

# NEURON AS AN AGENT

**Anonymous authors**

Paper under double-blind review

## ABSTRACT

Existing multi-agent reinforcement learning (MARL) communication methods have relied on a trusted third party (TTP) to distribute incentive to agents, leaving them inapplicable in peer-to-peer environments. This paper proposes incentive distribution using *Neuron as an Agent* (NaaA) in MARL without a TTP with two key ideas: (i) inter-agent reward distribution and (ii) auction theory. Auction theory is introduced because inter-agent reward distribution is insufficient for optimization. Agents in NaaA maximize their profits (the difference between reward and cost) and, as a theoretical result, the auction mechanism is shown to have agents autonomously evaluate counterfactual returns as the values of other agents. NaaA enables representation trades in peer-to-peer environments, ultimately regarding unit in neural networks as agents. Finally, numerical experiments (a single-agent environment from Open AI gym and a multi-agent environment from ViZDoom) confirm that NaaA framework optimization leads to better performance in RL.

## 1 INTRODUCTION

Now that intelligence with reinforcement learning has been shown to be superior to humans (Tesauro, 1995; Mnih et al., 2015; Silver et al., 2016), reinforcement learning is expected to expand into industry applications such as stock trading, autonomous cars, smart grids, and IoT. In future industrial applications, numerous companies will own agents used to improve their revenues. Such a situation can be considered as each agent independently solving the problems of a partially observed Markov decision process (POMDP).

These company agents are designed to independently maximize their own rewards; however, if such agents could exchange information, the overall revenue of stakeholders would increase. As each agent has limited visibility of the environment, information exchanges between agents would be helpful in solving their independent tasks. Thus, this paper aimed to actualize a society where stakeholders with conflicting interests willingly trade information.

We regard the situation as communication in multi-agent reinforcement learning (MARL). Multi-agent communication is addressed by several methods such as R/DIAL (Foerster et al., 2016) and CommNet (Sukhbaatar et al., 2016). CommNet is a state-of-the-art MARL method that considers communication between agents and features learning among agents with backpropagation.

When considering conditions in MARL such that different stakeholders have created different agents that communicate with each other, an incentive distribution design (e.g., monetary payment) and a framework without a *trusted third party* (TTP) are required. TTP (Wu & Soo, 1999; Sandholm & Wang, 2002) is a neutral administrator that performs reward distributions for all participants, and is supposed implicitly by most existing literature regarding MARL (Agogino & Tumer, 2006; Foerster et al., 2016; Sukhbaatar et al., 2016). Although there is a requirement for TTP neutrality towards all participants, several peer-to-peer trade configurations, such as inter-industry and -country trades, cannot place TTP. If reward distribution is performed by an untrusted third party, the rewards for partial participants may be undesirably altered.

To the best of our knowledge, no existing literature discusses reward distributions in the configuration described above. Because CommNet assumes an environment that distributes a uniform reward to all the agents, if the distributed reward is in limited supply (such as money), it causes the *Tragedy of the Commons* (Lloyd, 1833), where the reward of contributing agents will be reduced due to the participation of free riders. Although there are several MARL methods for distributing rewards ac-

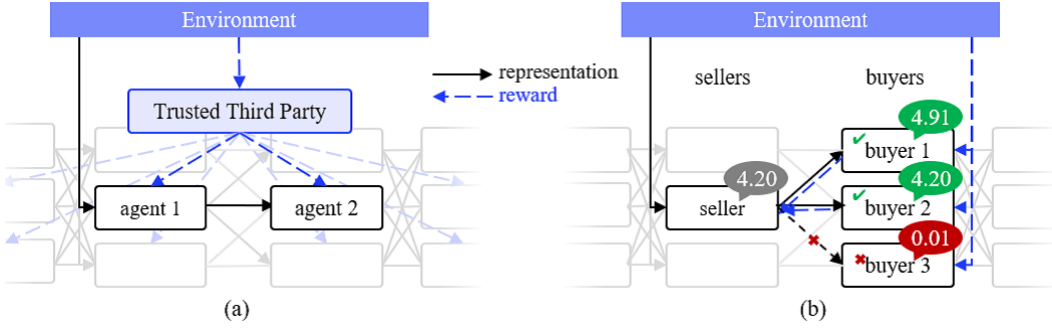


Figure 1: A schematic illustration of reward distribution models in MARL. **(a)** Centralized reward distribution model (Agogino & Tumer, 2006; Sukhbaatar et al., 2016; Foerster et al., 2016; 2017). They should suppose TTP to distribute the optimal reward to the agents. **(b)** Inter-agent reward distribution model (our model). Some agents receive reward from the environment directly, and redistribute to other agents. The idea to determine the optimal reward without TTP is playing auction game among the agents.

cording to agents’ contribution such as QUICR (Agogino & Tumer, 2006) and COMA (Sukhbaatar et al., 2016), they suppose the existence of TTP and hence cannot be applied to the situation investigated here.

The proposed method, *Neuron as an Agent* (NaaA), extends CommNet to actualize incentive distributions in MARL without TTP based on two key ideas: (i) inter-agent reward distribution and (ii) auction theory. Auction theory was introduced because inter-agent reward distributions were insufficient for optimization. Agents in NaaA maximize *profit*, the difference between their received rewards and the costs which they redistribute to other agents. If the framework is naively optimized, a trivial solution is obtained where agents reduce their costs to zero to maximize profits. Then, NaaA employs the auction theory in game design to prevent costs from dropping below their necessary level. As a theoretical result, we show that agents autonomously evaluate the *counterfactual return* as values of other agents. The counterfactual return is equal to the discounted cumulative sum of counterfactual reward (Agogino & Tumer, 2006) distributed by QUICR and COMA. NaaA actualizes reward distributions, making it more of a pareto improvement than inter-agent reward distribution method.

NaaA enables representation trades in peer-to-peer environments and, ultimately, regards neural network units as agents. As NaaA is capable of regarding units as agents without losing generality, this setting was utilized in the current study. The concept of the proposed method is illustrated in Figure 1.

An environment extending ViZDoom (Kempka et al., 2016), a POMDP environment, to MARL was used for the experiment. Two agents, a cameraman sending information and a main player defeating enemies with a gun, were placed in the environment. Results confirmed that the cameraman learned cooperative actions for sending information from dead angles (behind the main player) and outperformed CommNet in score.

