Doctoral Dissertation

Research on Decision-making Methods under Uncertain Information

- Assortment Optimization for Vending Machines -

Gaku Nemoto

Supervisor Kunihiko Hiraishi

Graduate School of Advanced Science and Technology
Japan Advanced Institute of Science and Technology
[Information Science]

March, 2025

Abstract

This dissertation addresses the optimization of product assortment in vending machines under conditions of uncertain information. In the retail industry, decision-making processes often require handling incomplete or noisy data, making it essential to develop methods that account for such uncertainties. This research formulates the vending machine assortment problem using Partially Observable Markov Decision Processes (POMDPs), enabling dynamic decision-making under limited observations.

The proposed methodology integrates a product selection model that captures consumer purchasing behavior and a POMDP-based optimization framework to improve vending machine operations. The study provides a comprehensive framework for modeling the state transitions, observation functions, and reward structures involved in assortment optimization. It also introduces practical strategies for assortment exchange that agents can implement in real-world scenarios. Numerical simulations are conducted to evaluate the performance of the proposed approach, demonstrating its effectiveness in maximizing expected rewards and improving vending machine operations.

The key contributions of this research are as follows: (1) the formulation of the assortment optimization problem as a decision-making process under uncertainty, (2) the development of a novel method for solving this problem in vending machine settings, and (3) an exploration of its applicability to other business sectors with similar decision-making challenges." The findings suggest that the proposed method offers a solution to the assortment optimization problem and provides valuable insights for improving decision-making processes in uncertain environments.

Keywords: Decision-making under uncertain information, Assortment optimization problem, Vending machine, Partially Observable Markov Decision Processes, Product selection model, Numerical simulation

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background

Decision-making in industrial practice is often based on judgments made under conditions of uncertain information.

For example, Traffic Alert and Collision Avoidance system (TCAS), it is an onboard collison avoidance system for airclafts [1]. TCAS provides pilots with ascent or descent instructions to avoid collisions with other aircraft through both audio and visual alerts. It receives replies from other aircraft via radio and estimates distance and bearing by measuring the delay of those replies. The system determines the optimal advice based on observed range, bearing, and altitude, while accounting for sensor imperfections and uncertainties in the future trajectories of the aircraft. It is designed to ensure high safety while not disrupting normal air traffic procedures.

Another example is attribute-based person search in surveillance video [1]. Attribute-based search refers to identifying a person based on noticeable features such as clothing, hair color, and carried items, without relying on facial recognition or biometric data. This method has the advantage of being able to search for individuals even without existing biometric data. However, accurate search becomes challenging due to variations in appearance caused by factors like clothing, lighting, and pose, which can result in different visual representations of the

same attributes. A probabilistic modeling approach is effective in handling such uncertainty. This approach probabilistically models the relationship between the attribute profile and the surveillance video, taking into account hidden factors like body orientation and clothing position. This allows for more accurate identification of individuals under varying conditions. Experimental results show that probabilistic models perform better than traditional methods, demonstrating their effectiveness in interactive searches at locations like airports. However, the accuracy of the search depends on factors such as scene resolution, lighting, and crowd density. To address these challenges, it is necessary to select appropriate video analysis targets and re-train the dataset. Additionally, this approach can be extended to search for other scene components, such as vehicles or luggage.

1.2 Aim and Contribution of this study

This study focuses on the decision-making process under uncertain information, and in particular deals with the optimization problem of product assortment in the retail industry. The product assortment problem is the problem of selecting products to place in a limited product display space (shelf), and a solution is sought that provides an assortment that maximizes or minimizes a specific index value.

In particular, this study addresses the problem of product assortment optimization under conditions where only uncertain information is available and where there are special constraints, such as limitations on the available products and actions. A specific example is the beverage vending machine (hereafter, vending machine). The product assortment optimization problem in vending machines is more complex than in typical retail stores due to unique constraints such as limited available information, a limited number of shelves, restricted stock levels for each product, and limitations on the frequency of assortment changes.

This problem requires the implementation of an appropriate assortment according to each situation based on information such as past sales data, product features, inventory status, consumer preferences, and changes in the environment.

At the same time, it is necessary to consider the next assortment based on the information obtained from the implemented assortment.

In this study, the following contributions will be made to address the assortment optimization problem under uncertain environments.

(1) Formulate the product assortment optimization problem as a decision-making process under uncertainty.

(2) Propose a unique optimization method for the product assortment optimization problem in the context of vending machines under uncertain information. Conduct numerical simulations to verify the effectiveness of the proposed method.

(3) Consider the possibility of a unified approach for product assortment problems in other business sectors by taking into account observable states and constraints, and provide a discussion on this.

1.3 Vending Machine Business in Japan

Here, we review the overview of the vending machine business, particularly the current situation and challenges in Japan.

1.3.1 Overview

The total number of vending machines and automated service machines (hereafter simply referred to as vending machines) in Japan is approximately 3.9 million. Of these, beverage vending machines make up the largest proportion, at approximately 2.20 million, or 56.4% of the total. By beverage vending machine type, soft drinks are the most common, accounting for approximately 89.1% of the total in terms of number, followed by cup-type machines selling coffee, cocoa, etc. (5.6%), milk beverage (4.4%), and alcoholic beverages and beer (0.9%) ([2], Fig.1.1).

1.3.2 Stakeholders

Nunokawa et al. [3] have summarized the current state of the vending machine business in Japan as follows:

The beverage vending machine business involves stakeholders on both the supply and demand sides.

On the supply side, there are vending machine manufacturers, beverage manufacturers, and operators who manage and restock the vending machines. On the demand side, there are location owners who provide the space for the vending machines and consumers who purchase the beverages from the machines. The vending machine business is mainly driven by the collaboration of vending machine manufacturers, beverage manufacturers, and operators.

Operators are generally divided into two types: specialized operators and diversified operators. Specialized operators are companies that focus solely on vending machine operations and typically handle products from multiple beverage manufacturers. On the other hand, diversified operators are companies that manage vending machines alongside other business activities. These operators are often beverage manufacturers themselves, running vending machines to promote their own products.

The practical work of vending machine operations is carried out by agents, also known as ”route men.” (Although Nunokawa and others refer to them as ”route men”, this study will refer to these workers as ”agents” for consistency.)

An agent works for an operator company and is responsible for restocking and maintaining the vending machines [4]. Specifically, agents load beverages into trucks from warehouses, then travel to different locations to refill and perform maintenance checks on the vending machines. After completing these tasks, they return to the depot to dispose of any waste and handle products. Since each agent is typically assigned to specific vending machines in a given area, the same vending machine is generally handled by one agent.

The workload of a agent varies depending on the number of vending machines

they are responsible for. On average, a agent manages 5 to 10 machines per day. During a single round of work, they may visit multiple vending machines, and if they work five days a week, they are responsible for about 25 to 50 machines. The tasks are often adjusted according to the agent's judgment, and there is no fixed route they follow every week. Additionally, the amount of work and frequency of visits can vary based on the location and number of machines, requiring a flexible approach.

1.3.3 Visit and Replenishment Operations

In conducting this study, we investigated how operators and agents in Japan manage the workload – visit and replenishment operations of vending machines.

Agents are typically responsible for managing several dozen to approximately 100 vending machines individually, adjusting the visit frequency based on the sales trends of each machine. Machines with high sales volumes are visited more frequently, while those with lower sales volumes are visited less frequently. High-frequency machines are often visited on a near-daily basis, whereas low-frequency machines may only be visited approximately once a month. In some cases, operators consider sales volume solely in terms of the number of units sold, without taking product prices into account.

Visit frequency is determined by considering the balance between sales revenue and visit costs, as well as the risk of stockouts. More frequent visits reduce the risk of stockouts but incur higher visit costs relative to revenue. Conversely, less frequent visits improve cost efficiency but increase the likelihood of lost sales opportunities due to stockouts. To optimize operational efficiency, operators tend to instruct agents to adjust visit frequency so that *sales per visit* remain relatively constant across machines. However, this calculation also considers the machine's storage capacity, i.e., the maximum inventory it can hold. Additionally, during each visit, the most common practice is to replenish the sold inventory in full. By adopting this approach, agents can stabilize both sales and replenishment quantities

per visit, provided that visit timing is appropriately aligned with sales trends.

In recent years, visit support systems that utilize past sales data to predict sales and suggest optimal visit frequencies to agents have been adopted by multiple companies.

1.3.4 Product Assortment Work

Regarding product assortment exchange, agents must first follow general instructions provided by the operator. The instructions from the operator typically involve seasonal considerations. For example, from spring to summer, the assortment is centered around cold beverages, while in the fall and winter, some items may be replaced with hot beverages. Additionally, new products are added, and discontinued products are removed. Moreover, the time and labor that can be devoted to a single visit are limited. For instance, due to physical, temporal, and endurance constraints, it is often infeasible to replace all products at once. In most cases, the exchange is limited to only two or three products.

Following these guidelines, agents have the discretion to adjust the product assortment for each vending machine. However, the results can vary significantly depending on the agent. Experienced agents take into account factors such as the location and sales trends for each product at specific vending machines, as well as the preferences of the location owner, to construct an appropriate assortment. They are able to improve sales by adapting the assortment in response to changes in sales trends.

On the other hand, for less experienced agents, adjusting product assortments is a challenging task. The operator typically provides a standard assortment, which, if followed, will ensure minimum sales. However, it is difficult for less experienced agents to perform flexible assortment changes based on various types of information, as experienced agents do.

1.4 Structure of our Proposal Method

Based on the industry situation described in the previous section, the main objective of this study is to model and propose the product assortment method that can help even less experienced agents achieve a certain level of sales performance.

An overview of the model structure for the proposed method addressing the vending machine product assortment optimization problem, as discussed in Section 1.2, will be provided here (Fig. 1.2). Further details will be explained in later chapters.

(1) A customized POMDP model is applied to represent agent decision-making.

(2) A product selection model is introduced to capture consumer purchasing behavior.

(3) For the vending machine model, a state transition model is applied to the vending machines state over time for simulation.

(4) Time in relation to purchasing behavior and restocking operations is handled flexibly, as purchases do not occur at fixed intervals.

1.5 Organization of this Dissertation

The structure of this dissertation is as follows.

In Chapter 1 (this chapter), we describe the research background and objectives, as well as an overview of the vending machine business in Japan. Chapter 2 discusses the fundamentals and basic methodologies for addressing decision-making under uncertainty.

In Chapter 2, related study is reviewed. It highlights key studies on assortment optimization problems, vending machines, and consumer behavior in product selection. Furthermore, it compiles and organizes representative solutions for POMDPs from existing literature.

Chapter 3 focuses on the general formulation of assortment optimization problems and explains the foundational concepts behind the proposed solutions.

Chapter 5 explains the details of our proposed POMDP-based method for addressing the vending machine assortment problem.

In Chapter 6, the proposed approach is applied to numerical simulations. Results from simulations conducted under various scenarios and conditions are presented to evaluate the performance of the method.

Chapter 7 provides a discussion of the proposed method, reflecting on the outcomes and findings presented in earlier chapters.

Finally, Chapter 8 concludes the work with a summary and an exploration of potential future research directions.

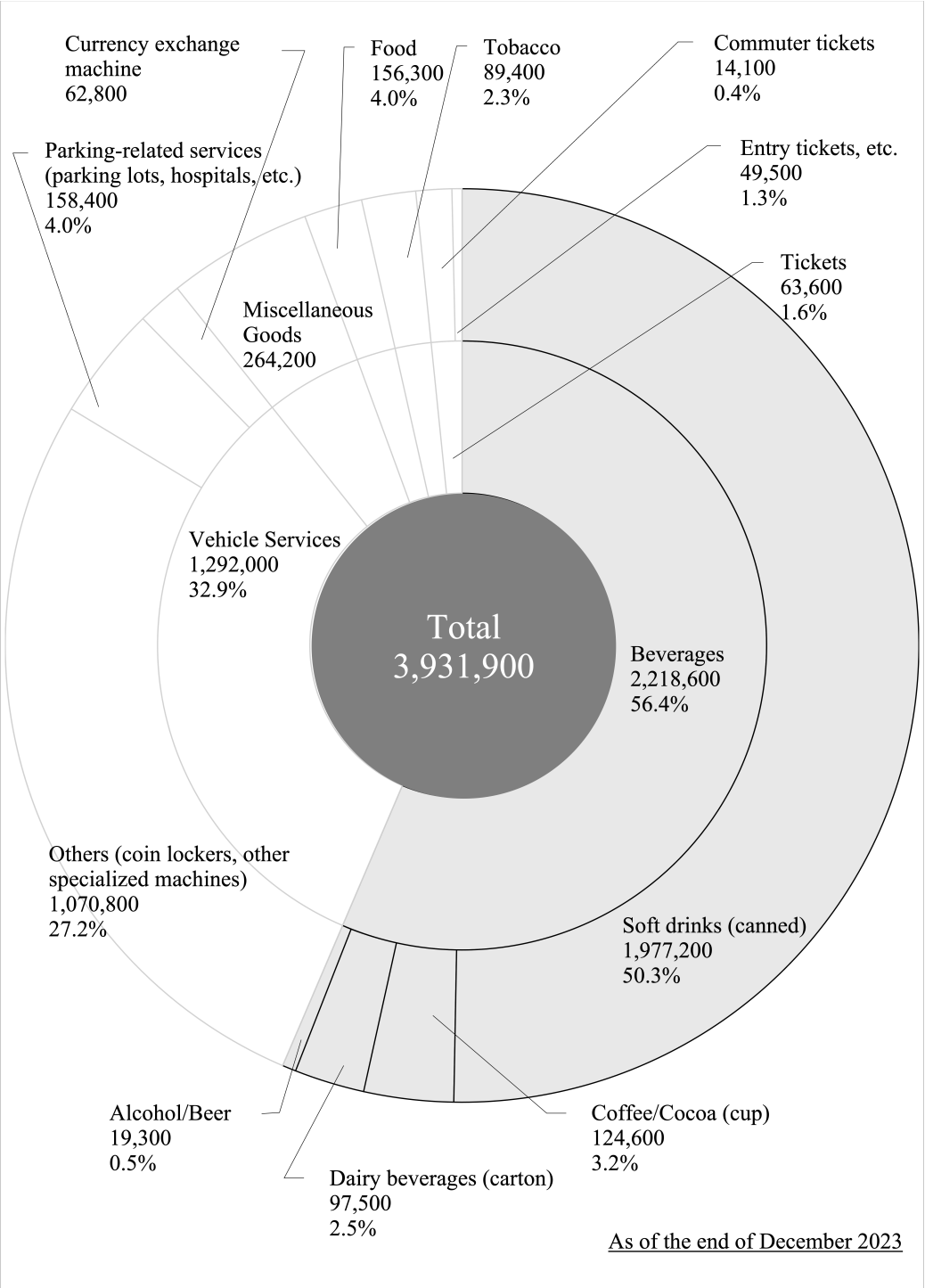

Figure 1.1: Deployment Status of Vending Machines (Compiled based on [2])

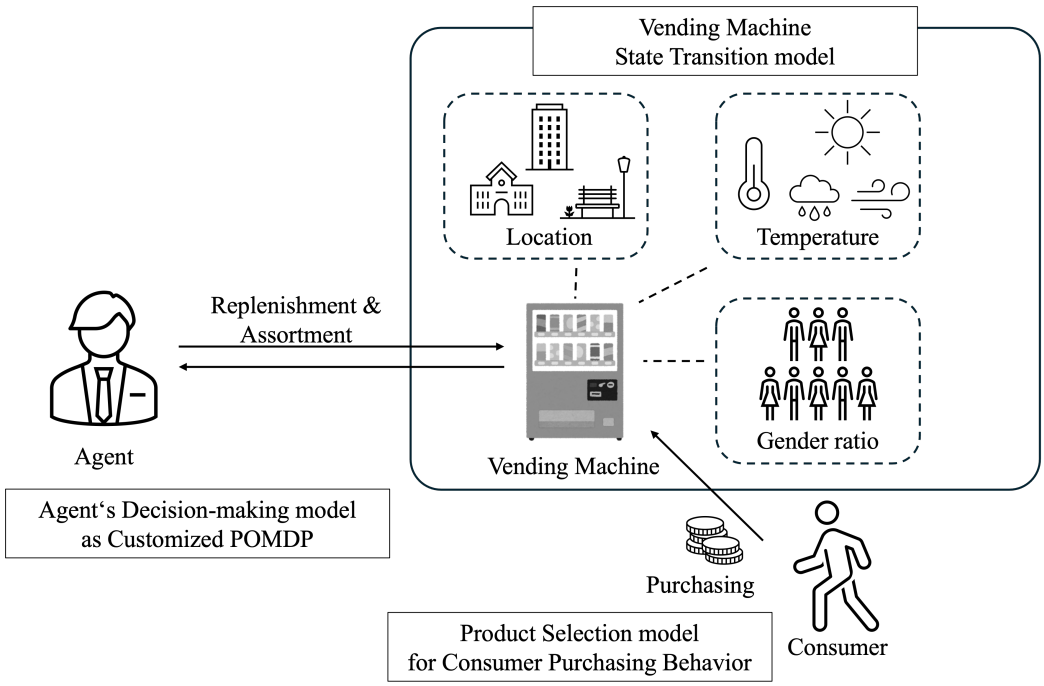

Figure 1.2: Structure of Proposal method

Chapter 2

Decision Making Under Uncertainty

2.1 Decision Making

In considering decision-making, we define the agent and the environment. An agent is an entity that makes decisions and takes actions based on its observations of the environment. Agents can be humans or robots, or they could be software, such as decision-support systems. The environment changes its state according to the agent's actions and other factors, while also providing information to the agent through observations.

Fig. 2.1 represents the interaction between the agent and the environment through actions and observations. At time t, the agent observes the environment and receives an observation $o(t)$. Observations may come from human senses, such as vision or hearing, from electrical signals like radar or sensors, or from numerical data, such as that from a point-of-sale (POS) system.

Observations are often incomplete or contain noise. Additionally, the observed values may change probabilistically. Through the decision-making process, which will be explained later, the agent selects an action $a(t)$. This action can potentially affect the environment. Our goal is to develop intelligent agents that interact with the environment over time to achieve specific objectives. Given a sequence of past observations $o(0), \ldots, o(t)$, the agent must consider its knowledge of the environment and select the optimal action $a(t)$ to achieve its goals.

