**Reinforcement Learning - Supply Chain Ordering Management: An application to the beer game**

**Set of possible states (inventory positions):**

State space S(t) = {S1(t), S2(t), S3(t), S4(t)} with Si(t) as inventory position of agent i in time step t

With actual inventory position space

Si(t) = {]−∞,−6[ ; [−6,−3[ ; [−3,0[ ; [0,3[ ; [3,6[ ; [6,10[ ; [10,15[ ; [15,20[ ; [20,∞[}

Coded inventory position space

Si(t)= {1, 2, 3, 4, 5, 6, 7, 8, 9}

**Start states Si(t):** actual Si(1) = {12, 12, 12, 12}; coded Si(1) = {7, 7, 7, 7}

**Set of possible actions (ordering sizes):** action space A = {0, 1, 2, 3} equal to y value in the X + Y rule

**Initial transition probabilities:** Prexploration = 0.98; Prexploitation = 0.02

**Reward the agent receives from the environment:** Loss function of whole supply chain costs

R = sum of all agents (**alpha** \* inventory costs **hi(t)** + **beta** \* backorder costs **Ci(t)**)

Ein Bild, das Schrift, weiß, Text, Typografie enthält.

Automatisch generierte Beschreibung

**Introduction**

**Goal:** determine the ordering size Oij(t) of each actor of the supply chain to the upstream actor aiming to minimize inventory costs of the whole supply chain (find optimal policy)

Ein Bild, das Schrift, Text, Typografie, weiß enthält.

Automatisch generierte Beschreibung

**Supply Chain:** linear with four echelons/agents (retailer, distributor, manufacturer and supplier)

Ein Bild, das Text, Reihe, Diagramm, Schrift enthält.

Automatisch generierte Beschreibung

**Environmental uncertainties:**

* customer demand CD(t) = [0, 15]
* lead times LTij(t) = [0, 4] from level i to level j

and LT10(t) = 0 (retailer to customer)

-> a fixed ordering rule cannot achieve the system’s goal!

-> how to react to a changing environment?

**Idea of RL:** Agent selects an action, and the environment responds to it and presents a new situation.

**Markov Decision Process (MDP):** A state signal that succeeds in holding all relevant decision-making information is said to be Markov, or to have Markov property. If the state signal has Markov property, then the environment’s response at time step t+1 depends only on the state and action representations at time step t.

**Trad-off between exploration and exploitation:** the agent in each time step has two choices: to explore or to exploit his knowledge. The agent mostly explores at first because of lack of knowledge about environment. In the next steps he improves his knowledge (learning the value of Q-functions for each state-action pair) and the probability of exploration decreases as the probability of exploitation increases.



**In every level at each time step, four events happen:**

1. Previous orders are received (according to the lead-times) from the upstream actor.
2. Order size is received from the downstream actor.
3. The received order is fulfilled from on-hand inventory (if possible).
4. Actor decides about placing order for stock replenishment.

**4. Problem description and modelling**

**Parameters:**

* time horizon: n = 35 weeks
* inventory holding costs: alpha = 1 for each actor/unit/period
* penalty/backorder costs: beta = 2 for each actor/unit/period
* discount factor: gamma = 1