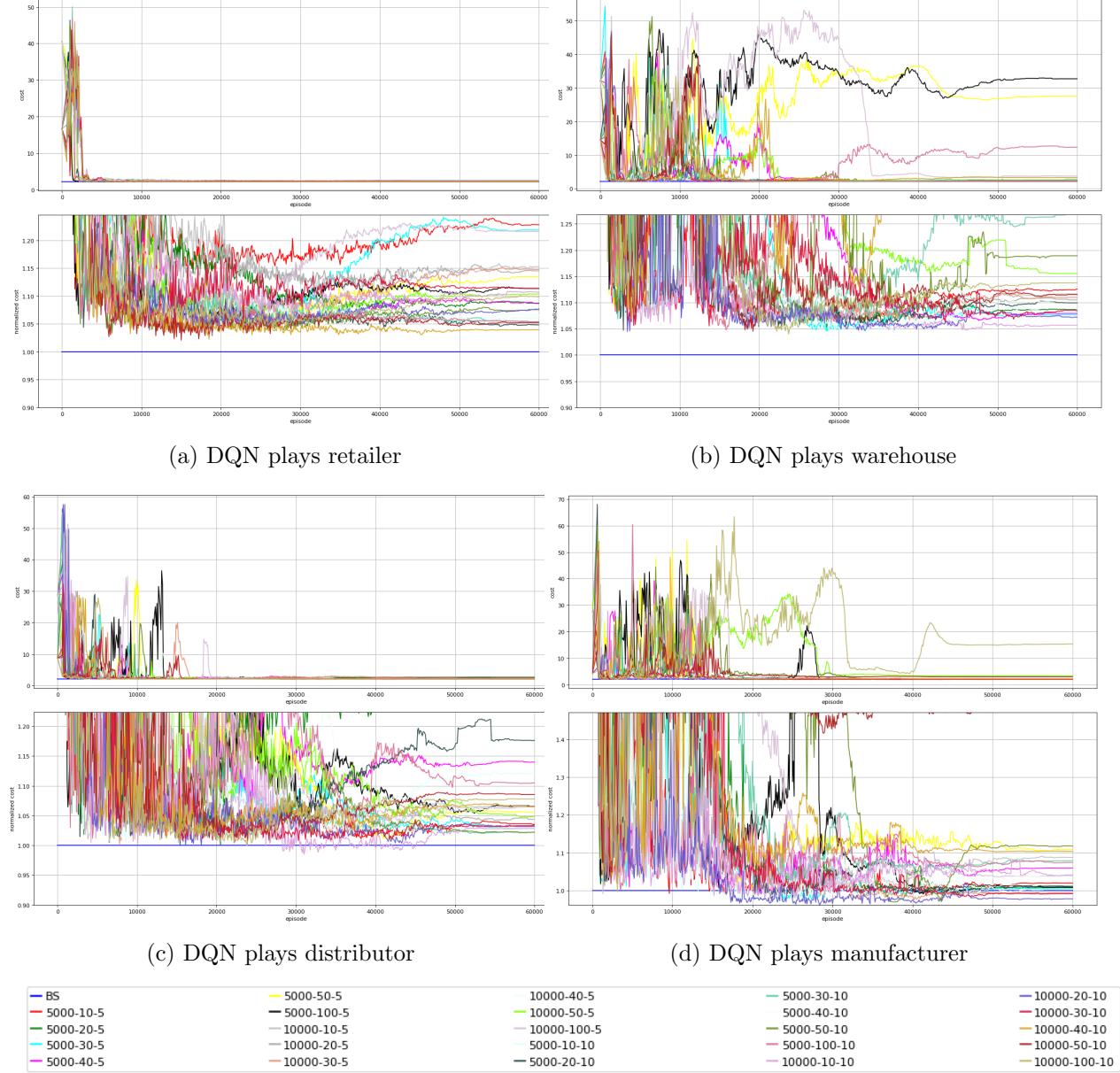
Figure 10 plots the training trajectories for DQN agents playing with three **BS** agents using various values of  $C$ ,  $m$ , and  $\beta$ . In each sub-figure, the blue line denotes the result when all players use a **BS** policy while the remaining curves each represent the agent using DQN with different values of  $C$ ,  $\beta$ , and  $m$ , trained for 60000 episodes with a learning rate of 0.00025.

