Title

Decision-Focused Learning and its Applications in Operations Management

Author

Liu, Mo

Publication Date

2024

Peer reviewed|Thesis/dissertation

Decision-Focused Learning and its Applications in Operations Management

By

Mo Liu

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Engineering – Industrial Engineering and Operations Research

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Zuo-Jun (Max) Shen, Chair
Assistant Professor Paul Grigas
Professor Terry Taylor

Spring 2024

Decision-Focused Learning and its Applications in Operations Management

Abstract

Decision-Focused Learning and its Applications in Operations Management

by

Mo Liu

Doctor of Philosophy in Engineering – Industrial Engineering and Operations Research

University of California, Berkeley

Professor Zuo-Jun (Max) Shen, Chair

This dissertation studies the data collection and model training problems in decision-focused learning, with applications in operations management. In Chapter 2, we consider a stochastic optimization problem with a linear objective function, where the coefficients in the objective function are treated as labels. The prediction model is built to predict the labels of the samples based on contextual information. Given unlabeled samples, we study how to sequentially select samples for labeling to minimize the number of acquired labels, while ensuring that the decision risk incurred by the prediction model is smaller than a given threshold. This is the first work to address active label acquisition within the predict-then-optimize framework. We demonstrate that, by utilizing the margin structure of the predict-then-optimize framework, our algorithm requires far fewer labeled samples than the naive supervised learning algorithm when achieving the same level of decision risk. In Chapter 3, we extend this idea of the active label acquisition algorithm to the personalized assortment optimization problem. When collecting data to learn customers' preferences for this problem, we seek to evaluate the importance of each data point. To quantify the importance of each customer, the expected marginal contribution to the risk reduction of each customer is defined as the value of one data point. We provide a feature-dependent upper bound for the value of one data point and utilize this upper bound to design a personalized incentive policy for acquiring the feedback of customers in the customer survey process. Both theoretical and numerical analyses show that our personalized incentive policy can reduce label acquisition costs while maintaining the same level of revenue from the downstream assortment optimization problem. In Chapter 4, motivated by the large amount of click data for online retailers, we study how to efficiently leverage the click transition data for the pricing decisions for multiple products. To capture the impact of product availability on the pricing decision, we propose a new dynamic attraction click model based on a Markov chain. This new model allows us to use click data to estimate customers' preferences and determine the optimal prices. To exploit the similarities between products, we leverage the low-rank structure of the transition matrix in the click model and propose efficient offline and online pricing algorithms. The experiments

on the real-world dataset demonstrate the advantage of our click model.

To Gengning, Qingyu, and Yue

Contents

List of Figures

List of Tables

Acknowledgments

I would like to express my profound gratitude to my advisor, Professor Zuo-Jun Max Shen, for his unwavering support throughout my PhD studies. I am incredibly fortunate to have been introduced to the world of research under his guidance. His encouragement and backing have been crucial as I navigated complex research challenges and paper revisions. I would also like to express my heartfelt gratitude to Professor Paul Grigas for his patience and enduring support, which pivoted my research focus towards decision-focused learning. His creativity, passion for teaching, and disciplined writing style have significantly shaped my academic journey. My sincere thanks also go to Professor Junyu Cao for her valuable contributions to my research papers and ongoing support during my doctoral journey. I would also like to thank my committee member, Professor Terry Taylor, for his insightful suggestions on my research and academic job search. I am also deeply grateful to Professor Ilan Adler for his assistance with my job search and teaching. I would also like to acknowledge my co-authors, Heyuan Liu and Meng Qi, for their support in our research papers.

I would like to thank all the faculty members at UC Berkeley who have provided generous support through my job search, including Professors Rhonda Righter, Zeyu Zheng, Anil Aswani, Ying Cui, Chiwei Yan, Candace Yano, Park Sinchaisri, Luyi Yang, and Rajan Udwani. Special thanks to my colleagues at IBM Research during my internship, Markus Ettl, Zack Xue, and Wei Sun, for the enjoyable time we shared.

I am profoundly grateful to my friends at UC Berkeley for their support during the challenging times of the COVID-19 pandemic and throughout my PhD studies. Thank you to Haixiang Zhang, Hyungki Im, Jingxu Xu, Ilgin Dogan, Hansheng Jiang, Mengxin Wang, Yiduo Huang, Shunan Jiang, Tomas Valencia Zuluaga, Tor Nitayanont, Yuhao Ding, Shuo Sun, Shaochong Lin, Meng Li, Haoting Zhang, Yunkai Zhang, Amy Guo, Donghao Ying, Ziyang Liu, and Wyame Benslimane, for the joy and companionship that brightened my days.

Lastly, I extend endless gratitude to my family, whose love and support have been my cornerstone. This dissertation is dedicated to them.

Chapter 1

Introduction

In most operations research and operations management (OR/OM) applications, operational decisions depend on our knowledge or assumptions about the uncertainty. These uncertainties include future product demand, customer service times, effects of prescriptions on patients, consumer purchase choices, and so on. With increased data availability, understanding and predicting these uncertainties has become easier through the power of Machine learning (ML).

ML has demonstrated its great power in predictive analysis, but in practice, when applying ML to OR/OM applications, the primary concern is the expected cost of operational decisions implied by these ML models. Importantly, a predictive model with higher accuracy does not directly translate to a decision with lower expected cost in OR/OM applications. The main reasons are two-fold:

1. Prediction error metrics may not align with the cost of the decisions in OR/OM applications.

2. If the true optimal decisions are already determined, a prediction model with a higher prediction accuracy will still result in the same decisions and expected cost.

The methodology of designing and training ML models that account for decision-making in downstream optimization problems is termed *decision-focused learning*, which has been the focus of this dissertation. Specifically, this dissertation centers around this question:

How do we efficiently use data to ensure that predictive models translate to good decisions in large-scale OR/OM applications?

In this dissertation, when answering the above question, we examine different applications, including the shortest-path problem, click behavior modeling, personalized pricing, assortment optimization, and general linear programming. Different applications have different structures and thus require different designs of prediction models. However, all these applications can be described as a stochastic optimization problem with some unknown parameters, where some contextual information (feature, side information, or covariates) for these unknown parameters is available.

In Chapters 2 and 3, we study the efficient data collection algorithms for decision-making. In Chapter 4, we consider both data collection and model training processes for the pricing problem by utilizing some special structure within the purchase model.

Specifically, in Chapter 2, we develop the first active learning method in the predict-then-optimize framework. Specifically, we develop a learning method that sequentially decides whether to request the "labels" of feature samples from an unlabeled data stream, where the labels correspond to the parameters of an optimization model for decision-making. Our active learning method is the first to be directly informed by the decision error induced by the predicted parameters, which is referred to as the Smart Predict-then-Optimize (SPO) loss. Motivated by the structure of the SPO loss, our algorithm adopts a margin-based criterion utilizing the concept of distance to degeneracy and minimizes a tractable surrogate of the SPO loss on the collected data. In particular, we develop an efficient active learning algorithm with both hard and soft rejection variants, each with theoretical excess risk (i.e., generalization) guarantees. We further derive bounds on the label complexity, which refers to the number of samples whose labels are acquired to achieve a desired small level of SPO risk. Under some natural low-noise conditions, we show that these bounds can be better than the naive supervised learning approach that labels all samples. Furthermore, when using the SPO+ loss function, a specialized surrogate of the SPO loss, we derive a significantly smaller label complexity under separability conditions. We also present numerical evidence showing the practical value of our proposed algorithms in the settings of personalized pricing and the shortest path problem.

In Chapter 3, we extend our active label acquisition algorithm to the assortment optimization problem. Predicting customers' preferences based on their features is crucial for personalized assortment optimization. When building this prediction model, using informative data can significantly increase the expected revenue from personalized assortments. This chapter studies how to sequentially and actively collect informative data to construct this prediction model. We introduce a novel concept, the 'value of one data point,' which evaluates the marginal contribution of acquiring a specific customer's preference to the expected revenue in personalized assortment optimization, given the existing training set. Notably, this value drops to zero once the optimal assortment for this specific customer is determined. To estimate this value and identify important customers for acquiring their preferences, we derive a feature-dependent upper bound. This bound provides significant insights into the importance of each data point for revenue growth. Based on this upper bound, we develop a personalized incentive policy for effectively collecting survey data from customers to obtain their preferences. We provide non-asymptotic guarantees for both the cumulative incentives and the revenue from the final prediction model. Theoretically, we show that our personalized incentive policy requires smaller cumulative incentives than any fixed incentive policy to achieve the same level of revenue. Furthermore, our numerical experiments with real-world and synthetic datasets validate the effectiveness of our personalized incentive algorithms over fixed strategies.