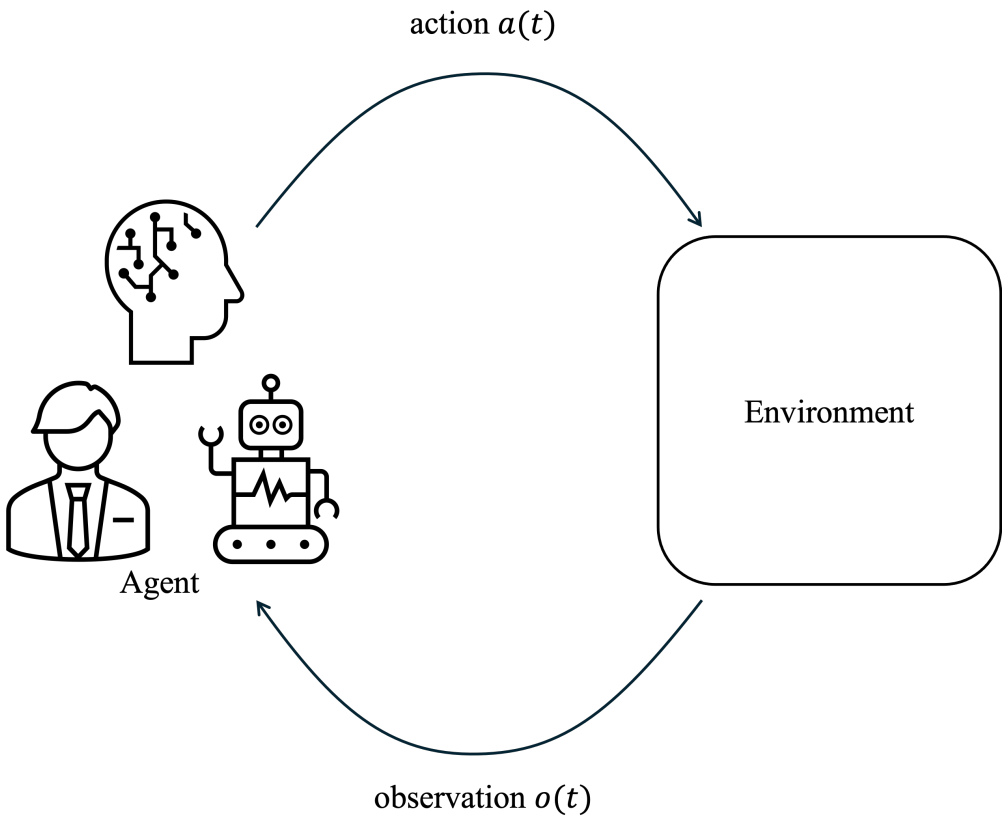

Figure 2.1: Interaction between the agent and the environment

2.2 Methods for Designing Decision Agents

There are various methods for designing decision agents. It is necessary to select an appropriate method depending on the task at hand. This section provides a brief overview of these methods.

2.2.1 Explicit Programming

The most direct method for decision-making is for the designer to anticipate various environments and states and explicitly pre-program the most desirable actions for the agent according to each situation. While this approach can be effective for simple environments or problems, it is challenging to achieve an optimal strategy for all possible states in complex problems.

2.2.2 Supervised Learning

One method for decision-making involves using supervised learning, a type of machine learning algorithm. This approach provides numerous training examples as learning data, where actions are selected based on observations, allowing the algorithm to learn these patterns. It is effective when the designer knows the optimal actions for representative states. However, when faced with new states not present in the training data, it can be difficult to achieve the desired level of performance.

2.2.3 Optimization

In addition, there are optimization methods. For optimization, the designer needs to specify a performance measure for decision-making. The optimization algorithm runs a series of simulations based on a decision strategy within the space of possible strategies and searches for the strategy that maximizes the performance measure. If the space of possible strategies is relatively low-dimensional and the performance measure does not have many local optima, it may be possible to explore appropriate local or global strategies. On the other hand, in complex problems, effectively utilizing knowledge of the dynamic model can often lead to better strategies. This is because, in complex problems, many factors are involved, and simply running simulations may make it difficult to find the optimal strategy.

2.2.4 Planning

Planning is an optimization method that utilizes a dynamic model to guide the search, and much research focuses on deterministic problems. For some problems, approximating the dynamic model with a deterministic model makes it easier to handle high-dimensional issues, but in other cases, accounting for future uncertainty is crucial.

2.2.5 Reinforcement Learning

In planning, it is assumed that the dynamic model is known, but reinforcement learning relaxes this assumption, allowing the agent to learn its decision-making strategy through interaction with the environment. The designer only provides a performance measure, and the optimization of the agent's behavior is left to the learning algorithm. One of the complexities of reinforcement learning is that the choice of action affects not only the achievement of goals but also the agent's ability to learn about the environment and discover features of the problem.

Chapter 3

Related Work

3.1 Assortment Optimization Problem

Assortment optimization problem (AOP) has been widely studied. Here, an overview of some of its aspects is presented.

The existing literature on assortment optimization is centered around two major themes: (1) static and dynamic substitution mechanisms, and (2) consumer behavior models.

Static substitution assumes that if the initially selected product is out of stock, then the consumer will not purchase another item instead [5]. In contrast, dynamic substitution assumes that if the product is out of stock, then the consumer purchases another item as an alternative [6, 7].

In [8], three models are shown for describing consumer behavior: exogenous demand, locational choice and multinomial logit. The exogenous demand model gives a method for describing consumer behavior from observable sales data of each product, such as Kök and Fisher [9]. The locational choice model was developed by Lancaster [10]. This model introduces multi-dimensional vectors where each dimension corresponds to a product characteristic and consumer's demand. The consumer's selection of products is determined by the proximity of the consumer's ideal vector to the Resear's vector. The multinomial logit (MNL) model is a random utility model that represents the selection probability

of each product as functions of consumer's utility. The basic MNL model was established by McFadden [11]. The MNL model has limitations because it assumes independence of irrelevant alternatives in the selection probabilities of products. To reduce these limitations, the nested MNL model was proposed by Williams [12].

3.1.1 Solutions for AOP

AOP involves various methods, each with distinct characteristics. The simplest approach is heuristic methods, such as ABC analysis and the 80/20 rule, which prioritize high-revenue products. These methods allow for quick decision-making with minimal computation, but they lack adaptability to market changes, often leading to suboptimal selections.

A more precise approach is mathematical optimization, which uses linear programming (LP) and integer linear programming (ILP) to determine the optimal assortment while considering constraints such as budget and shelf space. These methods provide high accuracy but can be computationally expensive, making them difficult to implement in large-scale environments.

Machine learning-based methods analyze past sales data to predict future demand. For example, collaborative filtering identifies customer purchasing trends and recommends high-demand products. However, these models require large datasets for accurate predictions and must be updated regularly to maintain effectiveness.

Another approach involves simulation and A/B testing. Monte Carlo simulation tests various demand scenarios virtually to evaluate the effectiveness of different assortments, while A/B testing compares different product lineups using real-world sales data. These methods offer realistic insights but require significant time and cost to implement effectively.

Finally, reinforcement learning adapts dynamically to market changes by learning the best assortment strategies through trial and error. Techniques like Q-learning and Multi-Armed Bandit algorithms refine decision-making over time.

Although reinforcement learning is highly flexible, it demands extensive computational resources and long training periods, making implementation challenging.

Each method has its strengths and weaknesses, and selecting the right one depends on data availability, computational capacity, and market volatility. A combination of multiple techniques can often yield the best results, allowing businesses to optimize their product assortment effectively.

3.2 Vending Machines

Research on vending machines has been conducted in various fields, with this discussion focusing primarily on issues related to the tasks of operators and management of agent's work.

3.2.1 Inventory Distribution Planning

There are studies addressing inventory distribution planning to enable efficient rounds across multiple vending machines. Many of these studies have taken an operations research perspective, with Miyamoto *et al.* [13] being a representative example.

3.2.2 AOP on Vending Machine

Studis on assortment planning for vending machines are progressing.

A column refers to a storage location within a vending machine, and through optimal allocation and inventory management of products within these limited columns, increased sales and reduced stockouts can be achieved. Miyamoto *et al.* [14] and Ito *et al.* [15] formulated the column allocation problem in vending machines as a combinatorial optimization problem, aiming to minimize total costs and reduce stockouts through integer programming. In contrast, Takeuchi *et al.* [16] modeled consumer purchase demand as following a Poisson process, differing from Miyamoto *et al.* by adopting long-term profitability as the objective

function. Anupindi *et al.* [17] proposed a demand estimation method that models consumer behavior in choosing substitute products when a stockout occurs.

Grzybowska *et al.* [18] proposed a model to optimize product allocation in vending machines under fixed restocking constraints. Using genetic algorithms, they evaluated revenue through simulation. A case study based on real-world data showed a 3.4% improvement in overall net revenue, with a maximum increase of 6% for highly popular vending machines.

Watanabe *et al.* [19] proposed a system for reducing the costs associated with product supply and release in next-generation vending machines. They improves the placement algorithm by utilizing a parameter-free genetic algorithm, aiming to enhance efficiency and reduce operational costs in vending machine systems.

The reasons why the assortment optimization problem for vending machines has not been well discovered are considered to be (i) demand for solving this problem was small because the assortment is usually decided by agents using their knowledge and experience on sales, and (ii) complexity of the problem. The complexity arises from the following notable characteristics of the problem:

- Sales of products can be observed only when the replenishment is done.
- The replenishment work is done on a regular basis. Therefore, solution to the problem is a decision making process based on past history of observations.
- Nature of customers is not observable and needs to be estimated.

Recently, beverage companies try to introduce information systems that support route men's work. Proposing a method that helps the route men's decision making is the main contribution of this study.

3.3 Consumer Product Selection Behavior

A representative recent study in the field of marketing on consumer product selection behavior when faced with a wide variety of products is the work of Sato

et al. [20] [21] [22], which formalized consumer product selection behavior using Bayesian modeling based on retail store sales performance data (point of sales: POS data). Fujita *et al.* [23] also formalized consumer utility for products in a logit model for restaurants, and attempted to forecast demand using parameters estimated from POS data using statistical methods. Another study is by Matsumura *et al.* [24], which used a multi-agent model to simulate consumer behavior when products are out of stock in convenience stores.

Both studies involve building statistical models based on sales data from retailers and restaurants and verifying their validity, but they only aim to predict consumer choice behavior for a fixed product lineup (shelf allocation in a retail store). There are still only a few studies that go as far as improving product lineups to increase store sales.

Vending machines differ from retail stores in that sales results are not known at the time of sale (data can only be obtained during the round visit), and because each machine is installed in a different environment, it is difficult to set uniform parameters. In addition, the route man himself is required to make appropriate judgments regarding the timing of product changes and the selection of target products. For this reason, a different approach from previous research is needed to model product selection behavior from vending machines.

3.3.1 Product Selection Model

In order to give the gain function, we introduce a consumer's product selection model. Based on MNL model, utility values give the probabilities that a consumer selects one from plural selectable products [25, 26]. Let $P_{q_i,s_j,k}$ denote the probability that consumer C_k tries to purchase product q_i in state s_j. The utility value when consumer C_k tries to purchase product q_i, denoted by $V_{q_i,s_j,k}$, is given by a linear regression model

$$V_{q_i,s_j,k} := \alpha_i^{j,k} + \sum_l \beta_{i,l}^{j,k} Y_l^j, \tag{3.1}$$

where we assume product q_1 is the reference, $\alpha_i^{j,k}$ is a constant, and $\beta_{i,l}^{j,k}$ is the coefficient of each explanatory variable Y_l^j. Then the probability that consumer C_k selects product q_i in state s_j is given by

$$P_{q_i,s_j,k} = \frac{\exp(V_{q_i,s_j,k})}{\sum_{l=1}^{n} \exp(V_{q_l,s_j,k})}. \tag{3.2}$$

Remark that utility values and the selection probabilities are defined not only for products in the assortment, but also for products not in the assortment.

3.3.2 Multinomial Logit Model

Using the probability Eq. (3.2), we give the purchase probability of product q_i by each consumer. Let N be the number of consumers and let X_i denote the stochastic variable representing the number of sales for product q_i without any restriction on the assortment. The purchase probability $\Pr(X_i = r \mid s_j)$, where the sales of product q_i amount to r as a result of N consumers attempting to purchase under state s_j, follows Poisson binomial distribution [27]. Poisson binomial distribution is explained as follows. We consider N independent trials each of which has its own success probability. Then Poisson binomial distribution is the discrete probability distribution of the number of successes from the N trials that can be computed recursively by

$$\Pr(X_i = r \mid s_j) = \begin{cases} \prod_{k=1}^{N} (1 - P_{q_i,s_j,k}) & \text{if } r = 0 \\ \\ \frac{1}{r} \sum_{l=1}^{r} (-1)^{l-1} \Pr(X_i = r - l \mid s_j) \Upsilon(l) & \text{if } r > 0 \end{cases}, \tag{3.3}$$

where $P_{q_i,s_j,k}$ represents the probability that the k-th consumer purchases product q_i under state s_j, defined by Eq. (3.2), and

$$\Upsilon(l) = \sum_{k=1}^{N} \left(\frac{P_{q_i,s_j,k}}{1 - P_{q_i,s_j,k}} \right)^l. \tag{3.4}$$

Expected value of X_i under state s_j is

$$E[X_i \mid s_j] = \sum_{k=1}^{N} P_{q_i,s_j,k}. \tag{3.5}$$

3.4 POMDP

Partially Observable Markov Decision Processes (POMDPs) are widely recognized as a powerful framework for modeling sequential decision-making problems under partial observability. Notable recent applications include clinical decision-making, dialogue management, and robot control policies.

However, despite their strengths, POMDPs face significant challenges. Solving large-scale POMDPs, even approximately, is notoriously difficult. This complexity arises primarily from the *curse of dimensionality*, where the computational cost increases exponentially with the size of the belief state space, and the *curse of history*, due to the vast number of possible state-observation-action sequences.

The foundational work by Kaelbling *et al.* [28] introduced key approaches for addressing infinite-horizon problems, focusing on maximizing the value function. Their methods form the basis of many modern solutions. While challenges remain, recent advancements, including Monte Carlo methods, deep reinforcement learning, and Point-Based Value Iteration (Value Iteration, PBVI), have significantly expanded the practical applicability of POMDPs in various domains.

Notation

POMDP is a model of an agent that synchronously interacts with environment. Given a discrete set Z, let $\Pi(Z)$ denote the set of all discrete probability distributions on Z. Formally, POMDP is defined as a tuple $POMDP = (S, A, \delta, R, \Omega, O)$, where S is the finite set of states, A is the finite set of actions, $\delta : S \times A \to \Pi(S)$ is the state transition function, $R : S \times A \to \mathbb{R}$ is the reward function, Ω is a finite set of observations, and $O : S \times A \to \Pi(\Omega)$ is the observation function.

Since the state has to be estimated through the observation function, Kaelbling's

method introduces a *belief*. A belief is a variable that represents what the current state is, and it is estimated from the history of observations.

At each time step, the agent choose an action to maximize the expected reward depending on the belief. A policy is a description of the behavior of the agent.

3.4.1 General Solution to POMDPs: Value Iteration

In POMDP, the agent cannot directly observe the true state of the system. Instead, it maintains a *belief state* $b(s)$, which represents a probability distribution over all possible states.

For a given belief state b, the POMDP policy is represented as a conditional plan structured as a tree. This conditional plan specifies rules for selecting actions based on observations received, allowing the agent to account for the uncertainty and dynamically adjust its behavior.

This approach enables the agent to operate effectively under partial observability by leveraging probabilistic reasoning about its environment.

The goal in a POMDP is to identify a policy π that maximizes the *value function*, defined as the expected sum of discounted rewards over a finite horizon h, starting from an initial belief state b_0:

$$V_\pi^h(b) = \mathbb{E}\left[\sum_{t=0}^{h} \gamma^t r_t \mid b_0 = b\right], \tag{3.6}$$

where r_t is the reward received at time t, γ is the discount factor ($0 < \gamma < 1$), and b_0 is the initial belief state.

Characteristics of the Value Function and Policy

For a finite horizon h :

- The optimal policy can be expressed as an h-step conditional plan (π_h).
- The value function for a finite horizon is *piecewise linear and convex* with respect to the belief state b. It can be represented as the maximum over a

finite set of α-vectors:

$$V^h(b) = \max_{\alpha_i^h \in \Gamma^h} (\alpha_i^h \cdot b), \tag{3.7}$$

where b and each α_i^h are vectors of dimension $|S|$ (the number of states).

Infinite horizon approximation For discount factors ($\gamma < 1$), the value function for an infinite horizon can be approximated arbitrarily closely using a sufficiently large finite horizon h [28].

Computation via Dynamic Programming

Initialization : At $h = 0$, the value function is initialized as $V^0(b) = 0$ with $\Gamma^0 = \{\alpha_1^0 = 0\}$.

Inductive computation : Using dynamic programming, the set of α-vectors for stage h, Γ^h, is computed from Γ^{h-1} through a backup operation. This operation updates the value function based on actions, observations, and transitions in the belief space.

This formulation provides a foundation for solving POMDPs by leveraging the structure of the belief space and the convexity of the value function, enabling efficient computation of optimal or near-optimal policies despite the problem's inherent complexity.

3.4.2 Other Solutions

Several solution methods have been proposed for solving POMDPs, in addition to Value Iteration. Bellow, we will illustrate the characteristics of these methods, including Value Iteration, and explain which types of problems each method is effective for.

Value Iteration is a method that iteratively updates the value function to determine the optimal policy. It represents the belief-state value function using α-vectors and guarantees an optimal solution. However, its computational cost is extremely high, making it impractical for large-scale problems.

Policy Iteration directly optimizes the policy through alternating evaluation and improvement steps. It can sometimes converge faster than Value Iteration, but it requires solving large linear programs, which can make it infeasible for problems with many states or observations.

Monte Carlo methods use random sampling to estimate the optimal policy. These methods are well-suited for large-scale problems because they approximate the solution without explicitly computing the full belief-state space. However, they do not guarantee an optimal solution, and their accuracy depends on the number of samples used [29].

POMCP (Partially Observable Monte Carlo Planning) improves upon standard Monte Carlo methods by integrating Monte Carlo Tree Search (MCTS) with belief-state updates. It is particularly effective for large and dynamic problems, as it can efficiently update its decisions based on new observations. While POMCP significantly reduces computational complexity, it only provides approximate solutions and does not guarantee global optimality [30].

Value Iteration and Policy Iteration are theoretically optimal but are only feasible for small problems due to their high computational cost. On the other hand, Monte Carlo methods and POMCP sacrifice optimality for efficiency, allowing them to handle large-scale and complex problems more effectively. Among them, POMCP is particularly useful for real-time decision-making in constantly changing environments.

Ultimately, the choice of a solution method depends on the problem size, computational resources, and the need for an exact solution.

Chapter 4

General Formulation for Assortment Optimization Problems (AOP)

Before discussing assortment optimization problem, we outline the general formulation. The assortment optimization problem for vending machines is defined by a 7-tuple $AOP = (\boldsymbol{A}, S, \delta, G, O, \pi, C)$, where $\boldsymbol{A}$ is the set of assortments, S is the set of states, δ is the state transition function, G is the gain function, O is the observation function, π is the policy, and C is the assortment constraints. Details are described in this chapter.