As shown in Figure 10a, when the DQN plays the retailer,  $\beta_1 \in \{20, 40\}$  works well, and  $\beta_1 = 40$  provides the best results. As we move upstream in the supply chain (warehouse, then distributor, then manufacturer),

**Figure 8**  $IL_t$ ,  $OO_t$ ,  $a_t$ , and  $r_t$  of all agents when DQN retailer plays with three BS co-players

**Figure 9**  $IL_t$ ,  $OO_t$ ,  $a_t$ , and  $r_t$  of all agents when DQN manufacturer plays with three Strm-BS co-players


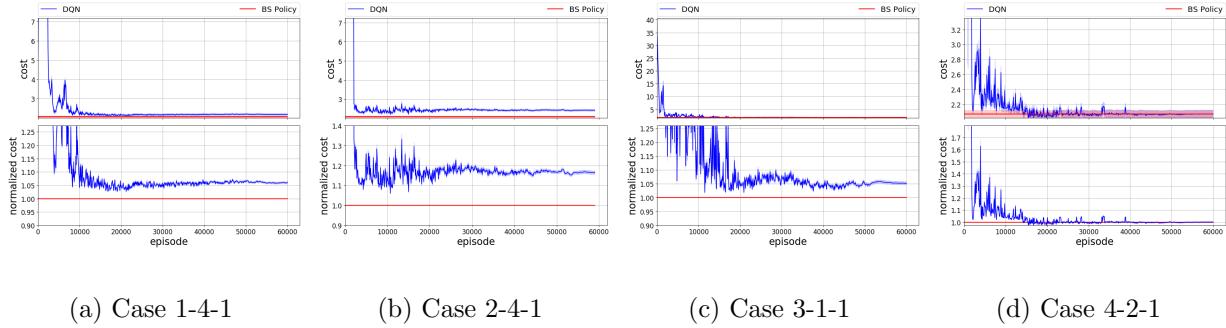
smaller  $\beta$  values become more effective (see Figures 10b–10d). Recall that the retailer bears the largest share of the optimal expected cost per period, and as a result it needs a larger  $\beta$  than the other agents.

**Figure 10 Total cost (upper figure) and normalized cost (lower figure) with one DQN agent and three agents that follow base-stock policy**



Not surprisingly, larger  $m$  values provide better costs since the DQN has more knowledge of the environment. Finally, larger  $C$  works better and provides a stable DQN model. However, there are some combinations for which smaller  $C$  and  $m$  also work well, e.g., see Figure 10d, trajectory 5000-20-5.

**Figure 11 Results of transfer learning between agents with the same cost coefficients and action space**



## Appendix D: Extended Results on Transfer Learning

### D.1. Transfer Knowledge Between Agents

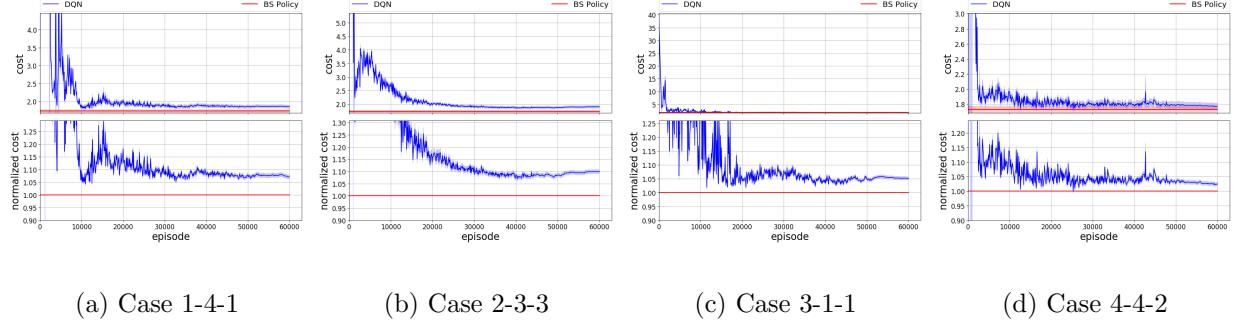
In this section, we present the results of the transfer learning method when the trained agent  $i \in \{1, 2, 3, 4\}$  transfers its first  $k \in \{1, 2, 3\}$  layer(s) into co-player agent  $j \in \{1, 2, 3, 4\}$ ,  $j \neq i$ . For each target-agent  $j$ , Figure 11 shows the results for the best source-agent  $i$  and the number of shared layers  $k$ , out of the 9 possible choices for  $i$  and  $k$ . In the sub-figure captions, the notation  $j$ - $i$ - $k$  indicates that source-agent  $i$  shares weights of the first  $k$  layers with target-agent  $j$ , so that those  $k$  layers remain non-trainable.