In Chapter 4, we study how to utilize random clicking behaviors of customers to optimize online retailers' pricing strategies, where product availability is constantly changing due to

various factors, including stockouts and the introduction of limited editions. We introduce a new dynamic attraction click model based on a Markov chain, which describes both purchase and click behaviors under dynamic product availability. To address the challenge of high-dimensional click transition data, we propose an efficient data-driven framework for learning customer's transition behaviors by exploiting the similarities in click transition patterns across products. These similar patterns are captured by the low-rank structure of the attraction matrix in our click model. When considering estimation and pricing decisions simultaneously, we demonstrate the effectiveness of a greedy online algorithm and derive a sublinear regret bound under dynamic product availability. Empirical investigations conducted on real-world data have validated the existence of low-rank structure in the attraction matrix, and shown that using click data along with purchase data can significantly reduce the prediction error associated with purchase behaviors, leading to a substantial increase in the anticipated revenue.

In this dissertation, the term 'predictive model' is used interchangeably with 'predictor,' which may function as either a regressor or a classifier, depending on its output. To facilitate readability, each chapter introduces its own set of notations independently. All mathematical proofs are provided in the appendices.

Chapter 2

Active Learning in the Predict-Then-Optimize Framework

2.1 Introduction

In many applications of operations research, decisions are made by solving optimization problems that involve some unknown parameters. Typically, machine learning tools are used to predict these unknown parameters, and then an optimization model is used to generate the decisions based on the predictions. For example, in the shortest path problem, we need to predict the cost of each edge in the network and then find the optimal path to route users. Another example is the personalized pricing problem, where we need to predict the purchase probability of a given customer at each possible price and then decide the optimal price. In this *predict-then-optimize* paradigm, when generating the prediction models, it is natural to consider the final decision error as a loss function to measure the quality of a model instead of standard notions of prediction error. The loss function that directly considers the cost of the decisions induced by the predicted parameters, in contrast to the prediction error of the parameters, is called the *Smart Predict-then-Optimize (SPO)* loss as proposed by Elmachtoub and Grigas (2022). Naturally, prediction models designed based on the SPO loss have the potential to achieve a lower cost with respect to the ultimate decision error.

In general, for a given feature vector x, calculating the SPO loss requires knowing the correct (in hindsight) optimal decision associated with the unknown parameters. However, a full observation of these parameters, also known as a label associated with x, is not always available. For example, we may not observe the cost of all edges in the graph in the shortest path problem. In practice, acquiring the label of one feature vector instance could be costly, and thus acquiring the labels of all feature vectors in a given dataset would be prohibitively expensive and time-consuming. In such settings, it is essential to actively select the samples for which label acquisition is worthwhile.

Algorithms that make decisions about label acquisition lie in the area of *active learning*. The goal of active learning is to learn a good predictor while requesting a small number

of labels of the samples, whereby the labels are requested actively and sequentially from unlabeled samples. Intuitively, if we are very confident about the label of an unlabeled sample based on our current predictor, then we do not have to request the label of it. Active learning is most applicable when the cost of acquiring labels is very expensive. Traditionally in active learning, the selection rules for deciding which samples to acquire labels for are based on measures of prediction error that ignore the cost of the decisions in the downstream optimization problem. Considering the SPO loss in active learning can hopefully reduce the number of labels required while achieving the same cost of decisions, compared to standard active learning methods that only consider measures of prediction error.

Considering active learning in the predict-then-optimize framework can bridge the gap between active learning and operational decisions, but there are two major challenges when designing algorithms to select samples. One is the computational issue due to the non-convexity and non-Lipschitzness of the SPO loss. When one is concerned with minimizing the SPO loss, existing active learning algorithms are computationally intractable. For example, the general importance weighted active learning (IWAL) algorithm proposed by Beygelzimer, Dasgupta, and Langford (2009) is impractical to implement, since calculating the "weights" of samples requires a large enumeration of all pairs of predictors. Other active learning algorithms that are designed for the classification problem cannot be extended to minimize the SPO loss directly. Another challenge is to derive bounds for the label complexity of the algorithms and to demonstrate the advantages over supervised learning. Label complexity refers to the number of labels that must be acquired to ensure that the risk of predictor is not greater than a desired threshold. To demonstrate the savings from active learning, label complexity should be smaller than the sample complexity of supervised learning, when achieving the same risk level with respect to the loss function of interest (in our case SPO). Kääriäinen (2006) shows that, without additional assumptions on the distributions of features and noise, active learning algorithms have the same label complexity as supervised learning. Thus, deriving smaller label complexity for an active learning algorithm under some natural conditions on the noise and feature distributions is a critical but nontrivial challenge.

In this chapter, we develop the first active learning method in the predict-then-optimize framework. We consider the standard setting of a downstream linear optimization problem where the parameters/label correspond to an unknown cost vector that is potentially related to some feature information. Our proposed algorithm, inspired by margin-based algorithms in active learning, uses a measure of "confidence" associated with the cost vector prediction of the current model to decide whether or not to acquire a label for a given feature. Specifically, the label acquisition decision is based on the notion of *distance to degeneracy* introduced by El Balghiti et al. (2022), which precisely measures the distance from the prediction of the current model to the set of cost vectors that have multiple optimal solutions. Intuitively, the further away the prediction is from degeneracy, the more confident we are that the associated decision is actually optimal. Our proposed margin-based active learning (MBAL-SPO) algorithm has two versions depending on the precise rejection criterion: soft rejection and hard rejection. Hard rejection generally has a smaller label complexity, whereas soft rejection is computationally easier. In any case, when building prediction models based on the actively

selected training set, our algorithm will minimize a generic surrogate of the SPO loss over a given hypothesis class. For each version, we demonstrate theoretical guarantees by providing non-asymptotic excess surrogate risk bounds, as well as excess SPO risk bounds, that hold under a natural consistency assumption.

To analyze the label complexity of our proposed algorithm, we define the near-degeneracy function, which characterizes the distribution of optimal predictions near the regions of degeneracy. Based on this definition, we derive upper bounds on the label complexity. We consider a natural low-noise condition, which intuitively says that the distribution of features for a given problem is far enough from degeneracy. Indeed, for most practical problems, the data are expected to be somewhat bounded away from degeneracy. Under these conditions, we show that the label complexity bounds are smaller than those of the standard supervised learning approach. In addition to the results for a general surrogate loss, we also demonstrate improved label complexity results for the SPO+ surrogate loss, proposed by Elmachtoub and Grigas (2022) to account for the downstream problem, when the distribution satisfies a separability condition. We also conduct some numerical experiments on instances of shortest path problems and personalized pricing problems, demonstrating the practical value of our proposed algorithm above the standard supervised learning approach. Our contributions are summarized below.

- We are the first work to consider active learning algorithms in the predict-then-optimize framework. To efficiently acquire labels to train a machine learning model to minimize the decision cost (SPO loss), we propose a margin-based active learning algorithm that utilizes a surrogate loss function.
- We analyze the label complexity and derive non-asymptotic surrogate and SPO risk bounds for our algorithm, under both soft-rejection and hard-rejection settings. Our analysis applies even when the hypothesis class is misspecified, and we demonstrate that our algorithms can still achieve a smaller label complexity than supervised learning. In particular, under some natural consistency assumptions, we develop the following guarantees.
  - In the hard rejection case with general surrogate loss functions, we provide generic bounds on the label complexity and the non-asymptotic surrogate and SPO risks in Theorem 2.4.1.
  - In the hard rejection case with the SPO+ surrogate loss, we provide a much smaller non-asymptotic surrogate (and, correspondingly, SPO) risk bound in Theorem 2.4.2 under a separability condition. This demonstrates the advantage of the SPO+ surrogate loss over general surrogate losses.
  - In the soft rejection case with a general surrogate loss, which is computationally easier, we provide generic bounds on the label complexity and the non-asymptotic surrogate and SPO risks in Theorem 2.4.3.