4.1 Products and Assortment

Consider n kinds of products $q_i (i = 1, ..., n)$ and m columns ($m > 0$, normally $n > m$), where columns of a vending machine are containers for stocking products. An assortment is a combination of selecting m products from n kinds of products allowing duplication. Such a combination is represented by a multiset. Multiset is a concept of set that combines the degree of duplication of how many elements are included when the set contains multiple elements of the same value. $\#X[e]$ represents the number of e included in the multiple set X. We denote $e \in X$ if $\#X[e] > 0$.

Let $\boldsymbol{A} = \{\boldsymbol{a}_1, \ldots, \boldsymbol{a}_L\}$ denote the set of all assortments, where L is the total number of assortments. The assortment given at time t is denoted by $\boldsymbol{a}(t)$. Note

that $\boldsymbol{a}(t+1)$ takes effect on sales between time t and $t+1$.

It should be noted that time t does not represent conventional time but rather a sequence of individual visit opportunities by the agent. As a result, the interval between visit opportunities t and $t+1$ is not constant in actual time.

As discussed in Section 1.3, agents can only obtain information at each visit, and in practice, they schedule visits when sales remain relatively stable. Considering this assumption, we adopt this approach.

Hereafter, we refer to these visit opportunities as time.

We assume that every column has the same capacity, and let *cap* denote the capacity of each column. Then the number of product q_i in the assortment $\boldsymbol{a}(t)$ is

$$stk(\boldsymbol{a}(t), q_i) := \#\boldsymbol{a}(t)[q_i] \cdot cap, \tag{4.1}$$

where $\#\boldsymbol{a}(t)[q_i]$ is the number of occurrences of q_i in the multiset $\boldsymbol{a}(t)$ and we assume each column is full after replenish work.

4.2 State Space

Let $S = \{s_1, \ldots, s_v\}$ be the set of states, where each state s_i is a u-dimensional vector and each component of a sate can be a real number, an integer or a discrete value. The states of vending machines consists of environment, weather, background population for the purchase at the vending machine, etc.

4.3 State Transition Function

The state at time t is denoted by $s(t)$. We define the state transition probability as a function $\delta\colon \mathbb{N} \times S \times S \to [0,1]$, where

$$\forall t, s_j : \sum_{j'} \delta(t, s_j, s_{j'}) = 1. \tag{4.2}$$

It means that the probability that $s(t) = s_j$ and $s(t+1) = s_{j'}$ is $\delta(t, s_j, s_{j'})$. When the state transition probability depends on time t, it is called time variant,

otherwise it is called time invariant. In the time invariant case, δ is defined as $\delta : S \times S \to [0, 1]$.

4.4 Gain Function

The gain function for assortments is defined as a function $G : S \times \boldsymbol{A} \to \mathbb{N}$ that gives the total sales (amount or unit) under a given state and an assortment. $G(s_j, \boldsymbol{a}_l)$ is given by the sum of the sales of all products:

$$G(s_j, \boldsymbol{a}_l) := \sum_{q_i \in \boldsymbol{a}_l} g_i, \tag{4.3}$$

where g_i is the sales of product q_i. The vector $\boldsymbol{g} := [g_1, \ldots, g_n]$ is called *the gain vector*. Note that we implicitly assume all products have the same price. How to derive the gain function is explained in the next section.

4.5 Observation Function

The observation function is defined as $O : S \to W$, where W is some set. The observation at time t is denoted by $o(t) := O(s(t))$. As we have defined, each state s is represented by a u-dimensional vector $s_i := [s_i^1, \ldots, s_i^u]$. In this study, we assume that the observation function masks some of the substates, e.g., for state $s_i = [s_i^1, s_i^2, s_i^3, s_i^4]$, $O(s_i) = [s_i^1, s_i^4]$, it means that the function masks the second and the third substates. Here the masked substates imply unobservable substates and the others imply observable ones.

4.6 Policy

At time t, given all past information about a vending machine, a function that outputs $\boldsymbol{a}(t)$ is called a policy π:

$$\boldsymbol{a}(t) = \pi(o(\tilde{t}), \boldsymbol{a}(\tilde{t}), \boldsymbol{g}(\tilde{t})), \quad \tilde{t} = 0, \ldots, t-1, \tag{4.4}$$

A policy is a description of the behavior of the agent. Here, $\tilde{t} = 0, \ldots, t-1$ represents all past time steps, and $o(\tilde{t}), \boldsymbol{a}(\tilde{t}), \boldsymbol{g}(\tilde{t})$ denote the sequences of all past observations, actions, and rewards, respectively.

4.7 Assortment Constraint

The assortment constraint is a set $C \subseteq \boldsymbol{A} \times \boldsymbol{A}$. For any time t, $(\boldsymbol{a}(t), \boldsymbol{a}(t+1)) \in C$ has to be satisfied. The reason why this constraint arises is that the number of products the route man can exchange at each time is limited. This constraint characterizes the assortment optimization problem for vending machines.

4.8 Assortment Optimization Problem

We now define the assortment optimization problem in this study.

Find a policy that satisfies the assortment constraint and maximizes the total gain during time $t = 0, \ldots, T$.

We can also classify the problem by the following characteristics: The state space is known / unknown for the agent, complete / incomplete observation, gain function is known / unknown, and transition probability is known / unknown. Examples are

- Stores (such as convenience stores, supermarkets): state is known, complete observation, gain function is known.
- Vending machines: state is known (or unknown), incomplete observation, gain function is known.

Chapter 5

Proposal method for AOP

This chapter describes the details of our proposed method for solving the vending machine assortment optimization problem. The method consists of two components. The first is the Product Selection Model, which models how consumers choose products from the vending machine, specifically by calculating the selection probability for each product. The second is the model for how the agent selects assortments, where we adopt a model based on POMDP.

5.1 Product Selection Model: Selection Probability

Next we consider the probability under a given assortment. We assume the static substitution. Since the amount of actual sales g_i is constrained by the assortment, the probability under state s_j and assortment $\boldsymbol{a}_h$ is obtained as follows

$$\Pr(g_i = r \mid s_j, \boldsymbol{a}_h) = \begin{cases} 0 & \text{if } r > stk(\boldsymbol{a}_h, q_i) \\ \sum_{l=r}^{N} \Pr(X_i = l \mid s_j) & \text{if } r = stk(\boldsymbol{a}_h, q_i) \\ \Pr(X_i = r \mid s_j) & \text{if } r < stk(\boldsymbol{a}_h, q_i) \end{cases} . \tag{5.1}$$

Due to the capacity constraint, all cases $r \leq X_i \leq N$ reduce to $X_i = r$. Also, the expected value of the gain function under state s_j and assortment $\boldsymbol{a}_h$ is given by

$$E[G(s_j, \boldsymbol{a}_h)] = \sum_{q_i \in \boldsymbol{a}_h} \min\left\{stk(\boldsymbol{a}_h, q_i), E[X_i \mid s_j]\right\} . \tag{5.2}$$

In other words, it is the sum, across all products, of the smaller value between the expected sales in state s_j and the inventory quantity for assortment $\boldsymbol{a}_h$.

5.2 POMDP model for AOP

Following Kaelbling *et al.* [28], we propose a POMDP-based method that select a good assortment policy from a given set of policies. POMDP is a stochastic process that deals with situations where the state can be partially observed, and these observations do not necessarily satisfy Markov process.

The POMDP model for the assortment optimization problem is described as follows.

5.2.1 State

The set of states in POMDP is given as the set of states in AOP. The state at time t is denoted by $s(t)$.

The change in state depending on t is referred to as "state transition", denoted as $s(t) \rightarrow s(t+1)$. It is assumed that the state transition $s(t) \rightarrow s(t+1)$ occurs immediately after the agent completes the assortment $a(t)$. In other words, $s(t+1)$ can be considered as the external environment (such as consumers and temperature) during the period $(t, t+1]$, while $a(t)$ represents the initial product assortment immediately after time t. The internal state of the vending machine, which depends on $s(t+1)$ and changes over time.

As a result, the state remains $s(t+1)$ throughout the period $(t, t+1]$, and consumers are assumed to make purchasing decisions based on $s(t+1)$ during this period.

5.2.2 Action

The action at time t is given as the assortment $\boldsymbol{a}(t)$ resulting from the exchange operations performed by the agent. Let us reiterate that $\boldsymbol{a}(t+1)$ takes effect on sales between time t and $t+1$.

5.2.3 State Transition Probability

The state transition probability from time t to $t+1$ is denoted by $\delta(t, s(t), s(t+1))$. We assume that assortments do not affect state transitions, and that the state transition probabilites are time invariant. So we denote the probabilites by $\delta(s(t+1) \mid s(t))$. We define the state transition probability for n time steps, denoted by δ^n, as

$$\begin{aligned} \delta^1(s' \mid s) &:= \delta(s' \mid s) \\ \delta^n(s' \mid s) &:= \sum_{s'' \in S} \delta^{n-1}(s' \mid s'')\delta(s'' \mid s). \end{aligned} \tag{5.3}$$

5.2.4 Observation

The possible observed state $o(t)$ and the sales vector $\boldsymbol{g}(t) = [g_1(t), \cdots, g_n(t)]$ at time t are stochastically given depending on the state $s(t)$ and the assortment $\boldsymbol{a}(t)$. Suppose that $\boldsymbol{g}(t) = [r_1, \ldots, r_n]$. Then the probability is given by

$$\begin{aligned} O(s(t), \boldsymbol{a}(t-1), o(t), \boldsymbol{g}(t)) &:= \Pr(o(t), \boldsymbol{g}(t) \mid s(t), \boldsymbol{a}(t-1)) \\ &= \prod_{i=1}^{n} \Pr(o(t), g_i(t) = r_i \mid s(t), \boldsymbol{a}(t-1)) \\ &= \prod_{i=1}^{n} \Pr(g_i(t) = r_i \mid s(t), \boldsymbol{a}(t-1)). \end{aligned} \tag{5.4}$$

Note that the observation $o(t)$ and gain $g(t)$ at time t are probabilistically determined by the state $s(t)$ at the same time and the assortment $a(t-1)$ at the previous time $t-1$. The last equality follows from the fact that $o(t)$ is uniquely determined from $s(t)$ as described in Section 4.5, i.e., the observation masks some substates.

$\Pr(g_i(t) = r_i \mid s(t), \boldsymbol{a}(t-1))$ is computed by Eq. (5.1).

5.2.5 Reward

At time t, the agent obtains the possible observed state $o(t)$ and the sales $g_i(t)$ of each product q_i. The total sales of products is regarded as the reward. Let $rw(t)$ denote the reward at time t, so

$$rw(t) := G(s(t), \boldsymbol{a}(t-1)) = \sum_{q_i \in \boldsymbol{a}(t-1)} g_i(t). \tag{5.5}$$

Here, $G(s(t), \boldsymbol{a}(t-1))$ represents the gain function obtained at the end of consumer purchasing behavior during the period $(t-1, t)$, given the state $s(t)$ and the initial product assortment $\boldsymbol{a}(t-1)$.

5.2.6 Policy

A policy on assortment exchange is a function that gives the assortment at each time. We assume that the policy π depends only on the latest state $s(t)$, the observed values $o(t)$, and the latest assortment $\boldsymbol{a}(t-1)$. Also, it is assumed that the agent can select one policy from a finite set of policies $\Pi := \{\pi^1, \ldots, \pi^M\}$.

5.2.7 Belief

The belief is represented by a function $b : S \to \mathbb{R}$ such that $0 \leq b(s) \leq 1$, and

$$\sum_{s_j \in S} b(s_j) = 1. \tag{5.6}$$

Let b_t denote the belief at time t. For each state $s \in S$, $b_t(s)$ is the strength that the agent believes $s(t) = s$.

At each time, the policy on assortment exchanges is decided by the current belief. Given a belief b_{t-1} at time $t-1$, the belief that $s(t) = s'$ under the observed

state $o(t)$ and the sales $\boldsymbol{g}(t)$ is given by

$$
\begin{aligned}
b_t(s') &= \Pr(s' \mid o(t), \boldsymbol{a}(t-1), \boldsymbol{g}(t), b_{t-1}) \\
&= \frac{\Pr(o(t), \boldsymbol{g}(t) \mid s', \boldsymbol{a}(t-1), b_{t-1}) \Pr(s' \mid \boldsymbol{a}(t-1), b_{t-1})}{\Pr(o(t), \boldsymbol{g}(t) \mid \boldsymbol{a}(t-1), b_{t-1})} \\
&= \frac{\Pr(o(t), \boldsymbol{g}(t) \mid s', \boldsymbol{a}(t-1))}{\Pr(o(t), \boldsymbol{g}(t) \mid \boldsymbol{a}(t-1), b_{t-1})} \\
&\quad \times \sum_{s \in S} \Pr(s' \mid \boldsymbol{a}(t-1), b_{t-1}, s) \Pr(s \mid \boldsymbol{a}(t-1), b_{t-1}) \\
&= \frac{O(s', \boldsymbol{a}(t-1), o(t), \boldsymbol{g}(t))}{\Pr(o(t), \boldsymbol{g}(t) \mid \boldsymbol{a}(t-1), b_{t-1})} \times \sum_{s \in S} \delta(s' \mid s) b_{t-1}(s), \qquad (5.7)
\end{aligned}
$$

where the denominator $\Pr(o(t), \boldsymbol{g}(t) \mid \boldsymbol{a}(t-1), b_{t-1})$ can be treated as a normalizing factor.

Based on the notation above, the state transition diagram for this model is illustrated in Fig. 5.1.

At time t, the vending machine is in state $s(t)$. At this time, the agent selects a policy π_t for the vending machine, which determines the assortment $\boldsymbol{a}(t)$ to be implemented.

During the transition from t to $t+1$, the state changes to $s(t+1)$ according to the transition probability $\delta(s(t+1) \mid s(t))$. After the transition to $s(t+1)$, consumers' purchasing behavior occurs at the vending machine. The sales of products depend on $s(t+1)$ and $\boldsymbol{a}(t)$ and are represented by the gain function $G(s(t+1), \boldsymbol{a}(t))$.

Following the purchasing behavior, the agent visits the vending machine to obtain an observation $o(t+1)$ and receives a reward $rw(t+1) = G(s(t+1), \boldsymbol{a}(t))$. Using $o(t+1)$, the agent estimates the current state $s(t+1)$ through the belief $b_{t+1}(s(t+1))$. The agent then selects a new policy π_{t+1} and implements the assortment $\boldsymbol{a}(t+1)$.

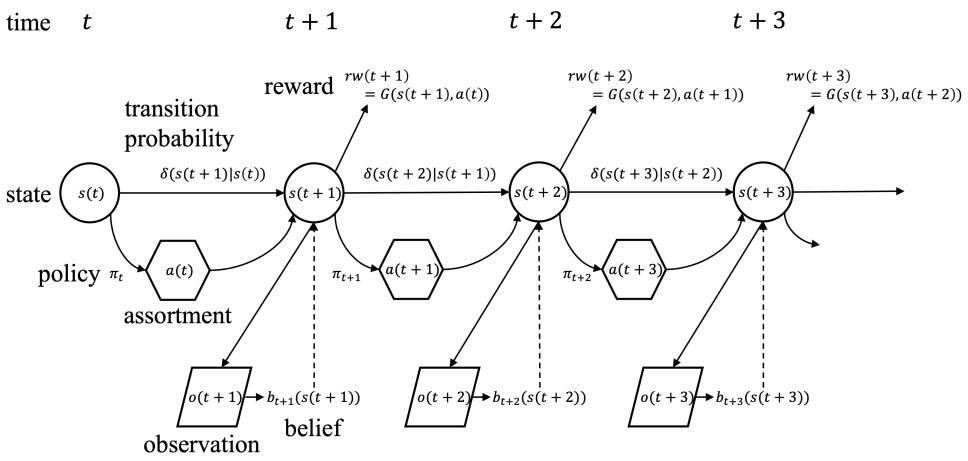

Figure 5.1: State Transition Diagram in the POMDP model

At time t, the vending machine is in state $s(t)$, and the agent executes assortment $\boldsymbol{a}(t)$ according to the policy π_t. When time t progresses to $t+1$, the state transitions to $s(t+1)$ based on the transition probability $\delta(s(t+1) \mid s(t))$. The agent receives a reward $rw(t+1)$ and observes $o(t+1)$, from which it estimates $s(t+1)$ using the belief $b_{t+1}(s(t+1))$. Based on the calculated belief, the agent selects the next policy $\pi(t+1)$ and executes assortment $\boldsymbol{a}(t+1)$. This process of state transitions, observations, policy selection, and assortment execution is repeated iteratively.

5.3 Strategy for Assortment Exchange

Based on the previous section, we describe the procedure for determining the strategy for assortment exchange at time t. In this strategy, the agent determines the policy for assortment exchange according to the following steps:

(1) Calculate the belief of the current vending machine state based on the observation obtained at time t.

(2) For each possible policy, calculate the expected rewards for the future states at time $t+1, t+2, \ldots, t+\tau$ assuming the assortment corresponding to the policy is implemented. Sum these expected rewards, applying a discount factor to account for the diminishing value of future rewards.

(3) Perform the calculation in step (2) for all possible policies and select the policy that maximizes the total expected reward.