Except for agent 2, all agents obtain costs that are very close to those of the BS policy, with a 6.06% gap, on average. (In Section 4.1.1, the average gap is 2.31%) However, none of the agents was a good source for agent 2. It seems that the acquired knowledge of other agents is not enough to get a good solution for this agent, or the feature space that agent 2 explores is different from other agents, so that it cannot get a solution whose cost is close to the BS cost.

In order to get more insight, consider Figure 4, which presents the best results obtained through hyper-parameter tuning for each agent. In that figure, all agents start the training with a large cost value, and after 25000 fluctuating iterations, each converges to a stable solution. In contrast, in Figure 11, each agent starts from a relatively small cost value, and after a few thousand training episodes converges to the final solution. Moreover, for agent 3, the final cost of the transfer learning solution is smaller than that obtained by training the network from scratch. And, the transfer learning method used one order of magnitude less CPU time than the approach in Section 4.1.1 to obtain very close results.

We also observe that agent  $j$  can obtain good results when  $k = 1$  and  $i$  is either  $j - 1$  or  $j + 1$ . This shows that the learned weights of the first DQN network layer are general enough to transfer knowledge to the

**Figure 12 Results of transfer learning between agents with different cost coefficients and same action space**



other agents, and also that the learned knowledge of neighboring agents is similar. Also, for any agent  $j$ , agent  $i = 1$  provides similar results to that of agent  $i = j - 1$  or  $i = j + 1$  does, and in some cases it provides slightly smaller costs, which shows that agent 1 captures general feature values better than the others.

## D.2. Transfer Knowledge for Different Cost Coefficients

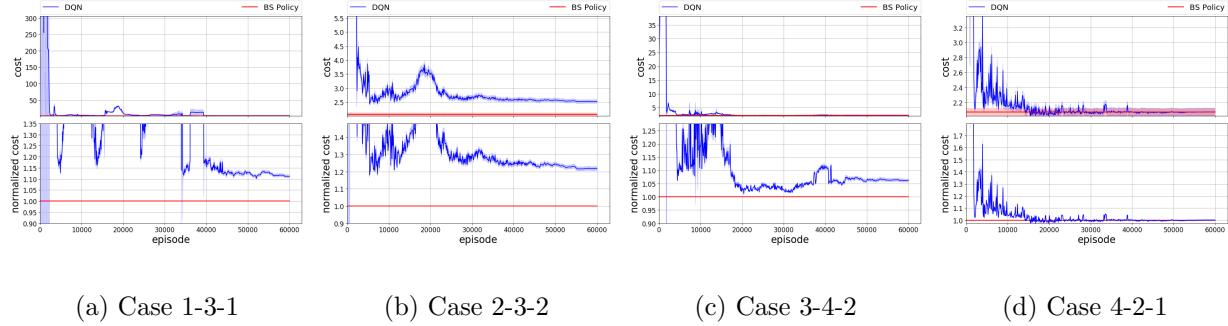
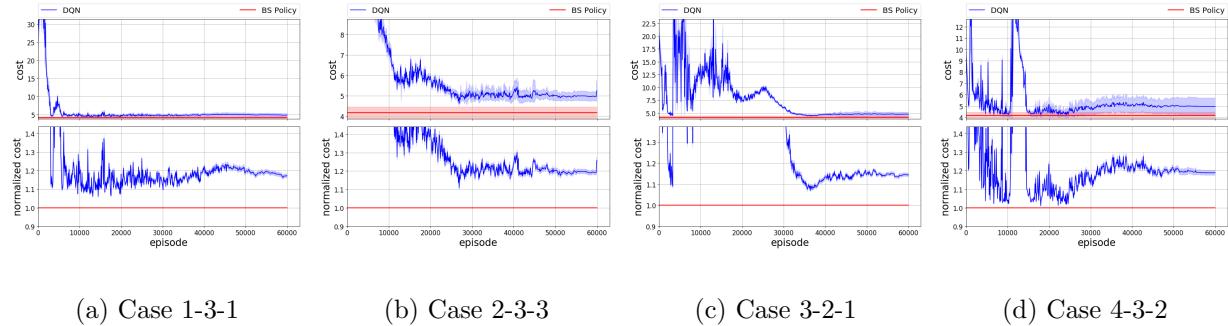
Figure 12 shows the best results achieved for all agents, when agent  $j$  has different cost coefficients,  $(c_{p_2}, c_{h_2}) \neq (c_{p_1}, c_{h_1})$ . We test target agents  $j \in \{1, 2, 3, 4\}$ , such that the holding and shortage costs are  $(5, 1)$ ,  $(5, 0)$ ,  $(5, 0)$ , and  $(5, 0)$  for agents 1 to 4, respectively. In all of these tests, the source and target agents have the same action spaces. All agents attain cost values close to the BS cost; in fact, the overall average cost is 6.16% higher than the BS cost.