  - For each case above, we characterize sufficient conditions for which we can specialize the above generic guarantees and demonstrate that the margin-based algorithm achieves sublinear or even finite label complexity. We provide concrete examples of these conditions, and we provide different non-asymptotic bounds in cases where the feasible region of the downstream optimization problem is either a polyhedron or a strongly convex region. In these situations, and under natural low-noise conditions, we demonstrate that our algorithm can achieve much smaller label complexity than the sample complexity of supervised learning.

- We demonstrate the practical value of our algorithm by conducting comprehensive numerical experiments in two settings. One is the personalized pricing problem, and the other is the shortest path problem. Both sets of experiments show that our algorithm achieves a smaller SPO risk than the standard supervised learning algorithm given the same number of acquired labels.

2.1.1 Example: Personalized Pricing Problem

To further illustrate and motivate the integration of active learning into the predict-then-optimize setting, we present the following personalized pricing problem as an example.

Example 2.1.1 (Personalized pricing via customer surveys)**.** *Suppose that a retailer needs to decide the prices of $\mathfrak{J}$ items for each customer, after observing the features (personalized information) of the customers. The feature vector of a generic customer is x, and the purchase probability of that customer for item j is $d_j(p^j)$, which is a function of the price p^j. This purchase probability $d_j(p^j)$ is unknown and corrupted with some noise for each customer. Suppose the price for each item is selected from a candidate list $\{p_1, p_2, ..., p_{\mathcal{I}}\}$, which is sorted in ascending order. Then, the pricing problem can be formulated as*

$$\max_{\mathbf{w}} \quad \mathbb{E}[\sum_{j=1}^{\mathfrak{J}} \sum_{i=1}^{\mathcal{I}} d_j(p_i) p_i w_{i,j} | x] \tag{2.1}$$

$$s.t. \quad \sum_{i=1}^{\mathcal{I}} w_{i,j} = 1, \quad j = 1, 2, ..., \mathfrak{J}, \tag{2.1a}$$

$$\mathbf{A}\mathbf{w} \leq b, \tag{2.1b}$$

$$w_{i,j} \in \{0, 1\}, \quad i = 1, 2, ..., \mathcal{I}, j = 1, 2, ..., \mathfrak{J}. \tag{2.1c}$$

Here, $\mathbf{w}$ encodes the decision variables with indices in the set $\mathcal{I} \times \mathfrak{J}$, where $w_{i,j}$ is a binary variable indicating which price for item j is selected. Namely, $w_{i,j} = 1$ if item j is priced at p_i, and otherwise $w_{i,j} = 0$. The objective (2.1) *is to maximize the expected total revenue of $\mathfrak{J}$ items by offering price p_i for item j. Constraints* (2.1a) *require each item to have one price selected. In constraint* (2.1b)*, $\mathbf{A}$ is a matrix with $\mathfrak{K}$ rows, and b is a vector with $\mathfrak{K}$ dimensions. Each row of constraints* (2.1b) *characterizes one rule for setting prices. For*

example, if the first row of $\mathbf{Aw}$ *is* $w_{i,j} - \sum_{i'=i}^{\mathcal{I}} w_{i',j+1}$ *and the first entry in* b *is zero, then this constraint further requires that if item* j *is priced at* p_i*, then the price for the item* $j+1$ *must be no smaller than* p_i*. For another example, if the second row of* $\mathbf{Aw}$ *is* $\sum_{i'=1}^{i-1} \sum_{j=1}^{\mathfrak{J}} w_{i',j}$*, and the second entry of* b *is 1, then it means that at most one item can be priced below the price* p_i*. Thus, constraints* (2.1b) *can characterize different rules for setting prices for* $\mathfrak{J}$ *items.*

Traditionally, the conditional expectation of revenue $\mathbb{E}[d_j(p_i)p_i|x]$ *must be estimated from the purchasing behavior of the customers. In this example, we consider the possibility that the retailer can give the customers surveys to investigate their purchase probabilities. By analyzing the results of the surveys, the retailer can infer the purchase probability* $d_j(p_i)|x$ *for each price point* p_i *and each item* j *for this customer. Therefore, whenever a survey is conducted, the retailer acquires a noisy estimate of the revenue, denoted by* $d_j(p_i)p_i|x$*, at each price point* p_i *and item* j*.*

In personalized pricing, first, the retailer would like to build a prediction model to predict $\mathbb{E}[d_j(p_i)p_i|x]$ *given the customer's feature vector* x*. Then, given the prediction model, the retailer solves the problem* (2.1) *to obtain the optimal prices. In practice, when evaluating the quality of the prediction results of* $d_j(p_i)p_i|x$*, the retailer cares more about the expected revenue from the optimal prices based on this prediction, rather than the direct prediction error. Therefore, when building the prediction model for* $d_j(p_i)p_i|x$*, retailers are expected to be concerned with minimizing SPO loss, rather than minimizing prediction error.*

One property of (2.1) *is that the objective is linear and can be further written as* $\max_{\mathbf{w}} \sum_{j=1}^{\mathfrak{J}} \sum_{i=1}^{\mathfrak{K}} \mathbb{E}[d_j(p_i)p_i|x]w_{i,j}$*. By the linearity of the objective, the revenue loss induced by the prediction errors can be written in the form of the SPO loss considered in Elmachtoub and Grigas (2022). In general, considering the prediction errors when selecting customers may be inefficient, since smaller prediction errors do not always necessarily lead to smaller revenue losses, because of the properties of the SPO loss examined by Elmachtoub and Grigas (2022).* □

In Example 2.1.1, in practice, there exists a considerable cost to investigate all customers, for example, the labor cost to collect the answers and incentives given to customers to fill out the surveys. Therefore, the retailer would rather intelligently select a limited subset of customers to investigate. This subset of customers should be ideally selected so that the retailer can build a prediction model with small SPO loss, using a small number of surveys.

Active learning is essential to help retailers select representative customers and reduce the number of surveys. Traditional active learning algorithms would select customers to survey based on model prediction errors, which are different from the final revenue of the retailer. On the contrary, when considering the SPO loss, the final revenue is integrated into the learning and survey distribution processes.

2.1.2 Literature Review

In this section, we review existing work in active learning and the predict-then-optimize framework. To the best of our knowledge, our work is the first work to bridge these two

streams.

Active learning. There has been substantial prior work in the area of active learning, focusing essentially exclusively on measures of prediction error. Please refer to Settles (2009) for a comprehensive review of many active learning algorithms. Cohn, Atlas, and Ladner (1994) shows that in the noiseless binary classification problem, active learning can achieve a large improvement in label complexity, compared to supervised learning. It is worth noting that in the general case, Kääriäinen (2006) provides a lower bound of the label complexity which matches supervised learning. Therefore, to demonstrate the advantages of active learning, some further assumptions on the noise and distribution of samples are required. For the agnostic case where the noise is not zero, many algorithms have also been proposed in the past few decades, for example, Hanneke (2007), Dasgupta, Hsu, and Monteleoni (2007),Hanneke (2011), Balcan, Beygelzimer, and Langford (2009), and Balcan, A. Broder, and T. Zhang (2007). These papers focus on binary or multiclass classification problems. Balcan, A. Broder, and T. Zhang (2007) proposed a margin-based active learning algorithm, which is used in the noiseless binary classification problem with a perfect linear separator. Balcan, A. Broder, and T. Zhang (2007) achieves the label complexity $\mathcal{O}(\epsilon^{-2\alpha}\ln(1/\epsilon))$ under uniform distribution, where $\alpha \in (0,1)$ is a parameter defined for the low noise condition and ϵ is the desired error rate. Krishnamurthy et al. (2017) and R. Gao and Saar-Tsechansky (2020) consider cost-sensitive classification problems in active learning, where the misclassification cost depends on the true labels of the sample.