Estimate Beliefs

At time $t = 0$, we assume that $s(0) = s_j$ is equally likely for all possible states $s_j \in S$. In other words, the belief $b_0(s(0))$ is

$$b_0(s_1) = b_0(s_2) = \cdots = b_0(s_v) = \frac{1}{v}. \tag{5.8}$$

At time $t > 0$, the observed value $o(t)$ and the sales $\boldsymbol{g}(t)$ are obtained. Using Eq. (5.7), we update the belief by

$$b_t(s_j) = O(s_j, \boldsymbol{a}(t-1), o(t), \boldsymbol{g}(t)) \times \sum_{s \in S} \delta(s_j \mid s) b_{t-1}(s). \tag{5.9}$$

After the update, b_t is normalized so that

$$\sum_{s_j \in S} b_t(s_j) = 1. \tag{5.10}$$

Calculate Expected Reward at time $t + 1$

The reward obtained at time t is determined by the gain function based on the assortment $\boldsymbol{a}(t-1)$ at time $t-1$ and the state $s(t)$ at time t, as shown in Eq. (5.5). The expected reward $E[rw(t)]$ is calculated as the product of the belief for all possible states $s(t)$ and the expected value of the gain function when each state $s(t)$ is realized.

$$E[rw(t)] = \sum_{s \in S} b_t(s) E[G(s, \boldsymbol{a}(t-1))]. \tag{5.11}$$

The policy on assortment exchanges $\pi_t^{k_0}(k_0 = 1, \ldots, M) \in \Pi$ is decided based on the expected reward obtained in the future. First, we consider the expected value of the reward $rw(t+1)$ at time $t+1$. In the case where $\pi_t^{k_0} : \boldsymbol{a}(t-1) \to \boldsymbol{a}^{k_0}(t)$ is

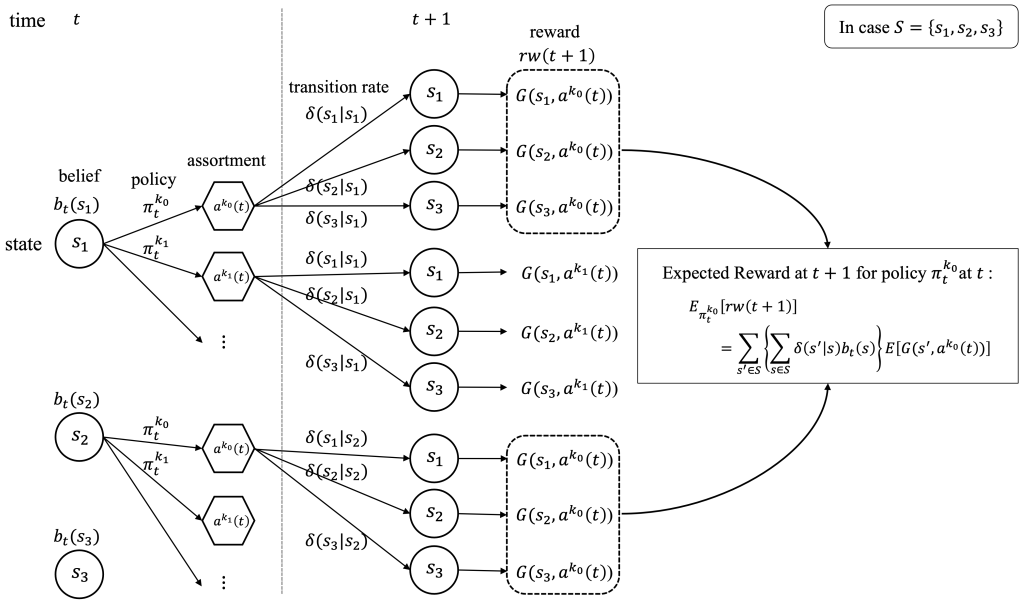

Figure 5.2: Sample process for expected reward at time $t+1$
Consider the agent's policy selection at time t. For simplicity, the set S is limited to three states: $\{s_1, s_2, s_3\}$.
At time t, the agent selects policy $\pi_t^{k_0}$ and executes assortment $\boldsymbol{a}^{k_0}(t)$.
The expected reward obtained at time $t+1$ is the sum of the expected rewards $E[G(s', \boldsymbol{a}^{k_0}(t))]$ for all possible combinations of paths $s(t) = s$ and $s(t+1) = s'$. In this calculation, the transition probabilities and beliefs for each path are multiplied and summed to obtain the final value.

selected, the expected reward at time $t+1$ is given as follows:

$$E_{\pi_t^{k_0}}[rw(t+1)] = \sum_{s' \in S} \left\{ \sum_{s \in S} \delta(s' \mid s) b_t(s) \right\} E[G(s', \boldsymbol{a}^{k_0}(t))]. \tag{5.12}$$

An example of the process described so far, for the case where $S = \{s_1, s_2, s_3\}$, is illustrated in Fig. 5.2.

Calculate Expected Reward at time $t+2$

Next, we consider the case that $\pi_{t+1}^{k_1}(k_1 = 1, \ldots, M)$ is selected at time $t+1$. The expected value of the reward $rw(t+2)$ is calculated for the each assortment $\boldsymbol{a}^{k_0}(t), \boldsymbol{a}^{k_1}(t+1)$. Note that the assortment at time $t+1$ depends on the assortment

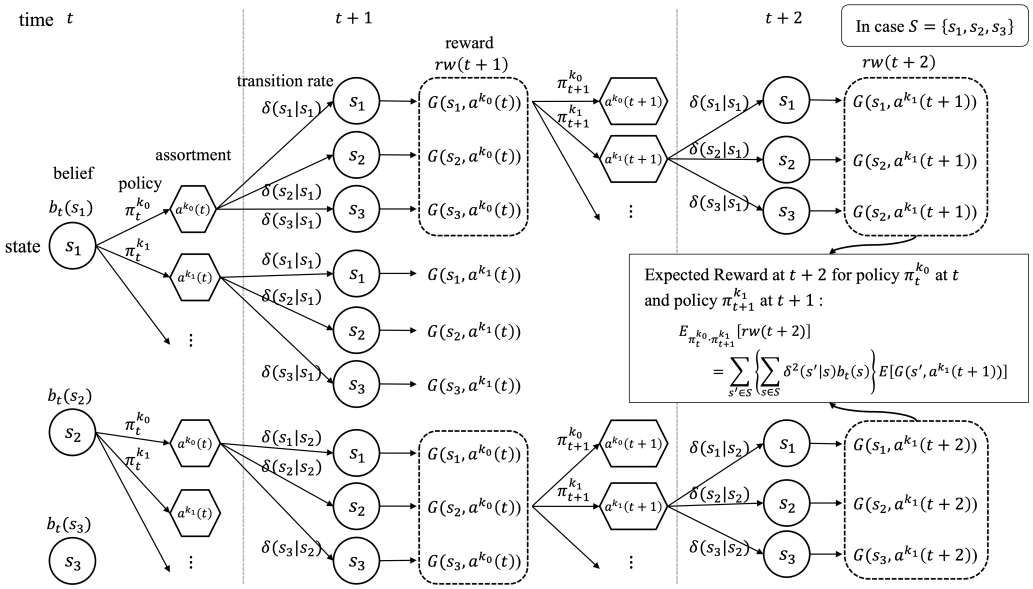

Figure 5.3: Sample process for expected reward at time $t + 2$

Continuing with the consideration of the agent's policy selection at time t, as shown in Figure 5.2, the agent is assumed to select policy $\pi_{t+1}^{k_1}$ and execute assortment $\boldsymbol{a}^{k_1}(t+1)$ at time $t+1$.
The expected reward obtained at time $t+2$ is the sum of the expected rewards $E[G(s', \boldsymbol{a}^{k_1}(t+1))]$ at time $t+2$ for $s(t+2) = s'$, calculated over all possible combinations of paths $s(t+1) = s$ and $s(t+2) = s'$. In this calculation, the transition probabilities and beliefs for each path are multiplied and summed.

at time t because of the assortment constraints.

$$
\begin{aligned}
&E_{\pi_t^{k_0} \cdot \pi_{t+1}^{k_1}} [rw(t+2)] \\
&\quad = \sum_{s' \in S} \left\{ \sum_{s \in S} \delta(s' \mid s) b_{t+1}(s) \right\} E[G(s', \boldsymbol{a}^{k_1}(t+1))] \\
&\quad = \sum_{s' \in S} \left\{ \sum_{s \in S} \sum_{s'' \in S} \delta(s' \mid s'') \delta(s'' \mid s) b_t(s) \right\} E[G(s', \boldsymbol{a}^{k_1}(t+1))] \\
&\quad = \sum_{s' \in S} \left\{ \sum_{s \in S} \delta^2(s' \mid s) b_t(s) \right\} E[G(s', \boldsymbol{a}^{k_1}(t+1))] \qquad (5.13)
\end{aligned}
$$

Here, Eq. (5.3) is applied. An example of this process is illustrated in Fig. 5.3

Total Expected reward from $t+1$ to $t+\tau$

Similarly at time $t+\tau$ ($\tau > 0$), we can obtain the expected value of the reward as follows:

$$E_{\pi_t^{k_0} \cdots \pi_{t+\tau-1}^{k_{\tau-1}}}[rw(t+\tau)] = \sum_{s' \in S} \left\{ \sum_{s \in S} \delta^{\tau}(s' \mid s) b_t(s) \right\} E[G(s', \boldsymbol{a}^{k_{\tau-1}}(t+\tau-1))]. \tag{5.14}$$

Therefore, we can calculate the maximum expected reward in the future when the policy $\pi_t^{k_0}$ is selected at time t. When the policy $\pi_t^{k_0}$ is selected, the total expected reward $E_{t \to t+\tau}(\pi_t^{k_0})$ is calculated as the sum of them from $t+1$ to $t+\tau$. As usual in POMDP, in order to make recent rewards more effective, we multiply the future

expected reward by the discount rate $\gamma(0 < \gamma < 1)$.

$$
\begin{aligned}
E_{t\to t+1}(\pi_t^{k_0}) &= E_{\pi_t^{k_0}}[rw(t+1)] \\
E_{t\to t+2}(\pi_t^{k_0}) &= E_{\pi_t^{k_0}}[rw(t+1)] + \gamma \max_{\pi_{t+1}^{k_1}\in\Pi}\left\{E_{\pi_t^{k_0}\cdot\pi_{t+1}^{k_1}}[rw(t+2)]\right\} \\
E_{t\to t+3}(\pi_t^{k_0}) &= E_{\pi_t^{k_0}}[rw(t+1)] + \gamma \max_{\pi_{t+1}^{k_1}\in\Pi}\left\{E_{\pi_t^{k_0}\cdot\pi_{t+1}^{k_1}}[rw(t+2)]\right. \\
&\qquad \left. + \gamma \max_{\pi_{t+2}^{k_2}\in\Pi}\left\{E_{\pi_t^{k_0}\cdot\pi_{t+1}^{k_1}\cdot\pi_{t+2}^{k_2}}[rw(t+3)]\right\}\right\} \\
&\vdots \\
E_{t\to t+\tau}(\pi_t^{k_0}) &= E_{\pi_t^{k_0}}[rw(t+1)] \\
&\qquad + \gamma \max_{\pi_{t+1}^{k_1}\in\Pi}\left\{E_{\pi_t^{k_0}\cdot\pi_{t+1}^{k_1}}[rw(t+2)]\right. \\
&\qquad\quad + \gamma \max_{\pi_{t+2}^{k_2}\in\Pi}\left\{E_{\pi_t^{k_0}\cdot\pi_{t+1}^{k_1}\cdot\pi_{t+2}^{k_2}}[rw(t+3)]\right. \\
&\qquad\qquad + \gamma \max_{\pi_{t+3}^{k_3}\in\Pi}\left\{E_{\pi_t^{k_0}\cdot\pi_{t+1}^{k_1}\cdot\pi_{t+2}^{k_2}\cdot\pi_{t+3}^{k_3}}[rw(t+4)]\right. \\
&\qquad\qquad \left.\left.\left. \cdots + \gamma \max_{\pi_{t+\tau-1}^{k_{\tau-1}}\in\Pi}\left\{E_{\pi_t^{k_0}\cdots\pi_{t+\tau-1}^{k_{\tau-1}}}[rw(t+\tau)]\right\}\cdots\right\}\right\}\right\}
\end{aligned}
\tag{5.15}
$$

Note that the expected sales in far future is not very important in vending machines, we consider rewards within a finite horizon.

Select the Ppolicy on the Assortment

According to the above procedure, we obtain the total expected rewards $E_{t\to t+\tau}(\pi_t^1), \ldots, E_{t\to t+\tau}(\pi_t^M)$ by the policies $\pi_t^1, \ldots, \pi_t^M$. Then, the policy on the assortment exchange π_t is decided as one that maximizes the expected reward:

$$
\pi_t = \arg\max_{\pi_t^{k_0}\in\Pi} E_{t\to t+\tau}(\pi_t^{k_0}). \tag{5.16}
$$

However, the π_t chosen here must ensure that the assortment implemented based on π_t satisfies the assortment constraint. That is,

$$\pi_t : \boldsymbol{a}(t-1) \to \boldsymbol{a}(t), \quad (\boldsymbol{a}(t-1), \boldsymbol{a}(t)) \in C. \tag{5.17}$$

Equation (5.16) represents the decision-making policy of the agent at time t. Specifically, the agent first selects one policy $\pi_t^{k_0} = \pi^0 \in \Pi$ from the available policies and assumes that the assortment based on this policy is executed. The agent then (mentally) calculates the expected future reward $E_{t\to t+\tau}(\pi_t^{k_0})$ over the period from $t+1$ to $t+\tau$ using Eq. (5.16). Next, the agent calculates the expected reward $E_{t\to t+\tau}(\pi_t^{k_0})$ for another policy $\pi_t^{k_0} = \pi^1 \in \Pi$. Similarly, the agent evaluates $E_{t\to t+\tau}(\pi_t^1), \ldots, E_{t\to t+\tau}(\pi_t^M)$. The agent then selects the policy $\pi_t^{\max}$ that maximizes the expected reward $E_{t\to t+\tau}(\pi_t^{\max})$ and determines it as the policy to execute at time t.

5.4 Baseline model

Thus far, we have explained our proposed model and the strategy selection policy for optimal assortment. In the next chapter, we will simulate and evaluate these models. However, before proceeding, we will discuss the metrics used for model evaluation.

Various approaches can be considered for model evaluation. In this study, we compare the performance of our model against that of an ideal agent executing the best possible assortment actions. Specifically, we assume an agent with perfect knowledge of all information

This ideal agent is assumed to have prior knowledge of the vending machine's state and state transitions at every time step, including unobservable factors, with 100% certainty. Additionally, the agent is assumed to know precisely which assortment yields the highest sales in each state. Such an agent can select the assortment that maximizes the reward at each time step based on the current and subsequent states, ensuring that the total expected reward across all time steps is

theoretically maximized.

It should be noted that for the evaluation of the proposed model, ideally, data should be collected from actual agents' product assortment processes and compared accordingly. However, as discussed in the Section 1.3, product assortment exchange by agents is largely unstructured, making it difficult to obtain data that would allow for a robust evaluation.

Therefore, in this study, we assumed an ideal agent and set the theoretically optimal assortment exchange they could achieve as the upper benchmark. The proposed model was then evaluated by comparing its results against this benchmark.

In this section, we examine two metrics based on the expected reward of such an ideal agent.

5.4.1 Theoretical Upper bound

First, we can derive the *theoretical upper bound* on the expected sales. If the agent explicitly knows the state $s(t)$ of the vending machine at time t, the agent can maximize the expected total sales by choosing an appropriate assortment from the set of the entire assortment $\boldsymbol{A}$ at time $t-1$. We can give the following upper bound of expected sales at each time t, without considering assortment constraint.

$$E_t^{Upper\ bound} = \max_{\boldsymbol{a}(t-1)\in A} E[G(s(t+1), \boldsymbol{a}(t))] \tag{5.18}$$

Figure 5.4 illustrates the concept of the theoretical upper bound. In this figure, for simplicity, the assortment $\boldsymbol{a}^{k_0}(t)$ implemented at time t according to the selected policy $\pi_t^{k_0}(k_0 = 1, \ldots, M)$ is denoted as $\boldsymbol{a}^{k_0}$.

In the calculation of the Theoretical Upper ound, it is assumed that an ideal agent "knows" the assortment that will maximize sales in the next time step and can select it accordingly. When considering the assortment at time $t+3$ in the diagram, the next optimal assortment $\boldsymbol{a}^{k_2}$ may not be selectable due to assortment constraints. However, for the calculation of the theoretical upper bound, this constraint is ignored, and it is assumed that $\boldsymbol{a}^{k_2}$ can be realized and the reward

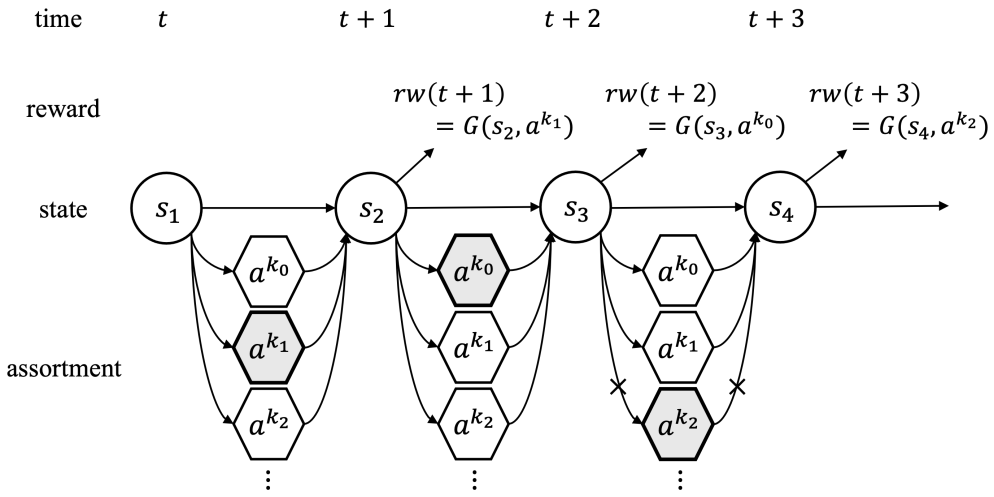

Figure 5.4: Sample process for Theoretical Upper bound

$rw(t+3) = G(s_4, \boldsymbol{a}^{k_2})$ is obtained. As a result, the expected reward obtained is maximized among all possible state transitions.

5.4.2 Feasible Maximum Value

By the assortment constraint, selection of the assortment at time t is constrained by the assortment at time $t-1$. We consider the *feasible maximum value* of expected sales under the assortment constraint. Let $\boldsymbol{a}_l$, $\boldsymbol{a}_m$ be two assortments and let $C(\boldsymbol{a}_l, \boldsymbol{a}_m)$ denote a Boolean variable such that $C(\boldsymbol{a}_l, \boldsymbol{a}_m) = 1$ if $(\boldsymbol{a}_l, \boldsymbol{a}_m) \in C$ and 0 otherwise. Then the expected sales considering assortment constraint at time t is given by

$$E_t(\boldsymbol{a}(0), \ldots, \boldsymbol{a}(t-1)) = \left(\prod_{i=1}^{t-1} C(\boldsymbol{a}(i-1), \boldsymbol{a}(i)) \right) \cdot E[G(s(t), \boldsymbol{a}(t-1))], \tag{5.19}$$

and the feasible maximum value at each time t is

$$E_t^{Feasible\ max} = \max_{\boldsymbol{a}(0),\ldots,\boldsymbol{a}(t-1) \in \boldsymbol{A}} E_t(\boldsymbol{a}(0), \ldots, \boldsymbol{a}(t-1)). \tag{5.20}$$

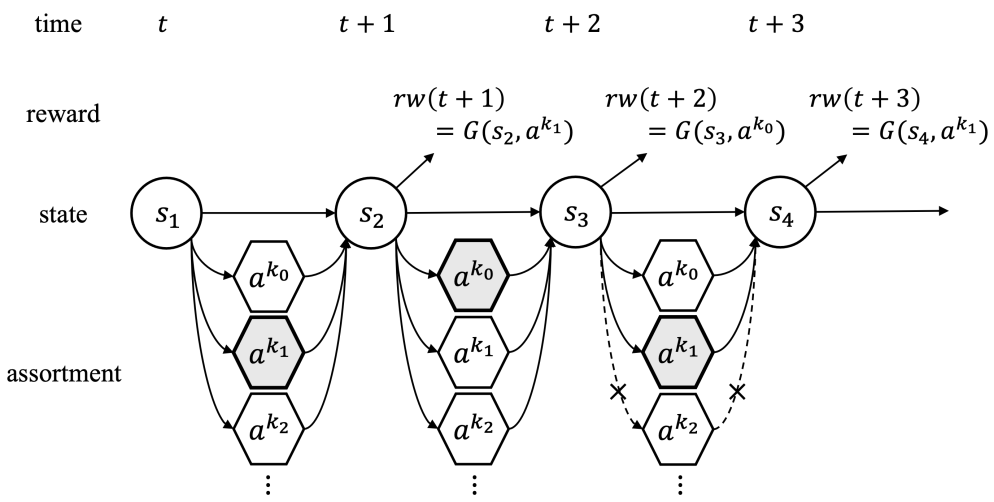

Figure 5.5: Sample process for Feasible max

In Fig. 5.4, consider the case where assortments that do not satisfy the assortment constraints cannot be implemented. The ideal agent selects the feasible assortment $\boldsymbol{a}^{k_2}$ within the constraints. The expected reward obtained in this case represents the maximum achievable value when the agent exerts its best effort under the given constraints.