In addition, similar to the results of Section D.1, base agent  $i = 1$  provides good results for all target agents. We also performed the same tests with shortage and holding costs  $(10, 1)$ ,  $(1, 0)$ ,  $(1, 0)$ , and  $(1, 0)$  for agents 1 to 4, respectively, and obtained very similar results.

## D.3. Transfer Knowledge for Different Size of Action Space

Increasing the size of the action space should increase the accuracy of the  $d + x$  approach. However, it makes the training process harder. It can be effective to train an agent with a small action space and then transfer the knowledge to an agent with a larger action space. To test this, we test target-agent  $j \in \{1, 2, 3, 4\}$  with action space  $\{-5, \dots, 5\}$ , assuming that the source and target agents have the same cost coefficients.

Figure 13 shows the best results achieved for all agents. All agents attained costs that are close to the BS cost, with an average gap of approximately 10.66%.

**Figure 13 Results of transfer learning for agents with  $|A_1| \neq |A_2|, (c_{p1}^j, c_{h1}^j) = (c_{p2}^j, c_{h2}^j)$** **Figure 14 Results of transfer learning for agents with  $|A_1| \neq |A_2|, (c_{p1}^j, c_{h1}^j) \neq (c_{p2}^j, c_{h2}^j), D_1 \neq D_2$ , and  $\pi_1 \neq \pi_2$** 

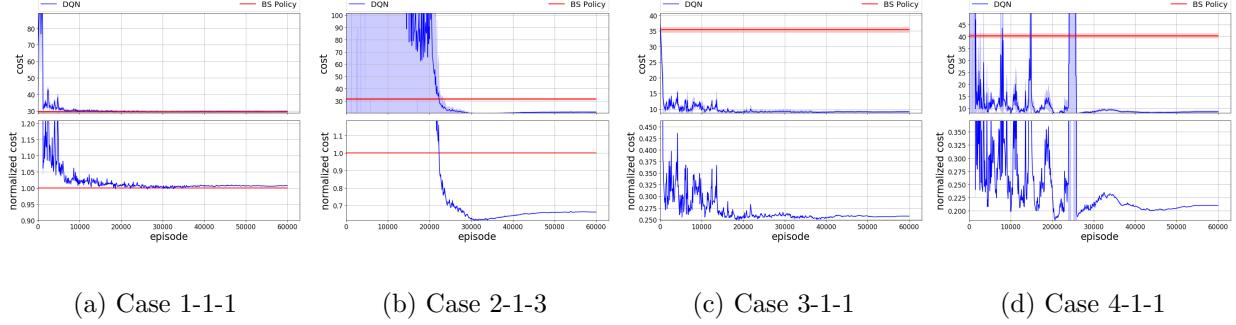
#### D.4. Transfer Knowledge for Different Action Space, Cost Coefficients, and Demand Distribution

This case includes all difficulties of the cases in Sections D.1, D.2, D.3, and 4.3, in addition to the demand distributions being different. So, the range of demand,  $IL$ ,  $OO$ ,  $AS$ , and  $AO$  that each agent observes is different than those of the base agent. Therefore, this is a hard case to train, and the average optimality gap is 17.41%; however, as Figure 14 depicts, the cost values decrease quickly and the training noise is quite small.