The above active learning algorithms in the classification problem do not extend naturally to real-valued prediction problems. However, the SPO loss is a real-valued function. When considering real-valued loss functions, Castro, Willett, and Nowak (2005) prove convergence rates in the regression problem, and Sugiyama and Nakajima (2009) and Cai, M. Zhang, and Y. Zhang (2016) also consider squared loss as the loss function. Beygelzimer, Dasgupta, and Langford (2009) propose an importance-weighted algorithm (IWAL) that extends disagreement-based methods to real-valued loss functions. However, it is intractable to directly use the IWAL algorithm in the SPO framework. Specifically, it requires solving a non-convex problem at each iteration, which may have to enumerate all pairs of predictor candidates even when the hypothesis set is finite.

Predict-then-optimize framework. In recent years, there has been a growing interest in developing machine learning models that incorporate the downstream optimization problem. For example, Bertsimas and Kallus (2020), Kao, Roy, and Yan (2009), Elmachtoub and Grigas (2022), T. Zhu, Xie, and Sim (2022), Donti, Amos, and Kolter (2017) and Ho and Hanasusanto (2019) propose frameworks that somehow relate the learning problem to the downstream optimization problem. In our work, we consider the Smart Predict-then-Optimize (SPO) framework proposed by Elmachtoub and Grigas (2022). Because the SPO loss function is nonconvex and non-Lipschitz, the computational and statistical properties of the SPO loss in the fully supervised learning setting have been studied in several recent works. Elmachtoub

and Grigas (2022) provide a surrogate loss function called SPO+ and show the consistency of this loss function. Elmachtoub, Liang, and McNellis (2020), Loke, Q. Tang, and Xiao (2022), Demirovic et al. (2020), Demirović et al. (2019), Mandi and Guns (2020), Mandi, Stuckey, Guns, et al. (2020), and B. Tang and Khalil (2022) all develop new applications and computational frameworks for minimizing the SPO loss in various settings. El Balghiti et al. (2022) consider generalization error bounds of the SPO loss function. Ho-Nguyen and Kılınç-Karzan (2022), H. Liu and Grigas (2021), and Hu, Kallus, and Mao (2022) further consider risk bounds of different surrogate loss functions in the SPO setting. There is also a large body of work more broadly in the area of decision-focused learning, which is largely concerened with differentiating through the parameters of the optimization problem, as well as other techniques, for training. See, for example, Amos and Kolter (2017), Wilder, Dilkina, and Tambe (2019), Berthet et al. (2020), and Chung et al. (2022), the survey paper Kotary et al. (2021), and the references therein. Recently there has been growing attention on problems with nonlinear objectives, where estimating the conditional distribution of parameters is often needed; see, for example, Kallus and Mao (2023) and Grigas, Qi, et al. (2021) and Elmachtoub, Lam, et al. (2023).

2.1.3 Organization

The remainder of the chapter is organized as follows. In Section 2.2, we introduce preliminary knowledge on the predict-then-optimize framework and active learning, including the SPO loss function, label complexity, and the SPO+ surrogate loss function. Then, we present our active learning algorithm, margin-based active learning (MBAL-SPO), in Section 2.3. We first present an illustration to motivate the incorporation of the distance to degeneracy in the active learning algorithm in 2.3.1. Next, we analyze the risk bounds and label complexities for both hard and soft rejection in Section 2.4. To demonstrate the strength of our algorithm over supervised learning, we consider natural low-noise conditions and derive sublinear label complexity in Section 2.5. We demonstrate the advantage of using SPO+ as the surrogate loss in some cases by providing a smaller label complexity. We further provide concrete examples of these low-noise conditions. In Section 2.7, we test our algorithm using synthetic data in two problem settings: the shortest path problem and the personalized pricing problem. Lastly, we point out some future research directions in Section 2.8. The omitted proofs, sensitivity analysis of the numerical experiments, and additional numerical results are provided in the Appendices.

2.2 Preliminaries

We first introduce some preliminaries about active learning and the predict-then-optimize framework. In particular, we introduce the SPO loss function, we discuss the goals of active learning in the predict-then-optimize framework, and we review the SPO+ surrogate loss.

2.2.1 Predict-then-Optimize Framework and Active Learning

Let us begin by formally describing the "predict-then-optimize" framework and the "Smart Predict-then-Optimize (SPO)" loss function. We assume that the downstream optimization problem has a linear objective, but the cost vector of the objective, $c \in \mathcal{C} \subseteq \mathbb{R}^d$, is unknown when the problem is solved to make a decision. Instead, we observe a feature vector, $x \in \mathcal{X} \subseteq \mathbb{R}^p$, which provides auxiliary information that can be used to predict the cost vector. The feature space $\mathcal{X}$ and cost vector space $\mathcal{C}$ are assumed to be bounded. We assume there is a fixed but unknown distribution $\mathcal{D}$ over pairs (x, c) living in $\mathcal{X} \times \mathcal{C}$. The marginal distribution of x is denoted by $\mathcal{D}_{\mathcal{X}}$. Let $w \in S$ denote the decision variable of the downstream optimization problem, where the feasible region $S \subseteq \mathbb{R}^d$ is a convex and compact set that is assumed to be fully known to the decision-maker. To avoid trivialities, we also assume throughout that the set S is not a singleton. Given an observed feature vector x, the ultimate goal is to solve the contextual stochastic optimization problem:

$$\min_{w \in S} \mathbb{E}_c[c^T w | x] \;=\; \min_{w \in S} \mathbb{E}[c|x]^T w. \tag{2.2}$$

From the equivalence in (2.2), observe that the downstream optimization problem in the predict-then-optimize framework relies on a prediction (otherwise referred to as estimation) of the conditional expectation $\mathbb{E}_c[c|x]$. Given such a prediction $\hat{c}$, a decision is made by then solving the deterministic version of the downstream optimization problem:

$$P(\hat{c}): \quad \min_{w \in S} \hat{c}^T w. \tag{2.3}$$

For simplicity, we assume $w^* : \mathbb{R}^d \to S$ is an oracle for solving (2.3), whereby $w^*(\hat{c})$ is an optimal solution of $P(\hat{c})$.

Our goal is to learn a cost vector predictor function $h : \mathcal{X} \to \mathbb{R}^d$, so that for any newly observed feature vector x, we first make prediction $h(x)$ and then solve the optimization problem $P(h(x))$ in order to make a decision. This predict-then-optimize paradigm is prevalent in applications of machine learning to problems in operations research. We assume the predictor function h is within a compact hypothesis class $\mathcal{H}$ of functions on $\mathcal{X} \to \mathbb{R}^d$. We say the hypothesis class is well-specified if $\mathbb{E}[c|x] \in \mathcal{H}$. In our analysis, the well-specification is not required. The active learning methods we consider herein rely on using a variant of empirical risk minimization to select $h \in \mathcal{H}$ by minimizing an appropriately defined loss function. Our primary loss function of interest in the predict-then-optimize setting is the SPO loss, introduced by Elmachtoub and Grigas (2022), which characterizes the regret in decision error due to an incorrect prediction and is formally defined as

$$\ell_{\text{SPO}}(\hat{c}, c) := c^T w^*(\hat{c}) - c^T w^*(c),$$

for any cost vector prediction $\hat{c}$ and realized cost vector c. We further define the SPO risk of a prediction function $h \in \mathcal{H}$ as $R_{\text{SPO}}(h) := \mathbb{E}_{(x,c)\sim\mathcal{D}}[\ell_{\text{SPO}}(h(x), c)]$, and the excess risk of h as $R_{\text{SPO}}(h) - \inf_{h' \in \mathcal{H}} R_{\text{SPO}}(h')$. (Throughout, we typically remove the subscript notation

from the expectation operator when it is clear from the context.) Notice that a guarantee on the excess SPO risk implies a guarantee that holds "on average" with respect to x for the contextual stochastic optimization problem (2.2).