An example of the feasible naximum value is shown in Fig. 5.5. In Fig. 5.4, consider the case where assortments that do not satisfy the assortment constraints cannot be implemented. But, for feasible maximum value, the ideal agent selects the feasible assortment $\boldsymbol{a}^{k_2}$ within the constraints. The expected reward obtained in this case represents the maximum achievable value when the agent exerts its best effort under the given constraints.

By comparing equations (5.18) and (5.20), it is clear that

$$E_t^{Feasible\ max} \leq E_t^{Upper\ bound}. \tag{5.21}$$

5.4.3 Assortment Constraints in Vending Machines

In the process of exchanging products in a vending machine, there are certain assortments that cannot be realized due to various constraints. For example, even

if the sales are good and the intention is to increase stock, there are limitations on the types and quantities of products that can be transported during restocking. Some columns cannot physically accommodate certain products due to their size, and the number of product swaps or columns that can be exchanged is limited by the employees' working hours. These are constraints that are specific to optimization problems when applied to vending machines.

Ignoring these constraints, the theoretical maximum expected reward is represented by $E_t^{Upper\ bound}$, while the maximum expected reward within the feasible range, considering the constraints, is $E_t^{Feasible\ max}$.

In the next chapter, these expected rewards will be used as baselines, and the rewards achieved by the model will be evaluated based on how closely they approach $E_t^{Upper\ bound}$ and $E_t^{Feasible\ max}$.

Chapter 6

Simulation model and Numerical Results

In this chapter, we present the numerical results obtained from computer simulations of assortment optimization problem for vending machines.

6.1 Models for Vending Machine

Based on the discussion in Chapter 5, we give the specific formulation of the assortment optimization problem for vending machines. In order to simplify the problem, we introduce some assumptions on products, agents and consumers.

6.1.1 Products and Agents

All products have the same shape and price, and the number of products that can be replenished in one column is also the same. The agent replenishes the vending machine with products and can exchange the assortment at its own discretion. The agent checks sales for each assortment at the next replenishment work. Vending machines have several states (e.g. environment, weather, background population for the purchase at the vending machine), and sales fluctuate depending on the state. The agent makes it possible to observe the state of all or part of the vending machine.

Note that these assumptions are based on the discussion in Section 1.3.

6.1.2 Consumer's Purchasing Behavior

We introduce simple assumptions on consumers who are purchasing products at the vending machine. A consumer has several attributes (gender, age, occupation, etc.), and the preferences for product selection probabilistically depend on these attributes together with the current state.

At time t, N consumers try to purchase products at the vending machine. Suppose that the k-th consumer $C_k (k = 1, \ldots, N)$ tries to purchase one of n kinds of products. The products the consumer C_k tries to purchase are determined probabilistically. We assume that the probabilities are determined by the attributes of the consumer C_k and the state of the vending machine $s(t)$. If the product to be purchased is present in the assortment $\boldsymbol{a}(t)$ and the inventory is sufficient, the consumer C_k purchases it. Otherwise (including in case of sold out), the consumer do not purchase any alternative products in that case, i.e., we assume the static substitution.

In the computer simulation, we assume that the attribute of the consumer is only gender: male or female. Other conditions and attributes are not considered.

6.2 Parameters and Assumptions

We show the parameters for the simulation and some assumptions.

6.2.1 Agent

In the simulation, we assume that the agent knows all attributes and parameters including transition probabilities between states. The agent cannot know the entire information on the current state and estimates it from the history of observations as the belief. Based on the belief, the agent selects the next policy.

6.2.2 State of Vending Machine

Originally, the state of vending machines can be considered to have many parameters and factors. The states can be classified into two types: observable and unobservable. In this simulation, we consider three states: location, temperature and gender ratio.

Location is observable and classified by three types: office, outdoor and school. In practice, the location is observable and fixed in the simulation targeting a single vending machine. In this study, the location functions as a factor that characterizes the elements defined later, such as products, utility values, other states, and state transitions.

Temperature is an external factor for vending machines. It is observable and is selected from one of $\{high, middle, low\}$ at each time.

The gender ratio is an internal factor for vending machines, and it means the ratio of males and females among consumers who are going to purchase at the vending machine. In the simulation, three patterns are assumed: $\{8:2, 5:5, 2:8\}$. The ratio is selected from one of them at each time. The ratio is unobservable and it cannot be known to the agent. The image of the vending machine states is shown in Fig.6.1.

6.2.3 Products and Assortment

All products have the same shape and price, and the number of products that can be replenished in one column is also the same. Products that can be in the assortment are 15 kinds: $\mathsf{A}, \ldots, \mathsf{O}$. The vending machine has 10 columns, and the capacity of each columns is 20 for any kind of products. It is possible to assign the same kind of products to multiple columns. Fig. 6.2 depicts an assortment and stocks.

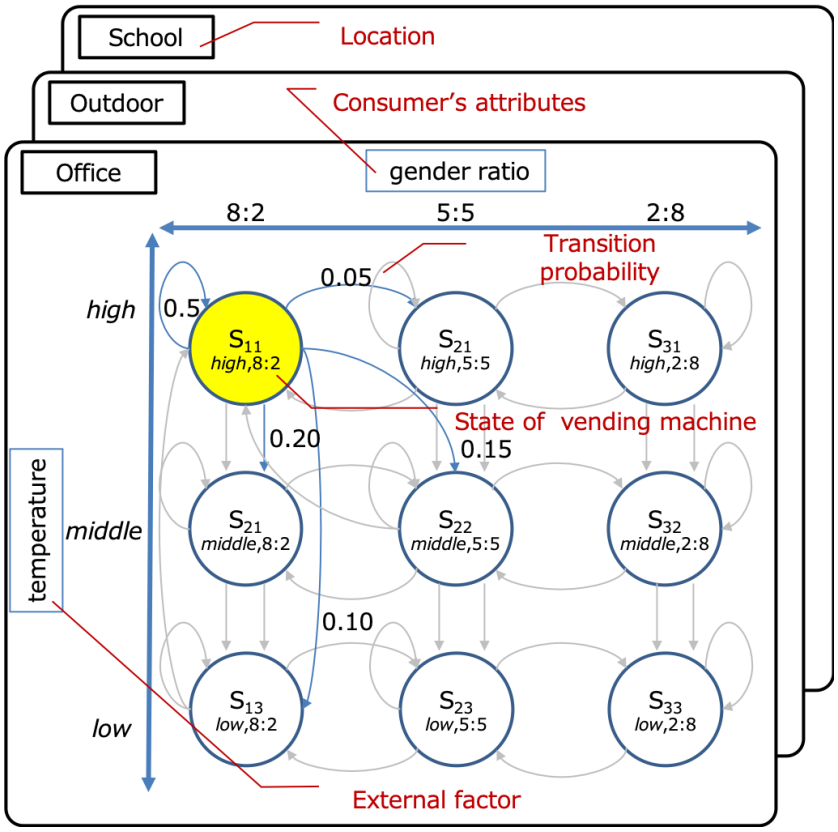

Figure 6.1: State of Vending Machine

A vending machine has three types of states: location, temperature, and gender ratio, each of which can be divided into three levels. Location is an observable and static state that does not change. Temperature is an observable and dynamic state influenced by external factors, while the gender ratio is an unobservable and dynamic state that changes over time. At each time step, the vending machine's state is assumed to move (or remain stationary) within a fixed 2D space defined by temperature and gender ratio at a given location, following a state transition probability.

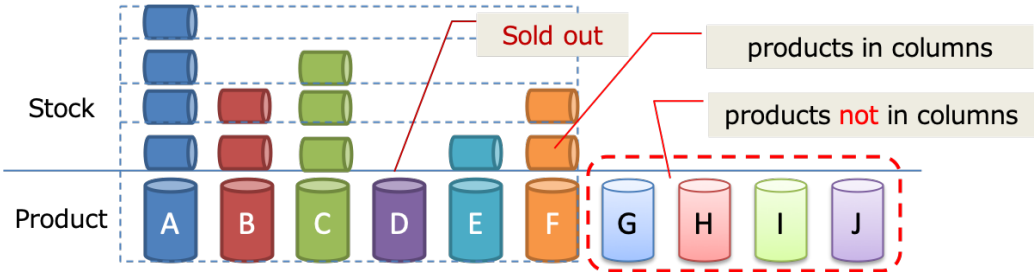

Figure 6.2: Products and Assortment

An illustration of vending machine columns and products. This figure depicts a case with 10 types of products and 6 columns. When products $\mathsf{A}, \ldots, \mathsf{F}$ are assigned to the 6 columns, these products are considered "available for sale". The remaining 4 products $\mathsf{G}, \mathsf{H}, \mathsf{I}, \mathsf{K}$, which are not assigned to any column, are regarded as "not available for sale". Moreover, even for products that are available for sale, if their stock reaches zero (as in the case of D in the figure), they are considered "sold out". While consumer purchase intentions, as modeled by the utility function, may still exist for products that are either not assigned to a column or sold out, these products cannot actually be purchased.

6.2.4 Selection Probability of Products

In the simulation, the utility value $V_{q_i,s_j,k}$ of product $q_i \in \{\mathsf{A}, \ldots, \mathsf{O}\}$ by the k-th consumer in state s_j is defined as:

$$V_{q_i,s_j,k} = V^0_{q_i} + V^M_{q_i} + \beta^M_{q_i} T_j \quad \text{for male} \tag{6.1}$$

$$V_{q_i,s_j,k} = V^0_{q_i} + V^F_{q_i} + \beta^F_{q_i} T_j \quad \text{for female,} \tag{6.2}$$

where $V^0_{q_i}$ is the gender-independent constant of utility value, $V^M_{q_i}, V^F_{q_i}$ is the gender-dependent constant, $\beta^M_{q_i}, \beta^F_{q_i}$ is the gender-dependent coefficient, and T_j is the temperature parameter in state s_j, such as $high \to 1$, $middle \to 0$, $low \to -1$.

These parameters are decided by characteristics of locations and products. Each product is classified into types of drink (coffee, green tea, etc.) with attribute COLD or HOT, and the feature of each product is reflected in the values of parameters. For example, the utility value for men is higher for coffee, more COLD products are sold as the temperature rises, etc. The parameter $V^0_{q_i}$ is

independent of the location, but $V_{q_i}^M, V_{q_i}^F, \beta_{q_i}^M, \beta_{q_i}^F$ are decided by characteristics of the location. $\beta_{q_i}^M, \beta_{q_i}^F$ are coefficients of temperature's contribution to the utility values. When these values are positive, they indicate that these products become easy to be sold as the temperature rises. The Sample of the parameter values we adopted for the simulation are shown in Table 6.1-6.3. (Note that in these tables, the 5 products $\mathsf{K}, \ldots, \mathsf{O}$ are assigned the same values as $\mathsf{F}, \ldots, \mathsf{J}$. This is based on the assumption that there are two different products within the same category. In later simulations, these will be assigned different values to test the effect of including products that are more strongly influenced by the state.)

Parameter estimation of the multinominal logit model can be done independently of the assortment optimization. In simulations, we use artificial values for parameters determined by the following way. We first consider a variety of real products that are for summer/winter, indoor / outdoor, and male / female. Next we assign values to each parameter that seem reasonable from qualitative point of view.

6.2.5 Transition Probability

The transition probabilities $\delta(s(t+1) \mid s(t))$ are decided by characteristics of locations. However, we assume that a transition of gender ratio and temperature are independent in any location. At outdoor, the transition probability of gender ratio to other states is large so that the variation of the ratio is increased. While at school, the probability of staying in the current state is increased because we consider that the variation is small. The probability in office is adopted an intermediate value. The parameter values for the simulation are shown in Table 6.4, 6.5.

Similarly to parameters in the multinominal logit model, the state transition probabilities should be estimated from empirical data. However, we here assume that the probabilities are already known. In this study, we aim to show the proposed

Table 6.1: Parameters of utility value: office

These are the utility value parameters for each product. They reflect the tendency for products to be purchased or not purchased depending on factors such as gender and temperature. For example, G (soda, COLD) is more likely to be purchased at higher temperatures ($\beta > 0$), but this tendency is stronger for men than for women ($\beta^M > \beta^F$).
The "office" location has been set with parameters that suggest a more neutral influence of gender and temperature compared to other locations.

item	type	COLD / HOT	V^0	V^M	V^F	β^M	β^F
A	coffee	COLD	1.0	1.0	0.5	0.0	0.5
B	coffee	HOT	0.5	0.5	0.5	-0.5	-0.5
C	café au lait	COLD	0.0	-0.5	0.5	0.0	-0.5
D	green tea	COLD	1.0	1.0	1.0	0.0	0.0
E	tea	COLD	0.5	0.5	1.0	0.0	0.5
F	enegy drink	COLD	-0.5	0.0	-1.0	0.0	0.0
G	soda	COLD	0.5	0.5	0.0	1.0	0.5
H	mineral water	COLD	1.0	0.0	0.0	0.0	0.0
I	tea	HOT	-0.5	-1.0	0.5	-1.0	-0.5
J	sports drink	COLD	0.5	0.5	0.0	0.0	0.5
K	enegy drink (2)	COLD	-0.5	0.0	-1.0	0.0	0.0
L	soda (2)	COLD	0.5	0.5	0.0	1.0	0.5
M	mineral water (2)	COLD	1.0	0.0	0.0	0.0	0.0
N	tea (2)	HOT	-0.5	-1.0	0.5	-1.0	-0.5
O	sports drink (2)	COLD	0.5	0.5	0.0	0.0	0.5

Table 6.2: Parameters of utility value: outdoor

The "outdoor" location is set with parameters that reflect a stronger influence of temperature compared to the "office" location.

item	type	COLD / HOT	V^0	V^M	V^F	β^M	β^F
A	coffee	COLD	1.0	0.5	-0.5	0.5	0.5
B	coffee	HOT	0.5	0.5	-1.0	-0.5	-1.0
C	café au lait	COLD	0.0	-1.0	0.0	0.5	1.0
D	green tea	COLD	1.0	0.5	0.5	0.0	0.5
E	tea	COLD	0.5	0.0	1.0	1.0	0.5
F	enegy drink	COLD	-0.5	-0.5	-1.0	0.0	0.5
G	soda	COLD	0.5	0.5	0.0	1.0	1.0
H	mineral water	COLD	1.0	-0.5	0.0	0.5	0.0
I	tea	HOT	-0.5	-1.0	0.5	-0.5	-1.0
J	sports drink	COLD	0.5	1.0	0.5	0.5	1.0
K	enegy drink (2)	COLD	-0.5	-0.5	-1.0	0.0	0.5
L	soda (2)	COLD	0.5	0.5	0.0	1.0	1.0
M	mineral water (2)	COLD	1.0	-0.5	0.0	0.5	0.0
N	tea (2)	HOT	-0.5	-1.0	0.5	-0.5	-1.0
O	sports drink (2)	COLD	0.5	1.0	0.5	0.5	1.0

Table 6.3: Parameters of utility value: school

The "school" location is set with parameters that result in a higher purchase intention for products aimed at younger consumers compared to the "office" location. Additionally, there are products where the gender influence is reversed compared to the office. For example, for V, J (sports drink, COLD) has a higher utility, while A, B (coffee, COLD / HOT) have lower utilities. Also, for E (tea, COLD), in the office, $\beta^M < \beta^F$, but in the school, $\beta^M > \beta^F$.

item	type	COLD / HOT	V^0	V^M	V^F	β^M	β^F
A	coffee	COLD	1.0	0.0	-0.5	0.0	0.5
B	coffee	HOT	0.5	-1.0	-1.0	-1.0	-1.0
C	café au lait	COLD	0.0	0.0	0.5	-0.5	0.0
D	green tea	COLD	1.0	1.0	1.0	0.0	-0.5
E	tea	COLD	0.5	0.5	1.0	0.5	0.0
F	enegy drink	COLD	-0.5	-0.5	-1.0	0.0	0.0
G	soda	COLD	0.5	0.5	0.0	1.0	0.5
H	mineral water	COLD	1.0	0.0	-0.5	0.0	0.5
I	tea	HOT	-0.5	-1.0	0.5	-0.5	-1.0
J	sports drink	COLD	0.5	1.0	0.5	1.0	0.5
K	enegy drink (2)	COLD	-0.5	-0.5	-1.0	0.0	0.0
L	soda (2)	COLD	0.5	0.5	0.0	1.0	0.5
M	mineral water (2)	COLD	1.0	0.0	-0.5	0.0	0.5
N	tea (2)	HOT	-0.5	-1.0	0.5	-0.5	-1.0
O	sports drink (2)	COLD	0.5	1.0	0.5	1.0	0.5

approach works if all the parameter values are known. If this is not true, then there is no sense to incorporate estimation of unknown factors in the model. Estimating unknown factors during the assort optimization process remains as future work.

6.2.6 Policy

In this simulation, we define the policy set Π includes $M = 8$ policies in Table 6.6. The assortment constraint is the most strict one that allows all of these policies.

As explained so far, a policy is a means of generating the next assortment from the previous one. If the agent had access to all the information, it would theoretically be possible to swap all products and columns at each operation, always maintaining the best possible state at any given time. However, However, as stated in Section 1.3.4, replacing a large number of products at once is often physically and temporally impossible. There are many hidden constraints on the exchange process.

Furthermore, even if a best-selling product is found and the agent were to only set that product in the columns, sales would drastically drop if the environment changes and the product no longer sells. Additionally, products that are not selling well now may become bestsellers in different environments.