Interestingly, NaaA can apply to single- and multi-agent settings, since it learns optimal topology between the units. *Adaptive DropConnect* (ADC), which combines DropConnect (Wan et al., 2013) (randomly masking topology) with an adaptive algorithm (which has a higher probability of pruning connections with lower counterfactual returns) was proposed as a further application for NaaA. Experimental classification and reinforcement learning task results showed ADC outperformed DropConnect.

The remainder of this paper is organized as follows. First, the two key ideas are described: inter-agent reward distribution and auction theory. After related works are introduced, the experimental results are shown in ViZDoom. Subsequently, ADC and its experimental results are shown to demonstrate a further application. Finally, a conclusion ends the paper.

## 2 PROBLEM DEFINITION

Suppose there are  $N$  agents interacting in an environment. The goal of this paper was to maximize the discounted cumulative reward the system obtained from the external environment. This was calculated as:

$$G = \sum_{i=1}^N \left[ \sum_{t=0}^T \gamma^t R_{it}^{\text{ex}} \right], \quad (1)$$

where  $R_{it}^{\text{ex}}$  is an external reward which  $i$ -th agent obtains at  $t$ , and  $\gamma \in [0, 1]$  is the discount rate and  $T$  is the terminal time.

## 3 PROPOSED METHOD

### 3.1 INTER-AGENT REWARD DISTRIBUTION

Some agents receive rewards from the environment, and they distribute these to other agents as incentives for giving precise information. Rewards are limited, so if an agent distributes  $\rho$  rewards, the reward of that agents is reduced by  $\rho$ , as it would be with currency. For this reason, agents other than a specified agent of interest can be considered a secondary environment for the agent giving rewards of  $-\rho$  instead of an observation  $x$ . This secondary environment was termed the *internal environment*, whereas the original environment was called the *external environment*.

Similarly to CommNet (Sukhbaatar et al., 2016), the communication protocol between agents was assumed to be a continuous quantity (such as a vector), the content of which could be trained by backpropagation. Under these conditions, it is possible to interpret multi-agent communications as a huge neural network. This allowed units to be interpreted as agents without losses of generality. This framework can be applied to both single- and multi-agent reinforcement learning.

A typical artificial neural network can be represented as a directed graph  $\mathfrak{G} = (\mathcal{V}, \mathcal{E})$  between units, where  $\mathcal{V} = \{v_1, \dots, v_N\}$  is a set of the units and  $\mathcal{E} \subset \mathcal{V}^2$  is a set of edges representing the connections between two units. If  $(v_i, v_j) \in \mathcal{E}$ , then connection  $v_i \rightarrow v_j$  holds, indicating that  $v_j$  observes the activation of  $v_i$ . Here, the activation of unit  $v_i$  at time  $t$  was denoted as  $x_{it} \in \mathbb{R}$ . Additionally, the set of units that unit  $i$  connects to was designated to be  $N_i^{\text{out}} = \{j | (v_i, v_j) \in \mathcal{E}\}$  and the set of units that unit  $i$  is connected from was  $N_i^{\text{in}} = \{j | (v_j, v_i) \in \mathcal{E}\}$ . Furthermore,  $N_i = N_i^{\text{in}} \cup N_i^{\text{out}}$ .

The proposed NaaA method interprets  $v_i$  as an agent, making  $\mathfrak{G}$  a multi-agent system. A  $v_i$  environment was comprised of an environment where the multi-agent system itself touched a set of units to which  $v_i$  directly connects:  $\{v_i \in \mathcal{V} | i \in N_i\}$ . The former was defined as the external environment, whereas the latter was termed the internal environment.  $v_i$  received rewards from both environments. The following assumptions were added to the  $v_i$  characteristics:

- N1: (Selfishness) The utility each unit  $v_i$  wants to maximize is its own return (cumulative discounted reward):  $G_{it} = \sum_{k=0}^T \gamma^k R_{i,t+k}$ .
- N2: (Conservation) The summation of internal rewards over all  $\mathcal{V}$  equals 0. Hence, the summation of rewards which  $\mathcal{V}$  (receive both internal and external environment  $R_{it}$ ) are equivalent to the reward  $R_t^{\text{ex}}$ , which the entire multi-agent system receives from the external environment:  $\sum_{i=1}^n R_{it} = \sum_{i=1}^n R_{it}^{\text{ex}}$ .
- N3: (Trade) A unit  $v_i$  will receive an internal reward  $\rho_{jit}$  from  $v_j \in \mathcal{V}$  in exchange for an activation signal  $x_i$  before transferring this signal to the unit. Simultaneously,  $\rho_{jit}$  will be subtracted from the reward of  $v_j$ .
- N4: (NOOP)  $v_i$  can select NOOP (no operation), for which the return is  $\delta > 0$ , as an action. In NOOP, the unit inputs and outputs nothing.

The social welfare function (total utility of the agents)  $G^{\text{all}}$  is equivalent to the objective function  $G$ . That is,

$$G^{\text{all}} = \sum_{i=1}^n G_{it} = \sum_{i=1}^n \left[ \sum_{k=0}^T \gamma^k R_{it} \right] = \sum_{k=0}^T \left[ \gamma^k \sum_{i=1}^n R_{it} \right]. \quad (2)$$

From N2,  $G^{\text{all}} = G$  holds.

### 3.1.1 DISCOUNTED CUMULATIVE PROFIT MAXIMIZATION

The external reward unit  $v_i$  received at time step  $t$  was denoted as  $R_{it}^{\text{ex}}$ , where  $\sum_{i=1}^n R_{it}^{\text{ex}} = R_t^{\text{ex}}$  holds. From N3, the reward  $R_{it}$  received by  $v_i$  at  $t$  can be written as:

$$R_{it} = R_{it}^{\text{ex}} + \sum_{j \in N_i^{\text{out}}} \rho_{jit} - \sum_{j \in N_i^{\text{in}}} \rho_{ijt}. \quad (3)$$

which can be divided into positive and negative terms, where the former is defined as revenue, and the latter as cost. These are respectively denoted as  $r_{it} = R_{it}^{\text{ex}} + \sum_{j \in N_i^{\text{out}}} \rho_{jit}$ ,  $c_{it} = \sum_{j \in N_i^{\text{in}}} \rho_{ijt}$ . Here,  $R_{it}$  represents profit.