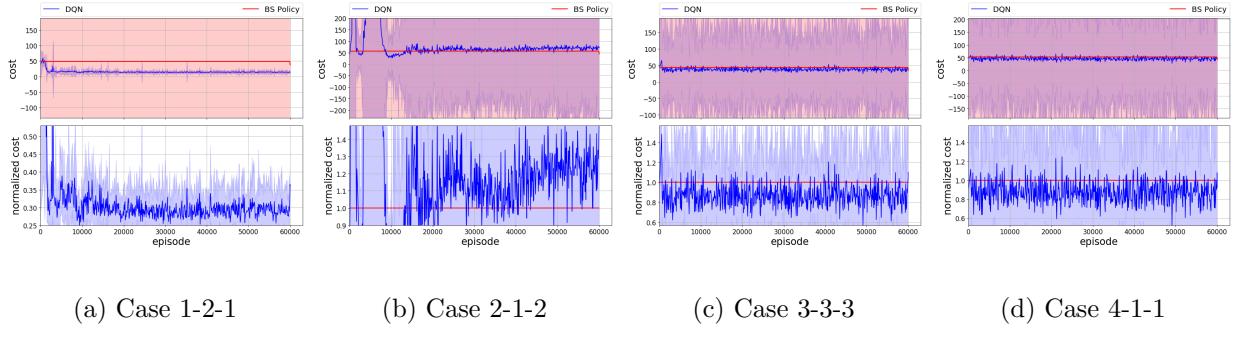
#### D.5. Transfer Knowledge for Different Action Space, Cost Coefficients, Demand Distribution, and $\pi_2$

Figures 15 and 16 show the results of the most complex transfer learning cases that we tested. Although the DQN plays with non-rational co-players and the observations in each state might be quite noisy, there are relatively small fluctuations in the training, and for all agents after around 40,000 iterations they converge.

**Figure 15 Results of transfer learning for agents with  $|A_1| \neq |A_2|, (c_{p1}^j, c_{h1}^j) \neq (c_{p2}^j, c_{h2}^j), D_1 \neq D_2$ , and  $\pi_1 \neq \pi_2$**



**Figure 16 Results of transfer learning for agents with  $|A_1| \neq |A_2|, (c_{p1}^j, c_{h1}^j) \neq (c_{p2}^j, c_{h2}^j), D_1 \neq D_2$ , and  $\pi_1 \neq \pi_2$**



## Appendix E: Pseudocode of the Beer Game Simulator

The DQN algorithm needs to interact with the environment, so that for each state and action, the environment should return the reward and the next state. We simulate the beer game environment using Algorithm 2. In addition to the notation defined earlier, the algorithm also uses the following notation:

$d^t$ : The demand of the customer in period  $t$ .

$OS_i^t$ : Outbound shipment from agent  $i$  (to agent  $i - 1$ ) in period  $t$ .

---

**Algorithm 2** Beer Game Simulator Pseudocode
 

---

```

1: procedure PLAYGAME
2:   Set  $T$  randomly, and  $t = 0$ , Initialize  $IL_i^0$  for all agents,  $AO_i^t = 0, AS_i^t = 0, \forall i, t$ 
3:   while  $t \leq T$  do
4:      $AO_i^{t+l_i^{f_i}} += d^t$                                  $\triangleright$  set the retailer's arriving order to external demand
5:     for  $i = 1 : 4$  do                                 $\triangleright$  loop through stages downstream to upstream
6:       get action  $a_i^t$                                  $\triangleright$  choose order quantity
7:        $OO_i^{t+1} = OO_i^t + a_i^t$                        $\triangleright$  update  $OO_i$ 
8:        $AO_{i+1}^{t+l_i^{f_i}} += a_i^t$                    $\triangleright$  propagate order upstream
9:     end for
10:     $AS_4^{t+l_4^{r_i}} += a_4^t$                        $\triangleright$  set manufacturer's arriving shipment to its order quantity
11:    for  $i = 4 : 1$  do                                 $\triangleright$  loop through stages upstream to downstream
12:       $IL_i^{t+1} = IL_i^t + AS_i^t$                    $\triangleright$  receive inbound shipment
13:       $OO_i^{t+1} -= AS_i^t$                            $\triangleright$  update  $OO_i$ 
14:      current_Inv =  $\max\{0, IL_i^{t+1}\}$            $\triangleright$  determine outbound shipment
15:      current_BackOrder =  $\max\{0, -IL_i^t\}$ 
16:       $OS_i^t = \min\{ \text{current\_Inv}, \text{current\_BackOrder} + AO_i^t \}$ 
17:       $AS_{i-1}^{t+l_i^{r_i}} += OS_i^t$                    $\triangleright$  propagate order downstream
18:       $IL_i^{t+1} -= AO_i^t$                            $\triangleright$  update  $IL_i$ 
19:       $c_i^t = c_i^p \max\{-IL_i^{t+1}, 0\} + c_i^h \max\{IL_i^{t+1}, 0\}$   $\triangleright$  calculate cost
20:    end for
21:     $t += 1$ 
22:  end while
23: end procedure
  
```

---