Thus, the assortment policy needs to be determined by considering a wide range of factors, including physical constraints and marketing requirements. In this simulation, taking these factors into account, 8 policies were defined with the following principles: (i) A maximum of two products can be replaced at a time. (ii) The general policy is to place products with high sales potential in the columns and remove products with low sales potential. (iii) To avoid biasing the assortment, a policy that randomly selects which products to replace is also included. (iv) If sales are expected, one product can be set in more than one column.

By determining assortments based on these policies, while it may be difficult to immediately achieve the optimal assortment in response to environmental changes or sudden sales fluctuations, the aim is to gradually adjust the assortment over time

Table 6.4: Transition probability of gender ratio

These tables show the transition probabilities for the gender ratio by location. In the "outdoor" location, the probability of transitioning to a different state is higher, while in the "school" location, the probability of staying in the same state is higher. The "office" location has transition probabilities set in between these two extremes.

office		$s(t+1)$		
$\delta(s(t+1) \mid s(t))$		$\{8:2\}$	$\{5:5\}$	$\{2:8\}$
$s(t)$	$\{8:2\}$	0.60	0.30	0.10
	$\{5:5\}$	0.25	0.50	0.25
	$\{2:8\}$	0.15	0.35	0.50

outdoor		$s(t+1)$		
$\delta(s(t+1) \mid s(t))$		$\{8:2\}$	$\{5:5\}$	$\{2:8\}$
$s(t)$	$\{8:2\}$	0.45	0.35	0.20
	$\{5:5\}$	0.30	0.40	0.30
	$\{2:8\}$	0.20	0.30	0.50

school		$s(t+1)$		
$\delta(s(t+1) \mid s(t))$		$\{8:2\}$	$\{5:5\}$	$\{2:8\}$
$s(t)$	$\{8:2\}$	0.80	0.15	0.05
	$\{5:5\}$	0.15	0.70	0.15
	$\{2:8\}$	0.10	0.15	0.75

Table 6.5: Transition probability of temperature

The transition probabilities for temperature are the same across all locations. Regardless of the previous state, the probability of transitioning to the intermediate state $\{middle\}$ is set to be the highest.

$\delta(s(t+1) \mid s(t))$		$s(t+1)$		
		$\{high\}$	$\{middle\}$	$\{low\}$
$s(t)$	$\{high\}$	0.35	0.50	0.15
	$\{middle\}$	0.20	0.60	0.20
	$\{low\}$	0.25	0.45	0.30

to get closer to the appropriate one. In other words, we expect to achieve a robust assortment that can withstand changes in the environment.

6.2.7 Models and Evaluation

For evaluations, we have performed simulations with baselimodels, comparative models, and POMDP models.

Baseline models are showned in Section 5.4. We have calculated the "theoretical upper bound" Eq. (5.18) and the "feasible maximum value" Eq. (5.20).

A comparative model $C_k (k = 1, \ldots, M)$ is a model that selects the same policy π^k at each time step. For example, π^1 for model C_1 is a policy that the agent exchanges one product with the lowest utility value and one product with the highest utility value, and π^7 for C_7 is that the agent exchanges the lowest two products and the highest two products. The same way, C_2 to C_6 are defined with π_2 to π_6. In this study, we have adopted four models: C_0, C_1, C_2, and C_7. In these cases, it is assumed that the agent considers the gender ratio of the vending machine to be constant at $\{5 : 5\}$ in the initial state. Since other policies showed lower performance than that by π^1, π^2, π^7, we have picked up these policies in the presentation of graphs and tables.

Table 6.6: Policy set Π

This table defines specific policies and shows how the assortment, or product exchanges, will be implemented based on each policy. Policy π^0 represents doing nothing (maintaining the current state), π^1 to π^4 are policies for swapping products within the columns and between columns, π^5 and π^6 are policies for setting or removing one product across two columns,and π^7 is a policy for swapping two products at a time (a total of 4 products).

Policy	Detail
π^0	Do nothing.
π^1	Exchange one of products in columns which has the lowest utility value for one of products not in columns which has the highest utility value.
π^2	Exchange one of products in columns which has the lowest utility value for one of products not in columns selected randomly.
π^3	Exchange one of products in columns selected randomly for one of products not in columns which has the highest utility value.
π^4	Exchange one of product in columns selected randomly for one of products not in columns selected randomly.
π^5	Add one column for one of products in columns which has the highest utility value, and remove one of that which has the lowest utility value.
π^6	Reduce one column from the multi-column products has the lowest utility value, and add one of products not in columns which has the highest utility value.
π^7	Exchange two products in columns which have the lowest utility values for two products not in columns products which have the highest utility values.

As POMDP models, we have prepared three models $\mathcal{M}_\tau$ based on how far into the future τ is to be predicted in the total expected reward $E_{t\to t+\tau}(\pi_t^{k_0})$ at each time t by Eq. (5.15). However, as τ increases, the computational complexity becomes enormous. In this study, for computational feasibility, simulations are limited to three cases: $\tau = 1, 2, 3$, and the improvement in accuracy when τ is larger is also evaluated.

Based on the above, in the following sections, simulations will be conducted under the same conditions for the models listed below, followed by analysis of the results and evaluation of their accuracy.

- Upper bound: $E_t^{Upper\ bound}$
- Feasible max: $E_t^{Feasible\ max}$
- Fix action = 0 C_0: Leave the initial assortment unchanged at all.
- Fix action = 1 C_1: (policy: π^1)
- Fix action = 2 C_2: (policy: π^2)
- Fix action = 7 C_7: (policy: π^7)
- Proposed model $\mathcal{M}_1$: $E_{t\to t+1}(\pi_t^{k_0})$ $(\tau = 1)$.
- Proposed model $\mathcal{M}_2$: $E_{t\to t+2}(\pi_t^{k_0})$ $(\tau = 2)$.
- Proposed model $\mathcal{M}_3$: $E_{t\to t+3}(\pi_t^{k_0})$ $(\tau = 3)$.

6.3 Basic case: 10 items, N=100

For the 2 baselines, 4 comparative models and 3 proposed models, 50 simulations were conducted at each of the 3 location. The length of each simulation is 20 steps, the number of consumers is $N = 100$ and the discount rate is $\gamma = 0.9$ in Eq. (5.15). For simplicity, in this case, the number of products is limited to 10 $(\mathsf{A}, \ldots, \mathsf{J})$,and the vending machine is assumed to accommodate only 6 columns.

Examples of simulation results at outdoor are shown in Fig. 6.3

Note: In Fig. 6.3, there are cases where the sales of the proposed models exceed that of the theoretical upper bound at some time steps. This is because the sales of proposed models are calculated by stochastic simulation at each time step while that of the theoretical upper bound is summed up the expected value of sales. Therefore, the proposed models are evaluated by the average values of the total sales in 50 simulations.

Table 6.7, 6.8, and 6.9 show the summary of "sales" and "sold out" in 50 simulations for each model. Here, the amount of "sold out" cases are counted for the number of consumers who wanted to purchase a product but could not because of sold out in columns.

In these tables, improvement efficiency of assortment for models are evaluated by "Achievement rate: Sales(Ave.) / UB", where UB is the theoretical upper bound. This means the rate of how close the expected sales value is to the upper bound. In all locations, these rates of "Feasible max" are almost close to 100%. It means that if the agent knows all the state in the future and can make the best exchange of products based on the information, the expected sales that the agent can obtain is almost close to the upper bound.

In comparative models, the rates of fix action C_0 is around 68-86%, and that of fix action = $1, 2, 7$ are around 90% on each locations. On the other hand, the rates of the proposed models 1 to 3 are over 90% in all locations, especially at office and school the rates are 92-94%. Therefore, we can conclude that the proposed models are effective in improving sales compared to the comparative models. Moreover, when state probabilities of state transitions are small like in school, we observe that the performance increases as the future time steps τ in estimation increases. However, the improvement is not very large. For office and outdoor, $\tau = 1$ seems enough.

Table 6.7: Result summary: Basic case, office

Baseline or Model	Sales (Ave.)	Sales (Std.)	Sold out (Ave.)	Achievement rate: Sales / Sales of UB(*)
Upper bound	77.81*	2.45	-	100.0%
Feasible max	77.81	2.45	-	100.0%
C_0	67.28	5.62	1.22	86.5%
C_1	72.34	6.61	1.49	93.0%
C_2	70.53	6.27	1.31	90.6%
C_7	71.21	6.83	1.43	91.5%
$\mathcal{M}_1$	73.55	6.71	1.52	94.5%
$\mathcal{M}_2$	73.49	6.85	1.52	94.4%
$\mathcal{M}_3$	73.43	6.59	1.48	94.4%

Table 6.8: Result summary: Basic case, outdoor

Baseline or Model	Sales (Ave.)	Sales (Std.)	Sold out (Ave.)	Achievement rate: Sales / Sales of UB(*)
Upper bound	79.24*	3.23	-	100.0%
Feasible max	78.94	3.41	-	99.6%
C_0	54.20	5.37	0.99	68.4%
C_1	70.50	8.73	2.69	89.0%
C_2	67.87	8.91	2.58	85.6%
C_7	70.15	8.16	2.90	88.5%
$\mathcal{M}_1$	72.54	6.73	3.18	91.5%
$\mathcal{M}_2$	72.15	6.19	3.34	91.1%
$\mathcal{M}_3$	72.05	6.30	3.11	90.9%

Table 6.9: Result summary: Basic case, school

Baseline or Model	Sales (Ave.)	Sales (Std.)	Sold out (Ave.)	Achievement rate: Sales / Sales of UB(*)
Upper bound	80.39*	3.19	-	100.0%
Feasible max	80.25	3.25	-	99.8%
C_0	55.30	7.52	2.68	68.8%
C_1	73.63	7.54	5.16	91.6%
C_2	70.82	7.51	4.78	88.1%
C_7	72.53	7.31	5.10	90.2%
$\mathcal{M}_1$	74.40	6.72	4.96	92.5%
$\mathcal{M}_2$	75.50	6.28	5.05	93.9%
$\mathcal{M}_3$	75.51	6.55	4.92	93.9%

6.4 Secondary case: 15 items, N=150

As next simulations, we adopt the parameters as follows: the number of consumers is $N = 150$, the number of vending machine columns is 10, sum of stocks is 200, time steps is $T = 20$ and the discount rate is $\gamma = 0.9$. The number of kinds of products is $n = 15$ ($\mathsf{A}, \mathsf{B}, \mathsf{C}, \ldots, \mathsf{O}$). Other parameters and conditions follow Section. 6.3.

We assume that outdoor is a location where temperature changes easily and school is a location where changes in gender ratio greatly affect selection behavior, and office is assumed as intermediate. In this study, we have added stadium, which has extremely high transition probabilities of gender ratio. That probabilities are shown Table 6.11.

The results of comparative models C_1, C_7 and POMDP models $\mathcal{M}_1$, $\mathcal{M}_2$, $\mathcal{M}_3$ are summarized in the Table 6.12. The simulations have been evaluated by *achievement rate*, that means the rate of proximity to the upper bound: “Sales

Table 6.10: Parameters of utility value: stadium

A stadium is parameterized such that the overall sales for each product are generally higher.

item	type	COLD / HOT	V^0	V^M	V^F	β^M	β^F
A	coffee	COLD	1.0	0.5	-0.5	0.5	0.5
B	coffee	HOT	0.5	0.5	-1.0	-0.5	-1.0
C	café au lait	COLD	0.0	-1.0	0.0	0.5	1.0
D	green tea	COLD	1.0	0.5	0.5	0.0	0.5
E	tea	COLD	0.5	0.0	1.0	1.0	0.5
F	enegy drink	COLD	-0.5	-0.5	-1.0	0.0	0.5
G	soda	COLD	0.5	0.5	0.0	1.0	1.0
H	mineral water	COLD	1.0	-0.5	0.0	0.5	0.0
I	tea	HOT	-0.5	-1.0	0.5	-0.5	-1.0
J	sports drink	COLD	0.5	1.0	0.5	0.5	1.0
K	enegy drink (2)	COLD	-0.5	-0.5	-1.0	0.0	0.5
L	soda (2)	COLD	0.5	0.5	0.0	1.0	1.0
M	mineral water (2)	COLD	1.0	-0.5	0.0	0.5	0.0
N	tea (2)	HOT	-0.5	-1.0	0.5	-0.5	-1.0
O	sports drink (2)	COLD	0.5	1.0	0.5	0.5	1.0

Table 6.11: Transition probability of gender ratio: stadium

The transition probabilities for the stadium are set to be more dynamic compared to other locations, allowing for more frequent state transitions.

stadium		$s(t+1)$		
$\delta(s(t+1)\|s(t))$		$\{8:2\}$	$\{5:5\}$	$\{2:8\}$
$s(t)$	$\{8:2\}$	0.10	0.30	0.60
	$\{5:5\}$	0.30	0.10	0.60
	$\{2:8\}$	0.60	0.30	0.10

/ Sales of UB", where UB is the theoretical upper bound in (5.18), that means the ratio of how close the expected sales value is to be upper bound. For all locations, the POMDP models perform few points higher achievement rate than the comparative models. In particular, stadium shows clearly $C_1, C_7 < \mathcal{M}_1 < \mathcal{M}_2, \mathcal{M}_3$, and sales increased when the future is predicted. On the other hand, other locations show $C_1, C_7 < \mathcal{M}_1, \mathcal{M}_2, \mathcal{M}_3$, but the effect of future prediction is not so much.

6.5 In case when the initial stock is biased

Next, we have verified whether the model can respond to changes in the condition. Specifically, a simulation has been performed for the case that the initial inventory is biased to 100 units each for two products (A and B).

The example of simulation result is shown in Fig.6.5. When t is small, the assortment does not meet consumer's demand, so sales of all models are low. Here, the POMDP models have a fast increase in sales, and can achieve sales close to UB at an earlier time step.

The summary of results are shown in the Table 6.13. Compared to the comparative models, the POMDP models have increased the achievement rate by more

than 10 points at office and about few points at outdoor and school.

6.6 In case the demand for the products varies significantly

As another case, we have simulated the case where the popularities of products are biased and these demands are greatly different. In this simulation, we varied the utility parameter significantly. In the Basic case, the parameters of the utility value $V^0, V_M, V^M, \beta^M, \beta^F$ are in the range from -1 to 1, but in this simulation, the values are set from -4 to 4. So, this simulation is conducted in which the utility value of a product changes greatly depending on the location, temperature, and gender ratio, and this affects sales.

The example of simulation result is shown in Fig.6.6 When the assortment does not match the demand for the product, the sales drop significantly, but the POMDP model reduces the drop.

The summary of results are shown in the Table 6.14. Compared to the comparative models, the POMDP models have increased the ac rate by more than 10-15 points at office and outdoor, and 3-5 points at school.

6.7 In case of alternative selection allowed

We examined the alternative selection for products when there is a sellout. So far, consumers have assumed that if their first selection was already sold out, they wouldn't purchase it. As a new simulation, if the first-selection product is sold out, a consumer will reselect a product with a high utility value other than the first-selection (second-selection product), and if it is in stock, he/she will purchase it.

Table 6.19 shows the results when alternative selection is allowed under the same conditions as before. In all locations, sales are higher than expected value for UB (not assuming second selection). The comparative models and the POMDP

models have almost the same value. We introduced the *first-selection ratio* as a new indicator. Even if the sales are the same, the higher the first-selection ratio, the higher the consumer satisfaction. According to this rate, the POMDP model is 1-2 points higher than the comparative model.

Table 6.21 shows the results when the initial stock is reduced ($200 \rightarrow 60$) to make it easier for sold out to occur. Sales of the POMDP models have increased by 2-3 points compared to the comparative models. And the first-selection ratio has increased by 2-3 points, which shows the effect of future prediction.

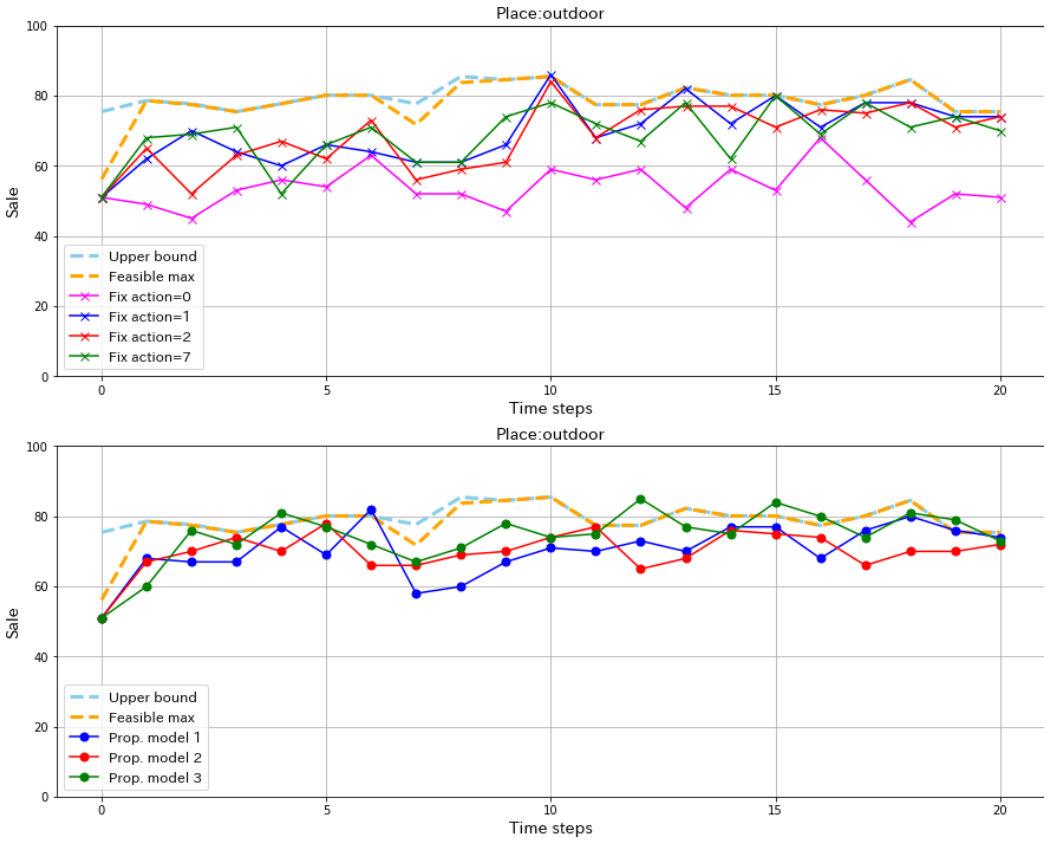

Figure 6.3: Example of Simulation: Basic case, outdoor

The horizontal axis represents time steps, and the vertical axis represents sales. Each model was simulated 50 times, and the average sales for each time step is plotted. In the top figure, Baseline and Comparative models are compared, while in the bottom figure, Baseline and Proposed models are compared. Among the Comparative models, all except C_0 exhibit similar trends. In the Proposed models, $\mathcal{M}_3$ shows behavior relatively close to the Baseline. There are certain days when the model revenues exceed the Baseline. This phenomenon is discussed further in the main text.