The unit  $v_i$  maximized its cumulative discounted profit,  $G_{it}$ , represented as:

$$G_{it} = \sum_{k=0}^T \gamma^k R_{i,t+k} = \sum_{k=0}^T \gamma^k (r_{i,t+k} - c_{i,t+k}) = r_t - c_t + \gamma G_{i,t+1}. \quad (4)$$

$G_{it}$  could not be observed until the end of an episode (the final time). Because predictions based on current values were needed to select optimal actions,  $G_{it}$  was approximated with the value function  $V_i^{\pi_i}(s_{it}) = \mathbb{E}_{\pi_i} [G_{it} | s_{it}]$  where  $s_{it} \in \mathcal{S}_O$ . Under these conditions, the following equation holds:

$$V_i^{\pi_i}(s_{it}) = r_{it} - c_{it} + \gamma V_i^{\pi_i}(s_{i,t+1}), \quad (5)$$

Thus, herein, we only consider maximization of revenue, the value function, and cost minimization. The inequality  $R_{it} > 0$  (i.e.,  $r_{it} > c_{it}$ ) indicates that the unit in question gave additional value to the obtained data. The unit selected the NOOP action because  $V_i^{\pi_i}(s_{it}) \leq 0 < \delta$  if  $R_{it} \leq 0$  for all  $t$  (because of N4).

### 3.1.2 SOCIAL DILEMMA

#### STUB

### 3.2 AUCTION THEORY

Mechanism design was introduced because, unlike the methods utilized in existing studies (Sukhbaatar et al., 2016), NaaA assumes that agents are selfish, not cooperative. If the optimization problem of NaaA is itself naively optimized, a trivial solution is obtained in which the internal rewards converge at 0, and all units (excepting output units) select NOOP as their action. This phenomenon occurs regardless of the network topology  $\mathcal{G}$ , as no nodes are incentivized to send payments  $\rho_{ijt}$  to other units. With this in mind, multi-agent systems must select actions with no information, achieving the equivalent of taking random actions. For that reason, the external reward  $R_t^{\text{ex}}$  shrinks markedly.

#### 3.2.1 ENVY-FREE AUCTION

To make the agents to evaluate other agents' value truthfully, the proposed objective function borrows its idea from the digital goods auction theory Guruswami et al. (2005). This auction theory is a part of mechanism design intended to unveil the true price of goods. Digital goods auctions are one mechanism developed from auction theory, specifically targeting goods that may be copied without cost such as digital books and music.

Although several variations of digital goods auctions exist, an envy-free auction (Guruswami et al., 2005) was used here because it required only a simple assumption: equivalent goods have a single simultaneous price. In NaaA, this can be represented by the following assumption:

N5: (Law of one price) If  $\rho_{j_1,i,t}, \rho_{j_2,i,t} > 0$ , then  $\rho_{j_1,i,t} = \rho_{j_2,i,t}$ .

The assumption above indicates that  $\rho_{jit}$  takes either 0 or a positive value, depending on  $i$  at an equal selected time  $t$ . Therefore, the positive side was named  $v_i$ 's *price*, and denoted as  $q_{it}$ .

The envy-free auction process is shown in the left section of Figure 2, displaying the negotiation process between one unit sending an activation (defined as a seller), and a group of units buying the activation (defined as buyers). First, a buyer places a bid with the unit at a bidding price  $b_{jit}$  (1). Next, the seller selects the optimal price  $\hat{q}_{it}$  and allocates the activation (2). Payment occurs if  $b_{jit}$  exceeds  $q_{it}$ . In this case,  $\rho_{jit} = H(b_{jit} - q_{it})q_{it}$  holds where  $H(\cdot)$  is a step function. For this transaction, the definition  $g_{jit} = H(b_{jit} - q_{it})$  holds, and is named *allocation*. After allocation, buyers perform payment:  $\rho_{jit} = g_{jit}\hat{q}_{it}$  (3). The seller sends the activation  $x_i$  only to the allocated buyers (4). Buyers who do not receive the activation approximate  $x_i$  with  $\mathbb{E}_\pi[x_i]$ . This negotiation is performed at each time step in RL.

The sections below discuss the revenue, cost, and value functions based on Eq:(5).

**Revenue:** The revenue of a unit is given as

$$r_{it} = \sum_{j \in N_i^{\text{out}}} g_{jit}q_{it} + R_i^{\text{ex}} = q_i d_{it} + R_i^{\text{ex}}, \quad (6)$$

where  $d_{it} = \sum_{j \in N_i^{\text{out}}} g_{jit}$  is some number of units for which the  $q_{it}$  bidding price is greater than or equal to  $q_{it}$  (designated as the demand). The equation maximizing  $q_{it}$  is designated as the optimal price, denoted as  $\hat{q}_{it}$ . Because  $R_i^{\text{ex}}$  is independent of  $q_t$ , the optimal price  $\hat{q}_{it}$  is given as:

$$\hat{q}_{it} = \operatorname{argmax}_{q \in [0, \infty)} q d_{it}(q). \quad (7)$$

The  $r_{it}$  curve is shown on the right side of Figure 2.

**Cost:** Cost is defined as an internal reward that one unit pays to other units. It is represented as:

$$c_{it} = \sum_{j \in N_i^{\text{in}}} g_{ijt}q_{jt} = \mathbf{g}_{it}^T \mathbf{q}_t, \quad (8)$$

where  $\mathbf{g}_{it} = (g_{i1t}, \dots, g_{iNt})^T$  and  $\mathbf{q}_t = (q_{1t}, \dots, q_{Nt})^T$ . Although  $c_{it}$  itself is minimized when  $b_{ijt} = 0$ , this represents a tradeoff with the following value function.

**Value Function:** The activation  $x_{it}$  depends on input from the units in  $N_i^{\text{in}}$  affecting the bidding price from units in  $N_i^{\text{out}}$ . If  $b_{ijt}$  is minimized and  $b_{ijt} = 0$ , the activation purchase fails and the reward obtained by the seller unit from the connected buyer units is lower in the future.

The effects of the value function were considered for both successful and unsuccessful  $v_j$  purchasing cases. The value function was approximated as a linear function  $\mathbf{g}_{it}$ :

$$V_i^{\pi_i}(s_{i,t+1}) \approx \mathbf{o}_{it}^T \mathbf{g}_{it} + V_{i,t+1}^0, \quad (9)$$

where  $\mathbf{o}_{it}$  is an implemented parameter (the difference between two  $v_i$  returns) regardless of whether  $x_i$  is observed. As  $\mathbf{o}_{it}$  is equivalent to the cumulative discount value of the counterfactual reward (Agogino & Tumer, 2006), it was named *counterfactual return*. As  $V_{it}^0$  (a constant independent of  $\mathbf{g}_{it}$ ) is equal to the value function when  $v_i$  takes an action without observing  $x_1, \dots, x_N$ , it was named *blind value function*.