As previously described, in many situations acquiring cost vector data may be costly and time-consuming. The aim of active learning is to choose which feature samples x to label sequentially and interactively, in contrast to standard supervised learning which acquires the labels of all the samples before training the model. In the predict-then-optimize setting, acquiring a "label" corresponds to collecting the cost vector data c that corresponds to a given feature vector x. An active learner aims to use a small number of labeled samples to achieve a small prediction error. In the agnostic case, the noise is nonzero and the smallest prediction error is the Bayes risk, which is $R^*_{\text{SPO}} = \inf_{h\in\mathcal{H}} R_{\text{SPO}}(h) > 0$. The goal of an active learning method is to then find a predictor $\hat{h}$ trained on the data with the minimal number of labeled samples, such that $R_{\text{SPO}}(\hat{h}) \leq R^*_{\text{SPO}} + \epsilon$, with high probability and where $\epsilon > 0$ is a given risk error level. The number of labels acquired to achieve this goal is referred to as the label complexity.

2.2.2 Surrogate Loss Functions and SPO+

Due to the potential non-convexity and even non-continuity of the SPO loss, a common approach is to consider surrogate loss functions ℓ that have better computational properties and are still (ideally) aligned with the original SPO loss. In our work, the surrogate loss function $\ell : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}_+$ is assumed to be continuous. The surrogate risk of a predictor $h \in \mathcal{H}$ is denote by $R_\ell(h)$, and the corresponding minimum risk is denoted by $R^*_\ell := \min_{h\in\mathcal{H}} R_\ell(h)$.

As a special case of the surrogate loss function ℓ, Elmachtoub and Grigas (2022) proposed a convex surrogate loss function, called the SPO+ loss, which is defined by

$$\ell_{\text{SPO+}}(\hat{c}, c) := \max_{w\in S} \left\{(c - 2\hat{c})^T w\right\} + 2\hat{c}^T w^*(c) - c^T w^*(c),$$

and is an upper bound on the SPO loss, i.e., $\ell_{\text{SPO}}(\hat{c}, c) \leq \ell_{\text{SPO+}}(\hat{c}, c)$ for any $\hat{c} \in \hat{\mathcal{C}}$ and $c \in \mathcal{C}$. Elmachtoub and Grigas (2022) demonstrate the computational tractability of the SPO+ surrogate loss, conditions for Fisher consistency of the SPO+ risk with respect to the true SPO risk, as well as strong numerical evidence of its good performance with respect to the downstream optimization task. H. Liu and Grigas (2021) further demonstrate sufficient conditions that imply that when the excess surrogate SPO+ risk of a prediction function h is small, the excess true SPO risk of a prediction function h is also small. This property not only holds for the SPO+ loss, but also for other surrogate loss functions, such as the squared ℓ_2 loss (see, for details, Ho-Nguyen and Kılınç-Karzan (2022)). Importantly, the SPO+ loss still accounts for the downstream optimization problem and the structure of the feasible region S, in contrast to losses like the ℓ_2 loss that focus only on prediction error. As will be shown in Theorem 2.4.2, compared to the general surrogate loss functions that satisfy Assumption 2.3.1 in our analysis, the SPO+ loss function achieves a smaller label complexity by utilizing the structure of the downstream optimization problem.

Notations. Let $\|\cdot\|$ on $w \in \mathbb{R}^d$ be a generic norm. Its dual norm is denoted by $\|\cdot\|_*$, which is defined by $\|c\|_* = \max_{w:\|w\|\leq 1} c^T w$. We denote the set of extreme points in the feasible region S by $\mathfrak{S}$, and the diameter of the set $S \subset \mathbb{R}^d$ by $D_S := \sup_{w,w' \in S}\{\|w - w'\|\}$. The "linear optimization gap" of S with respect to cost vector c is defined as $\omega_S(c) := \max_{w\in S}\{c^T w\} - \min_{w\in S}\{c^T w\}$. We further define $\omega_S(\mathcal{C}) := \sup_{c\in\mathcal{C}}\{\omega_S(c)\}$ and $\rho(\mathcal{C}) := \max_{c\in\mathcal{C}}\{\|c\|\}$, where again $\mathcal{C}$ is the domain of possible realizations of cost vectors under the distribution $\mathcal{D}$. We denote the cost vector space of the prediction range by $\hat{\mathcal{C}}$, i.e., $\hat{\mathcal{C}} := \{c \in \mathbb{R}^d : c = h(x), h \in \mathcal{H}, x \in \mathcal{X}\}$. For the surrogate loss function ℓ, we define $\omega_\ell(\hat{\mathcal{C}}, \mathcal{C}) := \sup_{\hat{c}\in\hat{\mathcal{C}}, c\in\mathcal{C}}\{\ell(\hat{c}, c)\}$. We also denote $\rho(\mathcal{C}, \hat{\mathcal{C}}) := \max\{\rho(\mathcal{C}), \rho(\hat{\mathcal{C}})\}$ for the general norm. We use $\mathcal{N}(\mu, \sigma^2)$ to denote the multivariate normal distribution with center μ and covariance matrix σ^2. We use $\mathbb{R}_+$ to denote $[0, +\infty)$. When conducting the asymptotic analysis, we adopt the standard notations $\mathcal{O}(\cdot)$ and $\Omega(\cdot)$. We further use $\tilde{\mathcal{O}}(\cdot)$ to suppress the logarithmic dependence. We use $\mathbb{I}$ to refer to the indicator function, which outputs 1 if the argument is true and 0 otherwise.

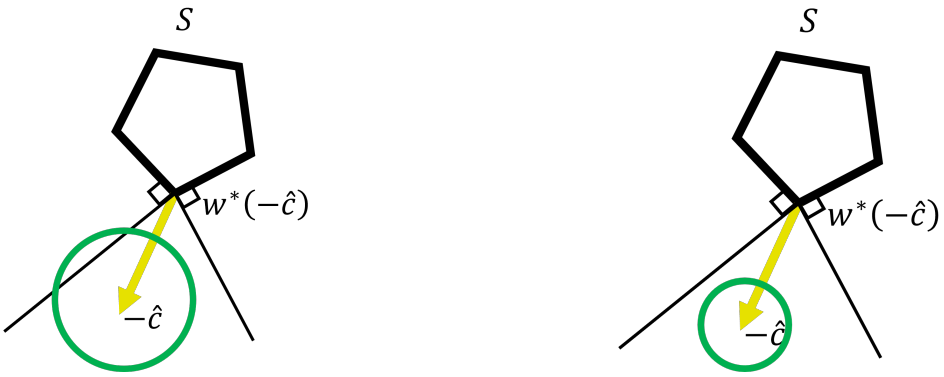
2.3 Margin-Based Algorithm

In this section, we develop and present the margin-based algorithm in the predict-then-optimize framework (MBAL-SPO). We first illustrate and motivate the algorithm in the polyhedral case. Then, we provide some conditions for the noise distribution and surrogate loss functions for our MBAL-SPO.

2.3.1 Illustration and Algorithm

Let us introduce the idea of the margin-based algorithm with the following two examples, which illustrate the value of integrating the SPO loss into active learning. Particularly, given the current training set and predictor, it is very likely that some features will be more informative and thus more valuable to label than others. In general, the "value" of labeling a feature depends on the associated prediction error (Figure 2.1) and the location of the prediction relative to the structure of the feasible region S (Figure 2.2). In Figure 2.1, the feasible region S is polyhedral and the yellow arrow represents $-\hat{h}(x)$. Within this example, for the purpose of illustration, let us assume the hypothesis class is well-specified. Our goal then is to find a good predictor h from the hypothesis class $\mathcal{H}$, such that $h(x)$ is close to $\mathbb{E}[c|x]$. However, because $c|x$ is random, the empirical best predictor $\hat{h}$ in the training set may not exactly equal the true predictor h^*, where $h^*(x) = \mathbb{E}[c|x]$. Given one feature x, the prediction is $\hat{c} = \hat{h}(x)$, the negative of which is shown in Figures 2.1a and 2.1b. Intuitively, when the training set gets larger, the empirical best predictor $\hat{h}$ should get closer to h^*, and $\hat{h}(x)$ should get closer to $\mathbb{E}[c|x]$. Thus, we can construct a confidence region around $\hat{h}(x)$, such that $\mathbb{E}[c|x]$ is within this confidence region with some high probability. Examples of confidence regions for the estimation of $\mathbb{E}[c|x]$ given the current training set are shown in the green circles in Figure 2.1. The optimal solution $w^*(\hat{c})$ is the extreme point indicated in