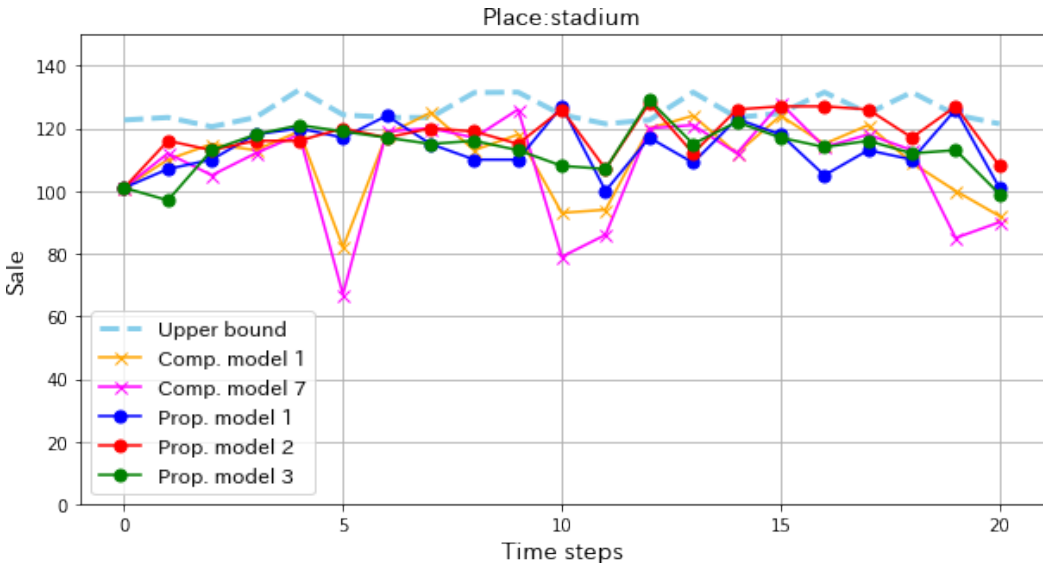

Figure 6.4: Example of Simulation in Stadium

Due to the high transition probabilities, the Comparative model struggles to adapt to state changes, resulting in significantly lower sales on certain days. In contrast, the Proposed model maintains relatively stable sales even on such days.

Table 6.12: Summary in basic case

Location	Sales of UB	Achievement rate: Sales / Sales of UB				
		C_1	C_7	$\mathcal{M}_1$	$\mathcal{M}_2$	$\mathcal{M}_3$
office	122.47	92.2%	92.7%	93.5%	93.2%	93.4%
outdoor	125.35	90.0%	90.6%	92.1%	92.6%	92.9%
school	125.81	92.6%	91.7%	93.3%	93.2%	93.5%
stadium	126.26	87.2%	85.5%	89.1%	91.3%	91.1%

Table 6.13: summary in case of biased stock at the initial

Location	Sales of UB	Achievement rate: Sales / Sales of UB				
		C_1	C_7	$\mathcal{M}_1$	$\mathcal{M}_2$	$\mathcal{M}_3$
office	122.47	63.1%	74.1%	87.2%	87.0%	87.5%
outdoor	125.35	71.4%	84.0%	87.0%	87.3%	87.0%
school	125.81	79.3%	85.7%	87.5%	87.0%	87.7%

Table 6.14: Summary in case the demand for the products varied significantly

Location	Sales of UB	Achievement rate: Sales / Sales of UB				
		C_1	C_7	$\mathcal{M}_1$	$\mathcal{M}_2$	$\mathcal{M}_3$
office	128.49	70.1%	69.2%	78.5%	78.9%	78.5%
outdoor	135.67	56.0%	55.3%	69.6%	69.8%	69.2%
school	129.34	75.0%	73.3%	78.7%	78.7%	78.8%

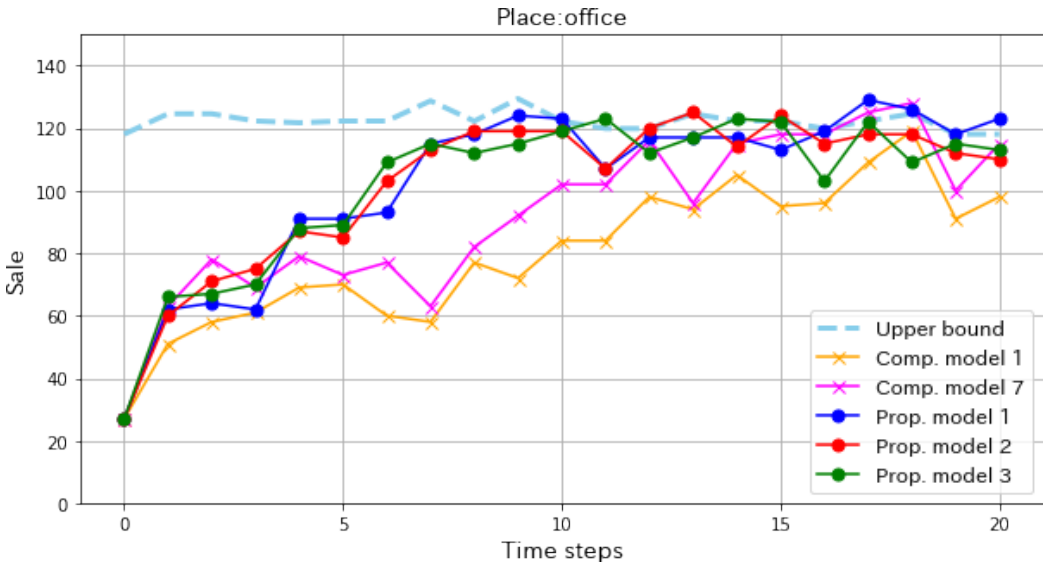

Figure 6.5: Example of Simulation: The initial stock is biased

At earlier stages when t is small, none of the models achieve high sales, but the Proposed model adapts to the environment more quickly and is able to increase sales.

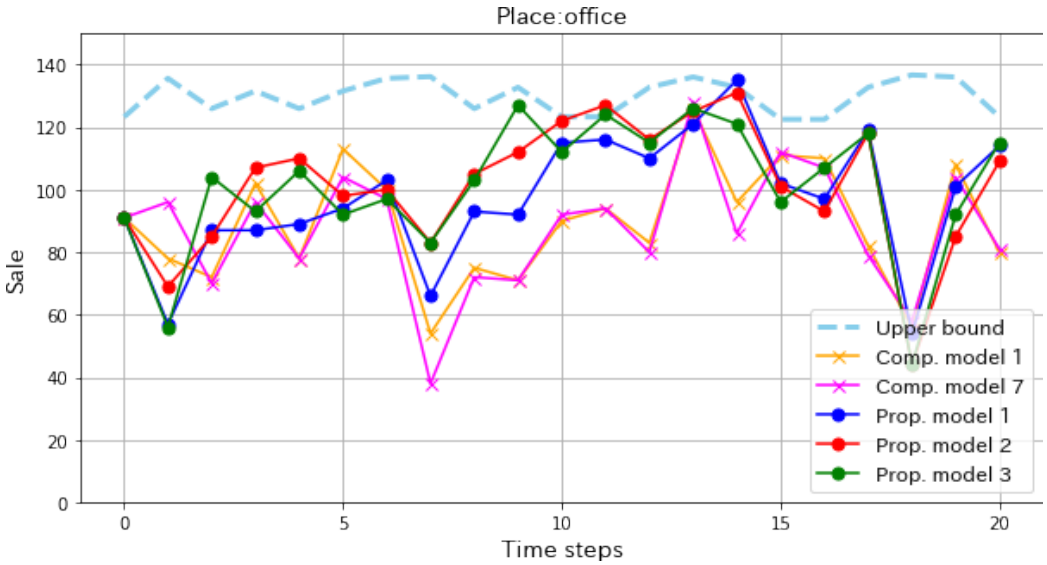

Figure 6.6: Example of Simulation: The products varies significantly

When the assortment does not match the demand for the products, sales drop significantly, but in the Proposed model, the decline is less pronounced.

Table 6.15: Result summary of office
6.19

Baseline or Model	Sales (Ave.)	Sales (Std.)	Sold out (Ave.)	Achievement rate: Sales / Sales of UB(*)
Upper bound	122.47*	3.54	-	100.0%
C_0	107.26	6.06	3.19	87.6%
C_1	112.93	9.22	3.41	92.2%
C_2	110.17	8.45	3.26	90.0%
C_7	113.58	9.37	3.46	92.7%
$\mathcal{M}_1$	114.47	8.10	3.28	93.5%
$\mathcal{M}_2$	114.16	8.56	3.47	93.2%
$\mathcal{M}_3$	114.38	7.51	3.31	93.4%

Table 6.16: Result summary of outdoor

Baseline or Model	Sales (Ave.)	Sales (Std.)	Sold out (Ave.)	Achievement rate: Sales / Sales of UB(*)
Upper bound	123.4*	3.79	-	100.0%
C_0	96.89	6.50	3.63	77.3%
C_1	112.78	11.41	5.16	90.0%
C_2	109.85	11.99	5.22	87.6%
C_7	113.56	10.84	5.27	90.6%
$\mathcal{M}_1$	115.42	8.25	6.38	92.1%
$\mathcal{M}_2$	116.08	8.81	6.22	92.6%
$\mathcal{M}_3$	116.41	8.28	6.00	92.9%

Table 6.17: Result summary of school

Baseline or Model	Sales (Ave.)	Sales (Std.)	Sold out (Ave.)	Achievement rate: Sales / Sales of UB(*)
Upper bound	125.81*	3.88	-	100.0%
C_0	96.66	7.09	6.13	76.8%
C_1	116.52	9.82	8.69	92.6%
C_2	111.89	10.65	8.07	88.9%
C_7	115.31	10.81	8.57	91.7%
$\mathcal{M}_1$	117.43	8.93	8.69	93.3%
$\mathcal{M}_2$	117.30	9.34	8.69	93.2%
$\mathcal{M}_3$	117.67	8.28	8.73	93.5%

Table 6.18: Parameters of utility value: office, varies significantly

Compared to Table 6.1 and others, the range of values is set larger to amplify the impact of state changes on sales.

item	type	COLD / HOT	V^0	V^M	V^F	β^M	β^F
A	coffee	COLD	2.0	4.0	0.5	-2.0	0.5
B	coffee	HOT	0.5	0.5	0.5	-0.5	-0.5
C	café au lait	COLD	0.0	-0.5	0.5	0.0	-0.5
D	green tea	COLD	1.0	1.0	1.0	0.0	0.0
E	tea	COLD	2.0	0.5	2.0	0.0	2.0
F	enegy drink	COLD	-2.0	0.0	-2.0	0.0	-1.0
G	soda	COLD	0.5	0.5	0.0	1.0	0.5
H	mineral water	COLD	1.0	0.0	0.0	0.0	0.0
I	tea	HOT	-4.0	2.0	0.5	-2.0	-0.5
J	sports drink	COLD	0.5	0.5	0.0	0.0	0.5
K	enegy drink (2)	COLD	-2.0	0.0	-4.0	0.0	1.0
L	soda (2)	COLD	0.5	0.5	0.0	1.0	0.5
M	mineral water (2)	COLD	2.0	1.0	0.0	2.0	0.0
N	tea (2)	HOT	-4.0	-2.0	0.5	-2.0	-0.5
O	sports drink (2)	COLD	2.0	0.5	1.0	0.0	4.0

Table 6.19: Summary in case of alternative selection

Location	Sales of UB	Achievement rate: Sales / Sales of UB				
		C_1	C_7	$\mathcal{M}_1$	$\mathcal{M}_2$	$\mathcal{M}_3$
office	122.47	112.5%	112.4%	112.9%	113.1%	113.1%
outdoor	125.35	109.2%	109.5%	109.8%	110.1%	110.1%
school	125.81	108.7%	108.7%	109.1%	109.4%	109.5%

Table 6.20: First-selection ratio in case of alternative selection

Location	First-selection ratio				
	C_1	C_7	$\mathcal{M}_1$	$\mathcal{M}_2$	$\mathcal{M}_3$
office	71.9%	71.9%	72.4%	72.7%	72.5%
outdoor	71.5%	71.6%	73.3%	73.5%	73.3%
school	72.2%	72.3%	73.4%	74.1%	73.4%

Table 6.21: Summary in case of alternative selection in less stock

Location	Sales of UB	Achievement rate: Sales / Sales of UB				
		C_1	C_7	$\mathcal{M}_1$	$\mathcal{M}_2$	$\mathcal{M}_3$
office	89.15	116.7%	114.8%	117.3%	117.0%	116.9%
outdoor	93.52	110.2%	109.1%	112.1%	112.5%	113.4%
school	94.95	109.7%	108.4%	110.8%	110.7%	111.2%

Table 6.22: First selection ratio in case of alternative selection in less stock

Location	First-selection ratio				
	C_1	C_7	$\mathcal{M}_1$	$\mathcal{M}_2$	$\mathcal{M}_3$
office	45.2%	44.0%	45.2%	45.5%	45.5%
outdoor	45.0%	44.2%	46.4%	46.6%	46.8%
school	46.3%	45.9%	47.1%	47.5%	47.5%

Chapter 7

Discussion

In the previous chapters, we proposed a method for product optimization in vending machines and analyzed and evaluated the results through simulation. In this chapter, we will discuss the accuracy evaluation of the proposed method, as well as the assessment of the proposed model and potential areas for improvement.

7.1 Evaluation of the Proposed model

In Chapter 6, the POMDP models slightly outperform the comparative models in terms of sales across various conditions. This suggests that sales can be enhanced by adopting a policy that accounts for future state transitions from the current state, rather than relying on a simple, uniform policy. This effect is particularly notable in locations such as outdoor and stadium environments, where sales fluctuate significantly based on the state (see Fig. 6.3, Fig. 6.4, and Table 6.12). The results also demonstrate that when the current assortment is mismatched with the location or state — such as when the initial inventory is skewed—faster adaptation to the appropriate assortment can be achieved.

However, the impact of predicting future states is not as significant, with noticeable effects only in the stadium scenario (Table 6.12). We believe that additional validation is required for situations where the transition probability is higher or the utility value has a substantial impact.

A limitation of this simulation is that the agent is aware of various parameters, such as transition rates and utility values, and the number of possible states is finite and small. As a result, the agent can quickly identify the optimal assortment for each state, and the corresponding actions can be executed within short time steps. Therefore, it is believed that even if future states are predicted, they will not significantly contribute to an increase in sales.

In the simulations where consumers are offered alternative selection, the POMDP models have been able to achieve higher sales (Table 6.19 and Table 6.21). We evaluated the first-selection rate as an index of consumer satisfaction, and it has been higher than the comparative models. On the other hand, the effect of predicting more future states was not significant.

7.2 Comparison with Existing Methods

In this study, we propose a method based on POMDP while employing an approach that differs from existing methods. When compared with general POMDP solutions such as Value Iteration, Policy Iteration, Monte Carlo methods, and POMCP (Section 3.4.2), we consider that our method exhibits the following characteristics:

- The search time into the future is finite; hence, optimization is not guaranteed.
- It does not involve any learning. Instead, at each time step, it estimates the future based solely on the observations obtained at that moment to search for an optimal policy.
- It is capable of flexibly handling parameters such as transition probabilities and consumers' utility functions; however, in this study, these parameters are fixed.
- The computational cost is not particularly high when the number of states is small, but it becomes enormous as the combinations of states increase.

Overall, we believe that our method offers a realistic solution at a relatively low computational cost for problems of moderate size. Moreover, the advantage of our approach is that it can build a flexible model structure independent of various parameters and the environment ' s structure.

On the other hand, compared to existing AOP methods (Section 3.1.1), the proposed method has the following characteristics:

- It does not guarantee an exact optimal solution.
- It can flexibly adapt to and track changes in the environment.
- It allows for the free configuration of consumer product selection behavior; however, somewhat arbitrary assumptions were adopted in the current simulation.
- It does not require past learning data for model construction.
- Its computational cost is relatively high.

Compared to existing methods, the proposed method is particularly notable for its focus on maximizing expected future sales. As discussed in Chapter 1, agents' visits are constrained by time and labor, and there are limitations on the number and variety of products that can be exchanged in a single product assortment exchange. Even if the optimal product assortment could be determined using mathematical methods, it would be ineffective if it could not be practically implemented. In the proposed method, despite these constraints, a forward-looking approach is employed to gradually converge toward an optimal product assortment through multiple product assortment exchanges. Indeed, the simulation results presented in the previous chapter demonstrate that even suboptimal initial assortments gradually adapt over time.

Future work should involve conducting similar simulations with existing methods to verify whether the observed differences in effectiveness can be detected.

Moreover, the influence of the agent's actions on consumer behavior is limited. Factors such as the impact of past assortments on current product selection (e.g., consumers avoiding purchases due to frequent stockouts) are not incorporated into the model. It is important to assess the impact of these assumptions on the simulation results and to determine whether they introduce significant discrepancies from real-world scenarios.

For comparison with real-world problems, the following validation methods are considered: To validate the model against real-world scenarios, the following methods are proposed:

- Performing simulation-based validation using real-world POS data.
- Implementing the proposed method in real-world assortment operations and comparing its performance with existing methods or the agents own assortment policies through A/B testing.Applying the proposed method to actual assortment operations and comparing it with existing methods or the agent's own assortment policies using A/B testing.

Through these validations, it is necessary to further evaluate the effectiveness and practicality of the proposed method. These validation steps will be essential for further evaluating the effectiveness and practical applicability of the proposed method.

7.3 Improvements of the Proposed Model

If the proposed model works properly, we expect that sales approaches the upper bound. One of possible improvements in the proposed model is to increase τ in Eq. (5.15), that is, the future time steps for summing up expected rewards. In this study, we have used $\tau = 1, 2, 3$, but we expect that sales approaches to the upper bound by increasing τ to $4, 5, \ldots$. However, as τ increases, the number of states that must be calculated increases exponentially. By this reason, we could

not try to use larger numbers for τ in the simulation. There are other ways for the improvements, such as increase in the types and patterns of policies.

One of the challenges of the proposed model is the effect of τ in Eq. (5.15), where τ represents the future time steps over which the expected rewards are summed. In this study, we conducted simulations for the cases $\tau = 1, 2, 3$. Although we anticipated that increasing τ would lead to sales approaching the upper bound, this was not demonstrated in the current study. As a future research direction, it is necessary first to theoretically examine the effect of increasing τ, and subsequently, to reexamine the simulation results when τ is increased. However, since the number of states that must be computed increases exponentially with τ, innovative simulation techniques will be required. Other potential improvement areas include verifying the simulation results for an increased variety and patterns of policies.