The optimization problem is, therefore, presented below.

$$\max_{\mathbf{a}} Q_i(s_{it}, \mathbf{a}) = \max_q q d_{it}(q) - \min_{\mathbf{b}} \mathbb{E}_{\hat{q}_t} [\mathbf{g}_{it}(\mathbf{b})^T (\hat{q}_t - \gamma \mathbf{o}_{it})] + \text{const.}, \quad (10)$$

where  $\mathbf{a} = (\mathbf{b}, q)$ . Note that  $\mathbf{g}_{it} = H(\mathbf{b} - \mathbf{q}_t)$ . The expectation  $\mathbb{E}_{\hat{q}_t}[\cdot]$  was taken because the asking price  $\hat{q}_t$  was unknown for  $v_i$ , except when  $\hat{q}_{it}$  and  $g_{iit} = 0$ .

Then, to identify the bidding price that  $b_{it}$  maximizes returns, the following theorem holds.

**Theorem 3.1.** (Truthfulness) the optimal bidding price for maximizing returns is  $\hat{\mathbf{b}}_{it} = \gamma \mathbf{o}_{it}$ .

This proof is shown in the Appendix.

This implies units should only consider their counterfactual returns!. When  $\gamma = 0$  it is equivalent to a case without auction. Hence, the bidding value is raised if each unit considers their long-time rewards. Consequently, when the NaaA mechanism is used units behave as if performing valuation for other units, and declare values truthfully.

Under these conditions, the following corollary holds:

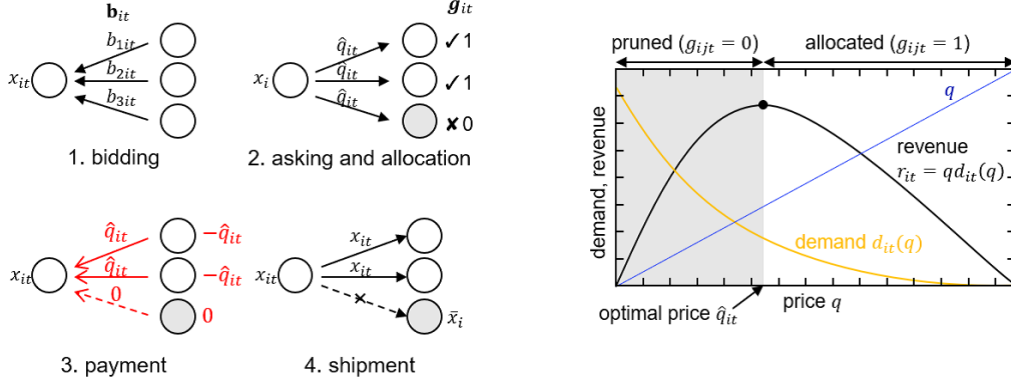


Figure 2: **Left:** Trade process in an envy-free auction. **Right:** Unit price determination curve. Unit revenue is a product of monotonically decreasing demands and prices. The price that maximizes revenue is the optimal price.

---

**Algorithm 1** NaaA: inter-agent reward distribution with envy-free auction

---

```

1: for  $t = 1$  to  $T$  do
2:   Compute a bidding price for every edge: for  $(v_j, v_i) \in \mathcal{E}$  do  $b_{ijt} \leftarrow Q^{\pi_i}(s_{it}, e_j) - Q^{\pi_i}(s_{it}, 0)$ 
3:   Compute an asking price for every node: for  $v_i \in \mathcal{V}$  do  $\hat{q}_{it} \leftarrow \operatorname{argmax}_{q \in [0, \infty)} q d_{it}(q)$ .
4:   for  $(v_i, v_j) \in \mathcal{E}$  do
5:     Compute allocation:  $g_{jit} \leftarrow H(b_{jit} - \hat{q}_{it})$ 
6:     Compute the price the agent should pay:  $\rho_{jit} \leftarrow g_{jit} \hat{q}_{it}$ 
7:   end for
8:   Make a payment: for  $v_i \in \mathcal{V}$  do  $R_{it} \leftarrow \sum_{j \in N_i^{\text{out}}} \rho_{jit} - \sum_{j \in N_i^{\text{in}}} \rho_{ijt}$ ,
9:   Make a shipment: for  $v_i \in \mathcal{V}$  do  $\tilde{x}_{ijt} = g_{ijt} x_{ijt} + (1 - g_{ijt}) \tilde{x}_{ijt}$ 
10:  for  $v_i \in \mathcal{V}$  do
11:    Observe external state  $s_{it}^{\text{ex}}$ 
12:     $s_{it} \leftarrow (s_{it}^{\text{ex}}, \tilde{x}_{it}, \theta_i)$ , where  $\tilde{x}_{it} = (\tilde{x}_{i1t}, \dots, \tilde{x}_{int})^T$  and  $\theta_i$  is  $v_i$ 's parameter.
13:    Sample action  $a_{it}^{\text{ex}} \sim \pi_i^{\text{ex}}(s_{it})$ 
14:    Receive external reward  $R_{it} \leftarrow R_{it} + R_{it}^{\text{ex}}(a_{it}^{\text{ex}})$ 
15:    Update  $Q^{\pi_i}$  under the manner of  $Q$ -learning by calculating the time difference (TD)-error
16:  end for
17: end for

```

---

**Corollary 3.1.** The Nash equilibrium of an envy-free auction is  $(\mathbf{b}_{it}, q_{it})$  is  $(\mathbf{o}_{it}, \operatorname{argmax}_q q d_{it}(q))$ .

The remaining problem is how to predict  $\mathbf{o}_t$ . Although several methods could be applied to this problem (including other RL such as SARSA and A3C ?),  $Q$ -learning was used to predict  $\mathbf{o}_t$  in this paper. As  $\mathbf{o}_{it}$  represented the difference between two  $Q$ s, each  $Q$  was approximated. The state was parameterized using the vector  $s_t$ , which contained input and weight. The  $\epsilon$ -greedy policy with  $Q$ -learning typically supposed that discrete actions. Thus the allocation  $g_{ijt}$  was employed as an action rather than  $\mathbf{b}_{it}$  and  $q_{it}$ .