Figure 2.1, and the normal cone at $w^*(\hat{c})$ illustrates the set of all cost vectors whose optimal solution is also $w^*(\hat{c})$. In addition, those cost vectors that lie on the boundary of the normal cone are the cost vectors that can lead to multiple optimal decisions (they will be defined as degenerate cost vectors in Definition 2.3.1 later). In cases when the confidence region is large (e.g., because the training set is small), as indicated in Figure 2.1a, the green circle intersects with the degenerate cost vectors, which means that some vectors within the confidence region for estimating $\mathbb{E}[c|x]$ could lead to multiple optimal decisions. When the confidence region is smaller (e.g., because the training set is larger), as indicated in Figure 2.1b, the green circle does not intersect with the degenerate cost vectors, which means the optimal decision of $\mathbb{E}[c|x]$ is the same as the optimal decision of $\hat{c} = \hat{h}(x)$, $w^*(\hat{c})$, with high probability. Thus, when the confidence region of $\mathbb{E}[c|x]$ does not intersect with the degenerate cost vectors, the optimal decision based on the current estimated cost vector will lead to the correct optimal decision with high probability, and the SPO loss will be zero. This in turn suggests that the label corresponding to x is not informative (and we do not have to acquire it), when the confidence region centered at the prediction $\hat{h}(x)$ is small enough to not intersect those degenerate cost vectors. Figure 2.2 further shows that considering the SPO loss function



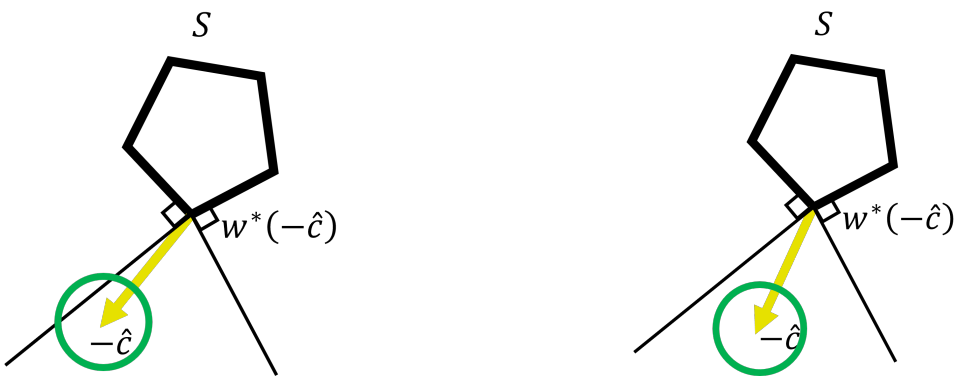
(a) Confidence region is large

(b) Confidence region is small

Figure 2.1: Active learning reduces the label complexity, given the same prediction.

reduces the label complexity when the confidence regions of the cost vector are the same size. In Figure 2.2, both green circles have the same radius but their locations are different. In Figure 2.2a, the confidence region for $\mathbb{E}[c|x]$ is close to the degenerate cost vectors, and thus the cost vectors within the confidence region will lead to multiple optimal decisions. In Figure 2.2b, the confidence region for $\mathbb{E}[c|x]$ is far from the degenerate cost vectors, and therefore acquiring a label for x is less informative, as we are more confident that $\hat{h}(x)$ leads to the correct optimal decision due to the more central location of the confidence region.

The above two examples highlight that the confidence associated with a prediction $\hat{h}(x)$ is crucial to determine whether it is valuable to acquire a true label c associated with x. Furthermore, confidence is related to both the size of the confidence region (which is often



(a) Prediction is close to degenerate cost vectors (b) Prediction is far from degenerate cost vectors

Figure 2.2: The SPO loss function reduces the label complexity, given the same size confidence region.

dictated by the number of labeled samples we have acquired) and the location of the prediction relative to the structure of S. El Balghiti et al. (2022) introduced the notion of "distance to degeneracy," which precisely measures the distance of a prediction $\hat{h}(x)$ to those degenerate cost vectors with multiple optimal solutions and thus provides the correct way to measure confidence about the location of a prediction. In fact, El Balghiti et al. (2022) argue that the distance to degeneracy provides a notion of confidence associated with a prediction that generalizes the notion of "margin" in binary and multiclass classification problems. El Balghiti et al. (2022) use the distance to degeneracy to provide tighter generalization guarantees for the SPO loss and its associated margin loss. In our context, we adopt the distance to degeneracy in order to determine whether or not to acquire labels. It is motivated by our intuition from the previously discussed examples wherein the labels of samples should be more informative if their predicted cost vectors are closer to degeneracy. In turn, we develop a generalization of margin-based active learning algorithms that utilize the distance to degeneracy as a confidence measure to determine those samples whose labels should (or should not) be acquired. Definition 2.3.1 reviews the notion of distance to degeneracy as defined by El Balghiti et al. (2022).

Definition 2.3.1. *(Distance to Degeneracy, El Balghiti et al. (2022)). The set of degenerate cost vector predictions is* $\mathcal{C}^o := \{\hat{c} \in \mathbb{R}^d : P(\hat{c})$ *has multiple optimal solutions*$\}$*. Given a norm* $\|\cdot\|$ *on* $\mathbb{R}^d$*, the distance to degeneracy of the prediction* $\hat{c}$ *is* $\nu_S(\hat{c}) := \inf_{c \in \mathcal{C}^o}\{\|c - \hat{c}\|\}$*.* □

The distance to degeneracy can be easily computed in some special cases, for example, when the feasible region S is strongly convex or in the case of a polyhedral feasible region with known extreme point representations. El Balghiti et al. (2022) provide the exact formulas of the distance to degeneracy function in these two special cases. In particular, in the case of a polyhedral feasible region with extreme points $\{v_j : j = 1, ..., K\}$, that

is, $S = \text{conv}(v_1, \ldots, v_K)$, Theorem 8 of El Balghiti et al. (2022) says that the distance to degeneracy of any vector $c \in \mathbb{R}^d$ satisfies the following equation:

$$\nu_S(c) = \min_{j:v_j \neq w^*(c)} \left\{ \frac{c^T(v_j - w^*(c))}{\|v_j - w^*(c)\|_*} \right\}. \tag{2.4}$$

Theorem 7 of El Balghiti et al. (2022), on the other hand, says that $\nu_S(c) = \|c\|$ whenever S is a strongly convex set. As mentioned, the distance to degeneracy $\nu_S(\hat{c})$ provides a measure of "confidence" regarding the cost vector prediction $\hat{c}$ and its implied decision $w^*(\hat{c})$. This observation motivates us to design a margin-based active learning algorithm, whereby if the distance to degeneracy $\nu_S(\hat{c})$ is greater than some threshold (depending on the number of iterations and samples acquired so far), then we are confident enough to label it using our current model without asking for the true label.

Our margin-based method is proposed in Algorithm 1. The idea of the margin-based algorithm can be explained as follows. At iteration t, we first observe an unlabeled feature vector x_t, which follows distribution $\mathcal{D}_\mathcal{X}$. Given the current predictor h_{t-1}, we calculate the distance to the degeneracy $\nu_S(h_{t-1}(x_t))$ of this unlabeled sample x_t. If the distance to degeneracy $\nu_S(h_{t-1}(x_t))$ is greater than the threshold b_{t-1}, then we reject x_t with some probability $1 - \tilde{p}$. If $\tilde{p} = 0$, this rejection is referred to as a hard rejection; when $\tilde{p} > 0$, this rejection is referred to as a soft rejection. If a soft-rejected sample is not ultimately rejected, we acquire a label (cost vector) c_t associated with x_t and add the sample (x_t, c_t) to the set $\tilde{W}_t$. On the other hand, if $\nu_S(h_{t-1}(x_t)) < b_{t-1}$, then we acquire a label (cost vector) c_t associated with x_t and add the sample (x_t, c_t) to the working training set W_t. At each iteration, we update the predictor h_t by computing the best predictor within a subset of the hypothesis class $H_t \subseteq \mathcal{H}$ that minimizes the empirical surrogate risk measured on the labeled samples. Note that Algorithm 1 maintains two working sets, $\tilde{W}_t$ and W_t, due to the two different types of labeling criteria. To ensure that the expectation of empirical loss is equal to the expectation of the true loss, we need to assign weight $\frac{1}{\tilde{p}}$ to the soft-rejection samples in the set $\tilde{W}_t$. It is assumed throughout that the sequence $(x_1, c_1), (x_2, c_2), \ldots$ is an i.i.d. sequence from the distribution $\mathcal{D}$.