Next we consider conditions in which the proposed model performs more effectively. The conditions may include the case that the numbers of products and columns are large enough, since when the numbers are small, the effect of future estimation reduces because the assortment reaches the optimal one immediately. When the absence of IoT devices and the current sales data cannot be obtained in real time, it is necessary to estimate the current state based on the limited data and to make accurate plans for stock replenishment and assortment exchanges. In this case, the proposed methods are reasonable and effective. When the environment and the consumer preferences of vending machines frequently change, methods based on demand forecasting from past history on sales can not keep up with the changes. Since the proposed method works adaptively to the current state, it works well even in such cases.

On the other hand, one of less effective cases is that the optimal assortment does not differ significantly on the environment and the consumer preferences. In such a case, the calculation of future expected reward does not necessarily work well.

As mentioned in the previous section, another issue is that the proposed model

assumes the agent has detailed information about consumer utility values and vending machine state transition probabilities, even though the state is only partially observable. The information may be unknown in actual situations and has to be estimated through past history of observations. Incorporating this factor into the model remains as future work.

What we have shown in this study is that there is a case in which the proposed method outperforms simple policies. Of course, the results will change when the parameter values are changed. From the above discussion, however, we can claim the following properties hold. Compared to simple policies,

- if the diversity of products increases, then the POMDP-based method works well;
- if the volatility of state change increases, then the POMDP-based method works well.

Chapter 8

Conclusion

8.1 Summary

This research investigates decision-making methods under uncertain information, focusing on assortment optimization for vending machines. The study addresses the unique challenges of making assortment decisions under constraints such as limited shelf stock, infrequent restocking opportunities, and uncertain consumer preferences.

Key contributions include:

Problem Formulation The vending machine assortment optimization problem is modeled as a decision-making process under uncertainty, using a Partially Observable Markov Decision Process (POMDP) framework. This formulation incorporates constraints and variability in vending machine states, consumer preferences, and sales patterns.

Proposed Method A POMDP-based method is proposed, allowing optimal assortment policies to be determined by observing consumer behavior, sales data, and the current state of the vending machine. The model emphasizes adaptability to changes in the environment and consumer demand.

Simulations and Results Numerical simulations are conducted to evaluate the effectiveness of the proposed model compared to baseline methods. Results show that the POMDP-based method achieves up to 2-3 points (in archievement rates to the theoretical upper bound) compared to comparative methods sales and adapts more effectively to changes in consumer preferences, particularly in dynamic and uncertain environments.

Implications The research highlights conditions where the POMDP model excels, such as environments with high state transition variability or significant product diversity.

8.2 Future work

There are several directions for future work. First, simulations under identical conditions using existing methods such as PODMP and AOP need to be conducted, and a quantitative comparison and evaluation with the proposed method is necessary. Next, it is important to evaluate the effects of simulations under various conditions, such as different numbers of products, columns, stock levels, and consumers. Additionally, to better reflect real-world assortment problems, the model should be extended to handle cases where the agent has limited information.

Future work should also involve conducting simulations under more realistic conditions. This includes developing methods where the agent determines actions by estimating unknown transition probabilities and utility values, as well as performing simulations with more diverse states and observational data.

Additionally, future research topics include the introduction of models that consider inventory and ordering conditions, as well as refining consumer choice models using sales data from other vending machines and stores. Furthermore, it is necessary to explore the application of the model to retail environments, particularly supermarkets and bookstores, which differ from vending machines. These research efforts are also expected to yield promising results.

Bibliography

[1] Mykel J. Kochenderfer. *Decision making under uncertainty : theory and application*. The MIT Press, 2015.

[2] Japan Vending Machine Manufacturers Association. Vending machine distribution and annual sales amount 2023 (reiwa 5 edition) (in japanese). https://www.jvma.or.jp/information/fukyu2023.pdf, 2024.

[3] Hiroshi Nunokawa and Kiwamu Sato. The Analysis of the Works of Route Man in Beverage Vending Machine Business (in Japanese). *Transactions of Japan Society of Kansei Engineering*, 15(4):471–478, 2016.

[4] Route sales job: Yuka co., ltd. for vending machine installation (in japanese). https://www.yukanet.co.jp/yuka_hp/c_jobs/j_work01.html. (Accessed on 2024/12/01).

[5] Stephen A. Smith and Narendra Agrawal. Management of Multi-Item Retail Inventory Systems with Demand Substitution. *Operations Research*, 48(1):50–64, 2000.

[6] Paat Rusmevichientong, Zuo-Jun Max Shen, and David B Shmoys. Dynamic Assortment Optimization with a Multinomial Logit Choice Model and Capacity Constraint. Technical report, 2009.

[7] Vishal Gaur and Dorothée Honhon. Assortment planning and inventory decisions under a locational choice model. *Management Science*, 52(10):1528–1543, 2006.

[8] Rebecca Chan, Zhaolin Li, and Dmytro Matsypura. Assortment optimisation problem: A distribution-free approach. *Omega*, jun 2019.

[9] A Gürhan Kök and Marshall L Fisher. Demand Estimation and Assortment Optimization Under Substitution: Methodology and Application. *Operations Research*, 55(6):1001–1021, 2007.

[10] Kevin J. Lancaster. A New Approach to Consumer Theory. *The Journal of Political Economy*, 74(2):132–157, 1966.

[11] Daniel L. McFadden. Conditional Logit Analysis of Qualitative Choice Behavior. *Frontiers in Econometrics*, 8:105–142, 1973.

[12] H C W L Williams. On the Formation of Travel Demand Models and Economic Evaluation Measures of User Benefit. *Environment and Planning A: Economy and Space*, 9(3):285–344, 1977.

[13] Yuichiro Miyamoto and Mikio Kubo. A case study on inventory and distribution planning for vending machines (in japanese). *Transactions of the Operations Research Society of Japan*, 44(4):378–389, 2001.

[14] Yuichiro Miyamoto and Kubo Mikio. A combinatorial optimization approach to the vending machine restocking problem (in japanese). *Bulletin of Tokyo University of Mercantile Marine. Natural Sciences*, 51:27–33, 2000.

[15] Shiho Ito, Mikio Kubo, Qiyun Shi, and Yuichiro Miyamoto. Vending machine column allocation problem (in japanese). *2000 Annual Operations Research Society of Japan Fall Research Conference*, pages 22–23, 2000.

[16] Toshiko Takeuchi, Hajime Ito, and Junichiro Fukuchi. Optimization of vending machine column allocation problem: Case of demand following a poisson process (in japanese). *The journal of Faculty of Economics, Gakushuin University*, 49:47–52, 2012.

[17] Ravi Anupindi, Maqbool Dada, and Sachin Gupta. Estimation of Consumer Demand with Stock-Out Based Substitution: An Application to Vending Machine Products. *Marketing Science*, 17(4):406–423, nov 1998.

[18] Hanna Grzybowska, Briscoe Kerferd, Charles Gretton, and S. Travis Waller. A simulation-optimisation genetic algorithm approach to product allocation in vending machine systems. *Expert Systems with Applications*, 145, 5 2020.

[19] Koji Watanabe and Motoo Tanaka. An improved algorithms of supplying and releasing for next generation vending machines using parameter-free genetic algorithm (in japanese). *Journal of the Faculty of Human Cultures and Sciences of Fukuyama University*, 14:1–7, 2014.

[20] Tadahiko Sato and Tomoyuki Higuchi. *Marketing in the Big Data Era: Utilizing Bayesian Modeling (in Japanese).* Kodansha, 2013.

[21] Tadahiko Sato. Bayesian modeling for dynamic utilization of outcome data in marketing (in japanese). *Operations Research = Communications of the Operations Research Society of Japan: Management Science*, 55(1):25–30, 01 2010.

[22] Tadahiko Sato and Tomoyuki Higuchi. Analysis of consumer store visit behavior using dynamic individual models (special feature: 75th anniversary of the japanese statistical society, part iii) (in japanese). *Journal of the Japan Statistical Society*, 38(1):1–38, September 2008.

[23] Marina Fujita, Takeshi Kawamoto, and Toshiko Aizono. Development of a product selection prediction method for demand forecasting in stores handling multiple product types (in japanese). *IEICE Technical Report: Communications Society Technical Report*, 112(466):59–64, 03 2013.

[24] Naoki Matsumura, Kiyoshi Izumi, and Kenta Yamada. A marketing simulation of a retail store with the consumer reactions to Outof- shelf based on a

POS data (in Japanese). *Transactions of the Japanese Society for Artificial Intelligence*, 31(2):1–8, 2016.

[25] R Duncan Luce. *Individual choice behavior.* John Wiley, Oxford, England, 1959.

[26] Ossama Elshiewy, Daniel Guhl, and Yasemin Boztug. Multinomial Logit Models in Marketing - From Fundamentals to State-of-the-Art. *Marketing ZFP*, 39(3):32–49, 2017.

[27] Jun S Chen, Xiang-Hui and Dempster, Arthur P and Liu. Weighted finite population sampling to maximize entropy. *Biometrika*, 81(3):457–69, 1994.

[28] Leslie Pack Kaelbling, Michael L. Littman, and Anthony R. Cassandra. Planning and acting in partially observable stochastic domains. *Artificial Intelligence*, 101(1-2):99–134, 1998.

[29] Sebastian Thrun. Monte carlo pomdps. In S. Solla, T. Leen, and K. Müller, editors, *Advances in Neural Information Processing Systems*, volume 12. MIT Press, 1999.

[30] David Silver and Joel Veness. Monte-carlo planning in large pomdps. In J. Lafferty, C. Williams, J. Shawe-Taylor, R. Zemel, and A. Culotta, editors, *Advances in Neural Information Processing Systems*, volume 23. Curran Associates, Inc., 2010.

[31] Alvaro Flores, Gerardo Berbeglia, and Pascal Van Hentenryck. Assortment optimization under the Sequential Multinomial Logit Model. *European Journal of Operational Research*, 273(3):1052–1064, mar 2019.

[32] Yasunari Maeda. Applying Markov Decision Processes to Role-playing Game (in Japanese). 53(6):1608–1616, 2012.

[33] Dorothée Honhon and Sridhar Seshadri. Probabilistic Analysis of Rumor Spreading Time. *ORSA Journal on Computing*, 00(0):1–20, 2017.

[34] Arne K. Strauss, Robert Klein, and Claudius Steinhardt. A review of choice-based revenue management: Theory and methods. *European Journal of Operational Research*, 271(2):375–387, dec 2018.

[35] Xi Chen and Yining Wang. A note on a tight lower bound for capacitated MNL-bandit assortment selection models. *Operations Research Letters*, 46(5):534–537, sep 2018.

[36] Alvaro Flores, Gerardo Berbeglia, and Pascal Van Hentenryck. Assortment optimization under the Sequential Multinomial Logit Model. *European Journal of Operational Research*, 273(3):1052–1064, mar 2019.

[37] Hidetaka Sakai, Hideki Nakajima, Minoru Higashihara, Masashi Yasuda, and Masato Oosumi. Development of a fuzzy sales forecasting system for vending machines. *Computers & Industrial Engineering*, 36(2):427–449, apr 1999.

[38] Katia Campo, Els Gijsbrechts, and Patricia Nisol. The impact of retailer stockouts on whether, how much, and what to buy. *International Journal of Research in Marketing*, 20(3):273–286, 2003.

[39] Emad Saad. Reinforcement learning in partially observable markov decision processes using hybrid probabilistic logic programs, 2010.

[40] Yuichiro Miyamoto, Mikio Kubo, Shiho Itoh, and Kenya Murakami. Algorithms for the Item Assortment Problem: An Application to Vending Machine Products (in Japanese). *Japan Journal of Industrial and Applied Mathematics*, 20(1):87–100, 2003.

[41] Daihan Zhang, Zhenghe Zhong, Chuning Gao, and Rui Chen. Capacitated Assortment Optimization with Pricing under the Paired Combinatorial Logit Model. *IEEE International Conference on Industrial Engineering and Engineering Management*, 2019-Decem:417–421, 2019.

[42] Baris Tan and Selcuk Karabati. Retail inventory management with stock-out based dynamic demand substitution. *International Journal of Production Economics*, 145(1):78–87, 2013.

[43] David W. Pentico. The assortment problem: A survey. *European Journal of Operational Research*, 190(2):295–309, oct 2008.

[44] Guy Shani, Joelle Pineau, and Robert Kaplow. A survey of point-based pomdp solvers. *Autonomous Agents and Multi-Agent Systems*, 27(1):1–51, July 2013.

[45] Hojung Shin, Soohoon Park, Euncheol Lee, and W.C. Benton. A classification of the literature on the planning of substitutable products. *European Journal of Operational Research*, 246(3):686–699, nov 2015.

[46] Yasuhide Kishimoto, Tetsuya Takiguchi, and Yasuo Ariki. Spoken dialogue manager in car navigation system using partially observable markov decision processes with hierarchical reinforcement learning (in japanese). volume 110, pages 121–126, July 2010.

[47] Yoshihide Yamashiro, Atsushi Ueno, and Hedeaki Takeda. A hybrid method of reinforcement learning and genetic algorithm for pomdp environment (in japanese). *Knowledge-Based Systems Research Group*, (47):37–42, March 2000.

[48] Selcuk Karabati, Baris Tan, and Ömer Cem Öztürk. A method for estimating stock-out-based substitution rates by using point-of-sale data. *IIE Transactions*, 41(5):408–420, 2009.

[49] Shipra Agrawal, Vashist Avadhanula, Vineet Goyal, and Assaf Zeevi. A near-optimal exploration-exploitation approach for assortment selection. In *Proceedings of the 2016 ACM Conference on Economics and Computation*, EC '16, pages 599–600, New York, NY, USA, 2016. Association for Computing Machinery.

[50] J.B. Oliveira, M. Jin, R.S. Lima, J.E. Kobza, and J.A.B. Montevechi. The role of simulation and optimization methods in supply chain risk management: Performance and review standpoints. *Simulation Modelling Practice and Theory*, 92:17–44, apr 2019.

[51] Felipe Caro and Jérémie Gallien. Dynamic Assortment with Demand Learning for Seasonal Consumer Goods. *MANAGEMENT SCIENCE*, 53(2):276–292, 2007.

[52] Hojung Shin, Soohoon Park, Euncheol Lee, and W.C. Benton. A classification of the literature on the planning of substitutable products. *European Journal of Operational Research*, 246(3):686–699, 2015.

[53] Xi Chen, Leonard N Stern, Yining Wang, and Yuan Zhou. Dynamic Assortment Optimization with Changing Contextual Information *. *Journal of Machine Learning Research*, 21:1–44, 2020.

[54] Feng-Cheng Lin, Hsin-Wen Yu, Chih-Hao Hsu, and Tzu-Chun Weng. Recommendation system for localized products in vending machines. *Expert Systems with Applications*, 38(8):9129–9138, aug 2011.

[55] Paat Rusmevichientong, Zuo Jun Max Shen, and David B. Shmoys. Dynamic assortment optimization with a multinomial logit choice model and capacity constraint. *Operations Research*, 58(6):1666–1680, 2010.

[56] James M Davis, Huseyin Topaloglu, and David P Williamson. Assortment optimization over time. *Operations Research Letters*, 43:608–611, 2015.

[57] Hajime Kimura and Leslie Pack Kaelbling. Reinforcement learning for pargially observable markov decision processes (in japanese). *Journal of the Japanese Society for Artificial Intelligence*, 12(6):822–830, 11 1997.

[58] Hajime Kimura, Masayuki Yamamura, and Shigenobu Kobayashi. Reinforcement learning in partially observable markov decision processes : A

stochastic gradient method (in japanese). *Journal of the Japanese Society for Artificial Intelligence*, 11(5):761–768, 1996.

[59] Yasuhiro Minami. Dialogue control using partially observable markov decision process (in japanese). *The Journal of the Acoustical Society of Japan*, 67(10):482–487, 2011.

[60] Tsutomu Maeda, Takeshi Kawai, Yoshio Ozawa, Masaaki Hitomi, and Yasuji Honjo. Proposal of a sales forecasting method for cup-type vending machines (general presentation) (in japanese). *Proceedings of the Kinki Chapter of The Society of Heating, Air-Conditioning and Sanitary Engineers of Japan*, 1999:117–120, 2000.

[61] Scott Sanner and Kristian Kersting. Symbolic Dynamic Programming for First-order POMDPs. *Proceedings of the AAAI Conference on Artificial Intelligence*, 24(1 SE - Reasoning about Plans, Processes and Actions):1140–1146, jul 2010.

Acknowledgement

I would like to express my heartfelt gratitude to everyone who supported me during my doctoral research. First and foremost, I deeply appreciate my supervisor, Professor Kunihiko Hiraishi, for his invaluable guidance, encouragement, and expert advice throughout this journey. His mentorship has been crucial in shaping my research and academic growth.

I also want to thank the faculty and staff of Japan Advanced Institute of Science and Technology for providing me with the opportunity and resources to pursue my research.

Most importantly, I extend my deepest thanks to my wife and daughter. Your patience, understanding, and unwavering love during this challenging time have been my greatest source of strength and motivation. I am truly grateful for your support.

Finally, I would like to acknowledge my friends and colleagues for their moral support and encouragement. This dissertation would not have been possible without all of you.

Thank you from the bottom of my heart.

List of Publications

Journal Paper

[1] Gaku Nemoto and Kunihiko Hiraishi, "A POMDP-Based Approach to Assortment Optimization Problem for Vending Machine," *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences*, Vol. E107.A, No.6, 2024, pp.909-918, doi: 10.1587/transfun.2023EAP1036.

International conference (refereed)

[1] Gaku Nemoto and Kunihiko Hiraishi, "Modeling and optimization of item changes in vending machines," *2017 11th Asian Control Conference (ASCC)*, Gold Coast, QLD, Australia, 2017, pp. 19-24,
doi: 10.1109/ASCC.2017.8287096.

[2] Gaku Nemoto and Kunihiko Hiraishi, "Validation of the POMDP-based Model for Assortment Optimization of Vending Machines," *2023 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM)*, Singapore, Singapore, 2023, pp. 1619-1623,
doi: 10.1109/IEEM58616.2023.10406456.

[3] Satoshi Okuda, Gaku Nemoto, Toshiki Mori, Norihiko Ishitani, Kazuhiko Nishimura and Naoshi Uchihira,
"Exploitation Pattern for Machine Learning Systems," *2021 36th International Technical Conference on Circuits/Systems, Computers and Communi-*

cations (ITC-CSCC), Jeju, Korea (South), 2021, pp. 1-4, doi: 10.1109/ITC-CSCC52171.2021.9501464.

International conference (oral, non-refereed)

[1] Gaku Nemoto and Kunihiko Hiraishi, "A POMDP-based Approach to Assortment Optimization Problem for Vending Machine," *2019 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM)*, Macao, China, 2019.