**Algorithm** The overall algorithm is shown in Algorithm 1.

### 3.3 NEURON AS AN AGENT

Stub

## 4 RELATED WORKS

Existing multi-agent reinforcement learning (MARL) communication methods have relied on a trusted third party (TTP) to distribute incentive to agents, leaving them inapplicable in peer-to-peer environments. R/DIAL (Foerster et al., 2016) is a communication method for deep reinforcement learning, which train the optimal communication among the agent with Q-learning. It focuses on that paradigm of centralized planning. CommNet (Sukhbaatar et al., 2016), which exploits the characteristics of a unit that is agnostic to the topology of other units, employs backpropagation to train multi-agent communication. Instead of reward  $R(a_t)$  of an agent  $i$  for actions at  $t$   $a_t$ , QUICR-learning (Agogino & Tumer, 2006) maximizes counterfactual reward  $R(a_t) - R(a_t - a_{it})$ , the difference in the case of the agent  $i$  takes an action  $a_{it}$  ( $a_t$ ) and not ( $a_t - a_{it}$ ). COMA (Foerster et al., 2017) also maximizes counterfactual rewards in an actor-critic setting. CommNet, QUICR and COMA have a centralized environment for distributing rewards through a TTP. In contrast, NaaA does not rely on a TTP, and hence, each agent calculates its reward.

While inter-agent reward distribution has not been considered in the context of communication, trading agents have been considered in other contexts. Trading agent competition (TACs), competitions for trading agent design, have been held in various locations regarding topics such as smart grids (Ketter et al., 2013), wholesale (Kuate et al., 2013), and supply chains (Pardoe & Stone, 2009), yielding innumerable trading algorithms such as Tesauro’s bidding algorithm (Tesauro & Bredin, 2002) and TacTex’13 (Urieli & Stone, 2014). Since several competitions employed an auction as optimal price determination mechanism (Wellman et al., 2001; Stone et al., 2003), using auctions to determine optimal prices is now a natural approach. Unfortunately, these existing methods cannot be applied to the present situation. First, their agents did not communicate because the typical purpose of a TAC is to create market competition between agents in a zero-sum game. Secondly, the traded goods are not digital goods but instead goods in limited supply, such as power and tariffs. Hence, this is the first paper to introduce inter-agent reward distribution to MARL communications.

Auction theory is discussed in terms of mechanism design (Myerson, 1983), also known as inverse game theory. Second-price auctions (Vickrey, 1961) are auctions including a single product and several buyers. In this paper, a digital goods auction (Guruswami et al., 2005) was used as an auction with an infinite supply. Several methods extend digital goods auction to address collusion, including the consensus estimate (Goldberg & Hartline, 2003) and random sample auction (Goldberg et al., 2006), which can be used to improve our method.

This paper is also related to DropConnect in terms of controlling connections between units. Adaptive DropConnect (ADC), proposed in a later section of this paper as a further application, extends the DropConnect (Wan et al., 2013) regularization technique. The finalized idea of ADC (which uses a skew probability correlated to the absolute value of weights rather than dropping each connection between units by a constant probability) is closer to Adaptive DropOut (Ba & Frey, 2013), although their derivation differs. The adjective “adaptive” is added with respect to the method. Neural network optimizing using RL was investigated by Andrychowicz et al. (2016); however, their methods used a recurrent neural network (RNN) and are therefore difficult to implement, whereas the proposed method is RNN-free and forms as a layer. For these reasons, its implementation is simple and fast and it also has a wide area of applicability.

## 5 EXPERIMENT

The proposed reward distribution method was confirmed to work as expected by a validation experiment using the multi-agent setting in ViZDoom (Kempka et al., 2016), an emulator of Doom containing a map editor where additional agents complement the main player. A main player in the ViZDoom environment aims to seek the enemy in the map and then defeat the enemy.

### 5.1 SETUP

A defend the center (DtC)-based scenario, provided by ViZDoom platform, was used for this experiment. Two players, a main player and a cameraman, were placed in the DtC, where they started in the center of a circular field and then attacked enemies that came from the surrounding wall. Although the main player could attack the enemy with bullets, the cameraman had no way to attack,

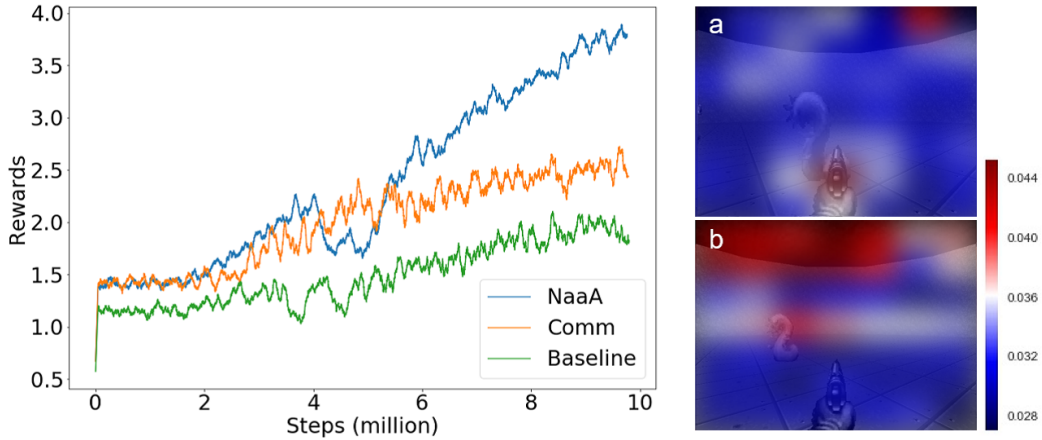


Figure 3: **Left:** Learning curve for ViZDoom multi-agent task. The proposed NaaA-based method outperformed the other two methods (baseline and Comm DQNs). **Right:** Visualizing reward from the main player to the cameraman shows us what is important information for the main player: (a) The pistol. (b) The point at which the enemy appeared and approached.

only scouting for the enemy. The action space for the main player was the combination of { attack, turn left, turn right }, giving a total number of actions  $2^3 = 8$ . The cameraman had two possible actions: { turn left, turn right }. Although the players could change direction, they could not move on the field. Enemies died after receiving one attack (bullet) from the main player, and then player received a score of +1 for each successful attack. The main player received 26 bullets by default at the beginning of each episode. The main player died if they received attacks from the enemy to the extent that their health dropped to 0, and received a score of -1 for each death. The cameraman did not die if attacked by an enemy. Episodes terminated either when the main player died or after 525 steps elapsed.

## 5.2 MODEL

Three models, described below, were compared: the proposed method and two comparison targets.

*Baseline:* DQN without communication. The main player learned standard DQN with the perspective that the player is viewing. Because the cameraman did not learn, this player continued to move randomly.

*Comm:* DQN with communication, inspired by Commnet. The main player learns DQN with two perspectives: theirs and that of the cameraman. The communication vector is learned with a feed-forward neural network.