Two versions of the MBAL-SPO have their own advantages. When using hard rejection, we update the set of predictors H_t according to Line 20 in Algorithm 1, and the value of $\tilde{p}$ is set to zero. In contrast, in the soft rejection case, we keep H_t as the entire hypothesis class $\mathcal{H}$ for all iterations, and the value of $\tilde{p}$ is non-zero. In comparison, hard rejection can result in a smaller label complexity because $\tilde{p} = 0$, while soft rejection can reduce computational complexity by keeping H_t as the whole hypothesis class $\mathcal{H}$. Please see the discussion in Sections 2.4.2 and 2.4.4 for further details.

In Algorithm 1, the case where $\nu_S(h_{t-1}(x_t)) \geq b_{t-1}$ intuitively corresponds to the case where the confidence region of $h_{t-1}(x_t)$ does not intersect with the degenerate cost vectors. Hence, we are sufficiently confident that the optimal decision $w^*(h_t(x_t))$ is equal to $w^*(h^*(x_t))$, where h^* is a model that minimizes the SPO risk. Thus, we do not have to ask for the label of x_t. Lemma 2.3.1 further characterizes the conditions when two predictions lead to the same decision when the feasible region S is polyhedral.

Lemma 2.3.1 (Conditions for identical decisions in polyhedral feasible regions.)**.** *Suppose that the feasible region S is polyhedral. Given two cost vectors $c_1, c_2 \in \mathbb{R}^d$, if $\|c_1 - c_2\| < \max\{\nu_S(c_1), \nu_S(c_2)\}$, then it holds that $w^*(c_1) = w^*(c_2)$. In other words, the optimal decisions for c_1 and c_2 are the same.*

Lemma 2.3.1 implies that given one prediction of the cost vector, when its distance to degeneracy is larger than the radius of its confidence region, then all the predictions within this confidence region will lead to the same decision. Moreover, if the optimal prediction is also within this confidence region, the SPO loss of this prediction is zero.

The computational complexity of Algorithm 1 depends on the choice of the surrogate loss we use. As discussed earlier, calculating the distance to degeneracy $\nu_S(h(x))$ is efficient in some special cases. In general, in the polyhedral case when a convex hull representation is not available, a reasonable heuristic is to only compute the minimum in (2.4) with respect to the neighboring extreme points of $w^*(c)$. Alternatively, we observe that the objective inside the minimum in (2.4) is quasiconcave. Therefore, we can relax the condition that v_j be an extreme point and still recover an extreme point solution. One can solve the resulting problem with a Frank-Wolfe type method, for example, see Yurtsever and Sra (2022). The computational complexity of updating h_t in Line 19 depends on the choice of hypothesis class $\mathcal{H}$. In the case of soft rejection, we maintain $H_t = \mathcal{H}$ for all t and the update is the same as performing empirical risk minimization in $\mathcal{H}$, which can be efficiently computed exactly or approximately for most common choices of $\mathcal{H}$, including linear and nonlinear models. In the case of hard rejections, H_t is now the intersection of t different level sets. Thus, $\min_{h \in H_t} \hat{\ell}^t(h)$ is a minimization problem with t level set constraints. The complexity of solving this problem again depends on the choice of $\mathcal{H}$ and can often be solved efficiently. For example, in the case of linear models or nonlinear models such as neural networks, a viable approach would be to apply stochastic gradient descent to a penalized version of the problem or to apply a Lagrangian dual-type algorithm. In practice, since the constraints may be somewhat loose, we may simply ignore them and still obtain good results. Finally, we note that in both cases of hard rejection and soft rejection, although we have to solve a different optimization problem at every iteration, these optimization problems do not change much from one iteration to the next, and therefore using a warm-start strategy that uses h_{t-1} as the initialization for calculating h_t will be very effective.

2.3.1.1 Surrogate Loss Function and Noise Distribution

Without further assumptions on the distribution of noise and features, the label complexity of an active learning algorithm can be the same as the sample complexity of supervised learning, as shown in Kääriäinen (2006). Therefore, we make several natural assumptions in order to analyze the convergence and label complexity of our algorithm. Recall that the optimal SPO and surrogate risk values are defined as:

$$R^*_{\text{SPO}} := \min_{h \in \mathcal{H}} R_{\text{SPO}}(h), \quad \text{and} \quad R^*_\ell := \min_{h \in \mathcal{H}} R_\ell(h).$$

Algorithm 1 Margin-Based Active Learning for SPO (MBAL-SPO)

Input: Exploration probability $\tilde{p}$, a sequence of cut-off values $\{b_t\}$, a sequence $\{r_t\}$, and a constant ϑ.
Initialize the working sets $W_0 \leftarrow \emptyset$, $\tilde{W}_0 \leftarrow \emptyset$ and $H_0 \leftarrow \mathcal{H}$.
Arbitrarily pick one $h_0 \in \mathcal{H}$, $n_0 \leftarrow 0$.
for t from $1, 2, ..., T$ **do**
  Draw one sample x_t from $\mathcal{D}_{\mathcal{X}}$.
  if $\nu_S(h_{t-1}(x_t)) \geq b_{t-1}$ **then**
    Flip a coin with heads-up probability $\tilde{p}$.
    if the coin gets heads-up **then**
      Acquire a "true" label c_t of x_t.
      Update working set $\tilde{W}_t \leftarrow \tilde{W}_{t-1} \cup \{(x_t, c_t)\}$. Set $n_t \leftarrow n_{t-1} + 1$.
    else
      Reject x_t. Set $n_t \leftarrow n_{t-1}$ and $\tilde{W}_t \leftarrow \tilde{W}_{t-1}$.
    end if
  else
    Acquire a "true" label c_t of x_t.
    Update working set $W_t \leftarrow W_{t-1} \cup \{(x_t, c_t)\}$. Set $n_t \leftarrow n_{t-1} + 1$.
  end if
  Let $\hat{\ell}^t(h) \leftarrow \frac{1}{t}\left(\sum_{(x,c)\in W_t} \ell(h(x), c) + \frac{1}{\tilde{p}} \sum_{(x,c)\in \tilde{W}_t} \ell(h(x), c)\right)$.
  Update $h_t \leftarrow \arg\min_{h\in H_{t-1}} \hat{\ell}^t(h)$ and $\hat{\ell}^{t,*} \leftarrow \min_{h\in H_{t-1}} \hat{\ell}^t(h)$.
  Optionally update the confidence set of the predictor H_t by $H_t \leftarrow \{h \in H_{t-1} : \hat{\ell}^t(h) \leq \hat{\ell}^{t,*} + r_t + \frac{\vartheta}{t}\sum_{i=0}^{t-1} b_i^2\}$.
end for
Return h_T.

We define $\mathcal{H}^*$ as the set of all optimal predictors for the SPO risk, i.e., $\mathcal{H}^* = \{h \in \mathcal{H} : R_{\text{SPO}}(h) \leq R_{\text{SPO}}(h'), \text{ for all } h' \in \mathcal{H}\}$ and $\mathcal{H}_\ell^*$ as the set of all optimal predictors for the risk of the surrogate loss, i.e., $\mathcal{H}_\ell^* = \{h \in \mathcal{H} : R_\ell(h) \leq R_\ell(h'), \text{ for all } h' \in \mathcal{H}\}$. We also use the notation $R_{\text{SPO+}}^*$ and $\mathcal{H}_{\text{SPO+}}^*$ when the surrogate loss ℓ is SPO+. We define the essential sup norm of a function $h : \mathcal{X} \to \mathbb{R}^d$ as $\|h\|_\infty := \inf\{\alpha \geq 0 : \|h(x)\| \leq \alpha \text{ for almost every } x \in \mathcal{X}\}$, with respect to the marginal distribution of x and where $\|\cdot\|$ is the norm defining the distance to degeneracy (Definition 2.3.1). Given a set $\mathcal{H}' \subseteq \mathcal{H}$, we further define the distance between a fixed predictor function h and $\mathcal{H}'$ as $\text{Dist}_{\mathcal{H}'}(h) := \inf_{h'\in\mathcal{H}'}\{\|h - h'\|_\infty\}$. Assumption 2.3.1 states our main assumptions on the surrogate loss function ℓ that we work with.