*NaaA:* The proposed method. The main player learned DQN with two perspectives: theirs and that of the cameraman. Transmissions of rewards and communications were performed using the proposed method.

## 5.3 RESULTS

Training was performed over the course of 10 million steps. Figure 3 Left demonstrates the proposed NaaA model outperformed the other two methods. Improvement was achieved by Adaptive DropConnect. It was confirmed that the cameraman observed the enemy through an episode, which could be interpreted as the cameraman reporting enemy positions. In addition to seeing the enemy, the cameraman observed the area behind the main player several times. This enabled the cameraman to observe enemy attacks while taking a better relative position.

To further interpret this result, a heatmap visualization of revenue earned by the agent is presented in Figure 3 Right. The background picture is a screen from Doom, recorded at the moment when



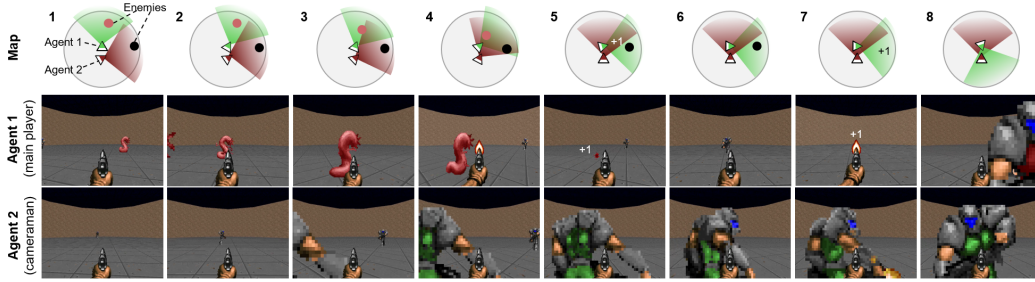


Figure 4: NaaA leads agents to enter a cooperative relationship. First, the two agents face different directions, and the cameraman sells their information to the main player (1). The main player (information buyer) starts to turn right to find the enemy. The cameraman (information seller) starts to turn left to seek new information by finding the blind area of the main player (2 and 3). After turning, the main player attacks the first, having already identified enemy (4 and 5). Once the main player finds the enemy, he attacks and obtains the reward (6 and 7). Both agents then return to watching the dead area of the other until the next enemy appears (8).

the CNN filter was most activated. Figure 4 shows an example of learnt sequence of actions by our method.

## 6 FURTHER APPLICATION

Actually, NaaA is useful not only for multi-agent RL, but also for training of the network. Typical training algorithms of a neural network such as those of RMSProp (Tieleman & Hinton, 2012) and Adam (Kingma & Ba, 2014) are based on a sequential algorithm such as stochastic gradient descent (SGD). Therefore, the problem can be interpreted as a problem to update the state (i.e., weight) to the goal, which is minimization of the expected likelihood.

### 6.1 ADAPTIVE DROPCONNECT

The learning can be accelerated by application of NaaA to the optimizer. We designate the application of NaaA to SGD as *Adaptive DropConnect* (ADC), which is eventually a combination of DropConnect (Wan et al., 2013) and Adaptive DropOut (Ba & Frey, 2013). We introduce ADC herein as one application of NaaA.

ADC uses NaaA for supervised optimization problem with several revisions. First, an environment has an input state such as an image. The agent is expected to update its parameters to maximize its reward obtained from the criterion calculator. The criterion calculator gives batch-likelihood as the reward to the agent. The agent is a classifier which updates its weights to maximize the reward from the criterion calculator. The weights are recorded as an internal state. As a counterfactual return  $o_{ijt}$ , we used a heuristic that uses the absolute value of weight  $|w_{ijt}|$ , which is the same technique as that used by Adaptive DropOut. We use the absolute value of weights because it is the update amount for which the magnitude of error of the output of units is proportional to  $|w_{ijt}|$ .

The algorithm is presented as Algorithm 2. Because the algorithm is quite simple, its implementation can be performed easily. For that reason, it can be widely applied for most general deep learning problems such as image recognition, sound recognition, and even for deep reinforcement learning.

### 6.2 EXPERIMENT

#### 6.2.1 SETUP

In this experiment, we confirm two tasks of classification and single-agent reinforcement learning.

For classification task, we used three types of datasets, MNIST, CIFAR-10 and STL-10. The task given here is to predict the label for each image. The number of class is 10 in those three datasets.

**Algorithm 2** Adaptive DropConnect

---

```

1: for  $t = 1$  to  $T$  do
2:   Compute a bidding price for every edge: for  $(v_j, v_i) \in \mathcal{E}$  do  $b_{ijt} \leftarrow |w_{ijt}|$ 
3:   Compute an asking price for every node: for  $v_i \in \mathcal{V}$  do  $\hat{q}_{it} \leftarrow \underset{q \in [0, \infty)}{\operatorname{argmax}} q d_{it}(q)$ .

4:   for  $(v_i, v_j) \in \mathcal{E}$  do
5:     Compute allocation:  $g_{jit} \leftarrow H(b_{ijt} - \hat{q}_{it})$ 
6:   end for
7:   Sample a switching matrix  $U_t$  from a Bernoulli distribution:  $U_t \sim \text{Bernoulli}(\varepsilon)$ 
8:   Sample the random mask  $M_t$  from a Bernoulli distribution:  $M_t \sim \text{Bernoulli}(1/2)$ 
9:   Generate the adaptive mask:  $M'_t \leftarrow U_t \circ M_t + (1 - U_t) \circ G_{ijt}$ 
10:  Compute  $\mathbf{h}_t$  for making a shipment:  $\mathbf{h}_t \leftarrow (M'_t \circ W_t)\mathbf{x}_t + \mathbf{b}_t$ 
11:  Update  $W_t$  and  $\mathbf{b}_t$  by backpropagation.
12: end for

```

---

The first dataset, MNIST, is a collection of black and white images of handwritten digits whose size is 28x28. The training set and test set are composed of 60,000 examples and 10,000 examples respectively. The images in CIFAR-10 dataset are colored and the size of each image is 32x32. The task is to predict what is shown in each picture. This dataset contains 6,000 images per class (5,000 for training and 1,000 for test). STL-10 is a dataset for image recognition, the number of which is 1,300 for each class (500 for training and 800 for test). The size of each image is 96x96. In this experiment, however, images were resized into 48x48, since the resolution is large compared to the datasets shown above and this dataset requires far more time and resource to compute.

Next, we set the single-agent reinforcement learning task. We used the CartPole task from OpenAI gym with visual input. In this setting, the agent must balance a pole while moving a cart. There is much non-useful information related to the image. For that reason, pruning the pixels is important.

### 6.2.2 MODEL

In this experiment, we compared two models, DropConnect and Adaptive DropConnect (proposed model in this paper). The baseline model is composed of two convolutional layers and two fully connected layers whose outputs are dropped out (we set the possibility as 0.5). The labels of input data are predicted using log-softmaxed value of last fully connected layer. In DropConnect model and Adaptive DropConnect model, first fully connected layer is replaced by DropConnected layer and Adaptive DropConnected layer respectively. Note that DropConnect model corresponds to the our method with  $\varepsilon = 1.0$  and this means agents do not perform their auctions, and randomly mask the weights.

### 6.2.3 RESULTS

For the MNIST datasets, the models are trained for 10 epochs and then evaluated with the test data. The numbers of epochs for CIFAR-10 and STL-10 are 20 and 40 respectively. Experiments are repeated 20 times for each condition, and the average and standard deviation of error rate was calculated. The results is shown in Table 1. As expected, with the model using Adaptive DropConnect, the classification error rate was lower than both the baseline and DropConnect regardless of the datasets given in this experiment.

Table 1: Experimental result for image classification tasks and single-agent RL

	MNIST	CIFAR-10	STL-10	CartPole
DropConnect (Wan et al., 2013)	$1.72 \pm 0.160$	$43.14 \pm 1.335$	$50.92 \pm 1.322$	$285 \pm 21.5$
Adaptive DropConnect	<b><math>1.36 \pm 0.132</math></b>	<b><math>39.84 \pm 1.035</math></b>	<b><math>42.17 \pm 2.329</math></b>	<b><math>347 \pm 29.4</math></b>

## 7 CONCLUDING REMARKS AND FUTURE WORK

This paper proposed a NaaA model to address communication in MARL without a TTP based on two key ideas: inter-agent reward distribution and auction theory. Existing MARL communication methods have assumed the existence of a TTP, and hence could not be applied in peer-to-peer environments. The inter-agent reward distribution, making agents redistribute the rewards they received from the internal/external environment, was reviewed first. When an envy-free auction was introduced using auction theory, it was shown that agents would evaluate the counterfactual returns of other agents. The experimental results demonstrated that NaaA outperformed a baseline method and a CommNet-based method.

Furthermore, a  $Q$ -learning based algorithm, termed Adaptive DropConnect, was proposed to dynamically optimize neural network topology with counterfactual return evaluation as a further application. To evaluate this application, experiments were performed based on a single-agent platform, demonstrating that the proposed method produced improved experimental results relative to existing methods.

Future research may also be directed toward considering the connection between NaaA and neuroscience or neuroevolution. Edelman propounded the concept of neural Darwinism (Edelman, 1987), in which group selection occurs in the brain. Inter-agent rewards, which were assumed in this paper, correspond to NTFs and could be used as a fitness function in genetic algorithms for neuroevolution such as hyperparameter tuning.

As NaaA can be applied in peer-to-peer environments, the implementation of NaaA in blockchain (Swan, 2015) is under consideration. This implementation would extend the areas where deep reinforcement learning could be applied. Bitcoin (Nakamoto, 2008) could be used for inter-agent reward distribution, and the auction mechanism could be implemented by smart contracts (Buterin et al., 2014). Using the NaaA incentive design, it is hoped that the world may be united, allowing people to share their own representations on a global scale.

## REFERENCES

- A. K. Agogino and K. Tumer. QUICR-learning for multi-agent coordination. AAAI’06, 2006.
- Marcin Andrychowicz, Misha Denil, Sergio Gomez, Matthew W Hoffman, David Pfau, Tom Schaul, and Nando de Freitas. Learning to learn by gradient descent by gradient descent. In *Advances in Neural Information Processing Systems*, pp. 3981–3989, 2016.
- Jimmy Ba and Brendan Frey. Adaptive dropout for training deep neural networks. In *Advances in Neural Information Processing Systems*, pp. 3084–3092, 2013.
- Vitalik Buterin et al. A next-generation smart contract and decentralized application platform. *white paper*, 2014.
- G. M Edelman. *Neural Darwinism: The theory of neuronal group selection*. Basic books, 1987.
- J. Foerster, G. Farquhar, T. Afouras, N. Nardelli, and S. Whiteson. Counterfactual multi-agent policy gradients. *arXiv:1705.08926*, 2017.
- Jakob Foerster, Yannis Assael, Nando de Freitas, and Shimon Whiteson. Learning to communicate with deep multi-agent reinforcement learning. In *Advances in Neural Information Processing Systems*, pp. 2137–2145, 2016.
- Andrew V Goldberg and Jason D Hartline. Competitiveness via consensus. In *Proceedings of the fourteenth annual ACM-SIAM symposium on Discrete algorithms*, pp. 215–222. Society for Industrial and Applied Mathematics, 2003.
- Andrew V Goldberg, Jason D Hartline, Anna R Karlin, Michael Saks, and Andrew Wright. Competitive auctions. *Games and Economic Behavior*, 55(2):242–269, 2006.
- Venkatesan Guruswami, Jason D Hartline, Anna R Karlin, David Kempe, Claire Kenyon, and Frank McSherry. On profit-maximizing envy-free pricing. In *ACM-SIAM symposium on Discrete algorithms*, 2005.

- Michał Kempka, Marek Wydmuch, Grzegorz Runc, Jakub Toczek, and Wojciech Jaśkowski. Viz-doom: A doom-based ai research platform for visual reinforcement learning. In *Computational Intelligence and Games (CIG), 2016 IEEE Conference on*, pp. 1–8. IEEE, 2016.
- Wolfgang Ketter, John Collins, and Prashant Reddy. Power tac: A competitive economic simulation of the smart grid. *Energy Economics*, 39:262–270, 2013.
- Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.
- Rodrigue Talla Kuate, Minghua He, Maria Chli, and Hai H Wang. An intelligent broker agent for energy trading: An mdp approach. In *IJCAI*, pp. 234–240, 2013.
- William Forster Lloyd. *Two lectures on the checks to population*. 1833.
- 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.
- Roger B Myerson. Mechanism design by an informed principal. *Econometrica: Journal of the Econometric Society*, pp. 1767–1797, 1983.
- Satoshi Nakamoto. Bitcoin: A peer-to-peer electronic cash system, 2008.
- David Pardoe and Peter Stone. An autonomous agent for supply chain management. *Handbooks in Information Systems Series: Business Computing*, 3:141–172, 2009.
- Tuomas Sandholm and XiaoFeng Wang. (im) possibility of safe exchange mechanism design. In *Eighteenth national conference on Artificial intelligence*, pp. 338–344. American Association for Artificial Intelligence, 2002.
- D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. Van Den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, et al. Mastering the game of go with deep neural networks and tree search. *Nature*, 529(7587):484–489, 2016.
- Peter Stone, Robert E Schapire, Michael L Littman, János A Csirik, and David McAllester. Decision-theoretic bidding based on learned density models in simultaneous, interacting auctions. *Journal of Artificial Intelligence Research*, 19:209–242, 2003.
- S. Sukhbaatar, R. Fergus, et al. Learning multiagent communication with backpropagation. In *NIPS’16*, 2016.
- Melanie Swan. *Blockchain: Blueprint for a new economy*. ” O’Reilly Media, Inc.”, 2015.
- Gerald Tesauro. Temporal difference learning and td-gammon. *Communications of the ACM*, 38(3): 58–68, 1995.
- Gerald Tesauro and Jonathan L Bredin. Strategic sequential bidding in auctions using dynamic programming. In *Proceedings of the first international joint conference on Autonomous agents and multiagent systems: part 2*, pp. 591–598. ACM, 2002.
- Tijmen Tieleman and Geoffrey Hinton. Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. *COURSERA: Neural networks for machine learning*, 4(2):26–31, 2012.
- Daniel Urieli and Peter Stone. Tactex’13: a champion adaptive power trading agent. In *Proceedings of the 2014 international conference on Autonomous agents and multi-agent systems*, pp. 1447–1448. International Foundation for Autonomous Agents and Multiagent Systems, 2014.
- William Vickrey. Counterspeculation, auctions, and competitive sealed tenders. *The Journal of finance*, 16(1):8–37, 1961.
- Li Wan, Matthew Zeiler, Sixin Zhang, Yann L Cun, and Rob Fergus. Regularization of neural networks using dropconnect. In *Proceedings of the 30th international conference on machine learning (ICML-13)*, pp. 1058–1066, 2013.

Michael P. Wellman, Peter R. Wurman, Kevin O'Malley, Roshan Bangera, Daniel Reeves, William E Walsh, et al. Designing the market game for a trading agent competition. *IEEE Internet Computing*, 5(2):43–51, 2001.

Shih-Hung Wu and Von-Wun Soo. Game theoretic reasoning in multi-agent coordination by negotiation with a trusted third party. In *Proceedings of the third annual conference on Autonomous Agents*, pp. 56–61. ACM, 1999.

## APPENDIX

### A.1 PROOF OF THEOREM 3.1

The optimization problem in Eq:10 is made of two terms except of the constant, and the only second term is depends on  $\mathbf{b}$ . Hence, we consider to optimize the second term. The optimal bidding prices  $\hat{\mathbf{q}}_t$  is given by the following equation.

$$\begin{aligned}\hat{\mathbf{b}}_{it} &= \underset{\mathbf{b}}{\operatorname{argmin}} \mathbb{E}_{\mathbf{q}_t} [\mathbf{g}_{it}(\mathbf{b})^T (\mathbf{q}_t - \gamma \mathbf{o}_{it})] = \underset{\mathbf{b}}{\operatorname{argmin}} \mathbb{E}_{\mathbf{q}_t} [H(\mathbf{b} - \mathbf{q}_t)^T (\mathbf{q}_t - \gamma \mathbf{o}_{it})] \\ &= \underset{\mathbf{b}}{\operatorname{argmin}} \mathbb{E}_{\mathbf{q}_t} \left[ \sum_{j=1}^N H(b_j - q_{jt})(q_{jt} - \gamma o_{ijt}) \right] = \underset{\mathbf{b}}{\operatorname{argmin}} \sum_{j=1}^N \mathbb{E}_{q_{jt}} [H(b_j - q_{jt})(q_{jt} - \gamma o_{ijt})],\end{aligned}\quad (11)$$

From independence, the equation is solved if we solve the following problem.

$$\hat{b}_{ijt} = \underset{b}{\operatorname{argmin}} \mathbb{E}_{q_{jt}} [H(b - q_{jt})(q_{jt} - \gamma o_{ijt})], \quad \forall j \in \{1, \dots, N\} \quad (12)$$

Hence,  $\hat{b}_{ijt}$  can be derived as the solution which satisfies the following equation.

$$\left. \frac{\partial}{\partial b} \mathbb{E}_{q_{jt}} [H(b - q_{jt})(q_{jt} - \gamma o_{ijt})] \right|_{b=\hat{b}_{ijt}} = 0, \quad \left. \frac{\partial^2}{\partial b^2} \mathbb{E}_{q_{jt}} [H(b - q_{jt})(q_{jt} - \gamma o_{ijt})] \right|_{b=\hat{b}_{ijt}} > 0$$

For simplicity, we let  $q = q_{jt}$  and  $o = o_{ijt,t+1}$ . Then, the following equation holds.

$$\begin{aligned}\frac{\partial}{\partial b} \mathbb{E}_q [H(b - q)(q - \gamma o)] &= \frac{\partial}{\partial b} \int_0^\infty H(b - q)(q - \gamma o)p(q) dq \\ &= \frac{\partial}{\partial b} \int_0^b (q - \gamma o)p(q) dq \\ &= (b - \gamma o)p(q = b),\end{aligned}\quad (13)$$

$$\frac{\partial^2}{\partial b^2} \mathbb{E}_q [H(b - q)(q - \gamma o)] = p(q = b) + (b - \gamma o) \frac{\partial}{\partial b} p(q = b) \quad (14)$$

Then, condition  $\left. \frac{\partial}{\partial b} \mathbb{E}_{q_{jt}} [H(b - q_{jt})(q_{jt} - \gamma o_{ijt})] \right|_{b=\hat{b}_{ijt}} = 0$  is satisfied only by  $\hat{b}_{ijt} = \gamma o_{ijt}$ . We substitute  $\hat{b}_{ijt}$  into Eq:14,

$$\left. \frac{\partial^2}{\partial b^2} \mathbb{E}_{q_{jt}} [H(b - q_{jt})(q_{jt} - \gamma o_{ijt})] \right|_{b=\hat{b}_{ijt}} = p(q = \hat{b}_{ijt}) + 0 > 0 \quad (15)$$

Therefore,  $\hat{b}_{ijt} = \gamma o_{ijt}$  is a unique solution as the minimum point. From generality,  $\hat{\mathbf{b}}_{it} = \gamma \mathbf{o}_{it}$  holds.