Assumption 2.3.1 (Consistency and error bound condition)**.** *The hypothesis class $\mathcal{H}$ is a nonempty compact set w.r.t. to the sup norm, and the surrogate loss function $\ell : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}_+$ is continuous and satisfies:*

(1) $\mathcal{H}_\ell^* \subseteq \mathcal{H}^*$, i.e., the minimizers of the surrogate risk are also minimizers of the SPO risk.

(2) There exists a non-decreasing function $\phi : \mathbb{R}_+ \to \mathbb{R}_+$ *with* $\phi(0) = 0$ *such that for any* $h \in \mathcal{H}$*, for any* $\epsilon > 0$*,*

$$R_\ell(h) - R_\ell^* \leq \epsilon \;\Rightarrow\; \text{Dist}_{\mathcal{H}_\ell^*}(h) \leq \phi(\epsilon).$$

Assumption 2.3.1.(1) states the consistency of the surrogate loss function. Note that since $\mathcal{H}$ is a nonempty compact set and ℓ is a continuous function, $\mathcal{H}_\ell^*$ is also a nonempty compact set. On the other hand, the SPO loss is generally discontinuous so $\mathcal{H}^*$ is not necessarily compact, although the consistency assumption $\mathcal{H}_\ell^* \subseteq \mathcal{H}^*$ ensures that $\mathcal{H}^*$ is nonempty. Assumption 2.3.1.(2) is a type of error bound condition on the risk of the surrogate loss, wherein the function ϕ provides an upper bound of the sup norm between the predictor h and the set of optimal predictors $\mathcal{H}_\ell^*$ whenever the surrogate risk of h is close to the minimum surrogate risk value. By Assumption 2.3.1.(2), when the excess surrogate risk of h becomes smaller, h becomes closer to the set $\mathcal{H}_\ell^*$, which implies that the prediction $h(x)$ also gets closer to an optimal prediction $h^*(x)$ for any given x. As a consequence, the distance to degeneracy $\nu_S(h(x))$ also converges to $\nu_S(h^*(x))$ for almost all $x \in \mathcal{X}$. This property enables us to analyze the performance of MBAL-SPO under SPO+ and surrogate loss function respectively in the next two sections. Assumption 2.3.1 is related to the uniform calibration property studied in Ho-Nguyen and Kılınç-Karzan (2022) in the SPO context. Next, to measure how the density of the distribution $\nu_S(h^*(x))$ is allocated near the points of degeneracy, we define the near-degeneracy function Ψ in Definition 2.3.2.

Definition 2.3.2 (Near-degeneracy function)**.** *The near-degeneracy function* $\Psi : \mathbb{R}_+ \to [0, 1]$ *with respect to the distribution of* $x \sim \mathcal{D}_\mathcal{X}$ *is defined as:*

$$\Psi(b) := \mathbb{P}\left(\inf_{h^* \in \mathcal{H}^*} \{\nu_S(h^*(x))\} \leq b\right).$$

□

The near-degeneracy function Ψ measures the probability that the distance to degeneracy of $h^*(x)$ is smaller than b, when x follows the marginal distribution of x in $\mathcal{D}_\mathcal{X}$. If $\mathcal{H}^*$ contains more than one optimal predictor, the near-degeneracy function Ψ considers the distribution of the smallest distance to degeneracy of all optimal predictors h^*. Intuitively, when $\Psi(b)$ is smaller, the density allocated near the points of degeneracy becomes smaller, which means Algorithm 1 has a larger probability to reject samples, and achieves smaller label complexity. This intuition is characterized in Lemma 2.3.2.

Lemma 2.3.2 (Upper bound on the expected number of acquired labels)**.** *Suppose that Assumption 2.3.1 holds. In Algorithm 1, if* h_t *satisfies* $\text{Dist}_{\mathcal{H}_\ell^*}(h_t) \leq b_t$ *for all iterations* $t \geq 0$*, then the expected number of acquired labels after* T *total iterations is at most* $\tilde{p}T + \sum_{t=1}^{T} \Psi(2b_{t-1})$.

Lemma 2.3.2 provides an upper bound for the expected number of acquired labels up to time t, by utilizing the near-degeneracy function Ψ. Note that in the soft rejection case, if $\tilde{p} > 0$ and $\tilde{p}$ is independent of T, Lemma 2.3.2 implies that this upper bound grows linearly in T. However, if we know the value of T before running the algorithm, then this upper bound can be reduced to a sublinear order by setting $\tilde{p}$ as a function of T. On the other hand, if we can set $\tilde{p} = 0$, i.e., in the hard rejection case, the upper bound in Lemma 2.3.2 is sublinear if $\sum_{t=1}^{T} \Psi(2b_t)$ is sublinear. As will be shown later in Proposition 2.5.1, in the hard rejection case, we achieve a sublinear and sometimes even finite label complexity when the near-degeneracy function Ψ satisfies certain conditions.

2.4 Guarantees and Analysis for the Margin-Based Algorithm

In this section, we analyze the convergence and label complexity of MBAL-SPO in various settings. We first review some preliminary information about generalization error bound in Section 2.4.1. Next, we analyze the label complexity under hard rejections and soft rejections in Sections 2.4.2 and 2.4.4, respectively. In both sections, we develop non-asymptotic surrogate and SPO risk error bounds. We also develop bounds for the label complexity that, under certain conditions, can be much smaller than supervised learning. In Section 2.4.3, we further provide tighter SPO+ and SPO risk bounds when using the SPO+ surrogate under a separability condition. At the end of this section, we discuss how to set the values of the parameters in practice in MBAL-SPO.

2.4.1 Reweighted loss function

In Algorithm 1, the samples in the training set are not i.i.d., instead, whether to acquire the label at iteration t depends on the historical label results. One of the challenges in analyzing the convergence and label complexity of the margin-based algorithm stems from the non i.i.d. samples. In this section, we review some techniques that characterize the convergence of non i.i.d. random sequences.

In Algorithm 1, the random variables in one iteration can be written as (x_t, c_t, d_t^M, q_t), where $d_t^M \in \{0, 1\}$ represents whether the sample is near degeneracy or not, i.e. if $\nu_S(h_{t-1}(x_t)) < b_{t-1}$ then $d_t^M = 1$, otherwise $d_t^M = 0$. The random variable $q_t \in \{0, 1\}$ represents the outcome of the coin flip that determines if we acquire the label of this sample or not, in the case when $d_t^M = 0$. For simplicity, we use random variable $z_t \in \mathcal{Z} := \mathcal{X} \times \mathcal{C} \times \{0, 1\} \times \{0, 1\}$ to denote the tuple of random variables $z_t := (x_t, c_t, d_t^M, q_t)$. Thus, z_t depends on $z_1, ..., z_{t-1}$ and the classical convergence results for i.i.d. samples do not apply in the margin-based algorithm. We define $\mathcal{F}_{t-1}$ as the σ-field of all random variables until the end of iteration $t - 1$ (i.e., $\{z_1, ..., z_{t-1}\}$). In Algorithm 1, the re-weighted loss function at iteration t is

$\ell^{\texttt{rew}}(h; z_t) := d_t^M \ell(h(x_t), c_t) + (1 - d_t^M) q_t \frac{\mathbb{I}\{\tilde{p}>0\}}{\tilde{p}} \ell(h(x_t), c_t)$. It is easy to see that $\ell^{\texttt{rew}}(h; z)$ is upper bounded by $\frac{\omega_\ell(\hat{\mathcal{C}}, \mathcal{C})}{\tilde{p}^{\mathbb{I}\{\tilde{p}>0\}}} < \infty$.

Then, to analyze the uniform convergence of $\frac{1}{T}\sum_{t=1}^T \ell^{\texttt{rew}}(h; z_t)$ to $\frac{1}{T}\sum_{t=1}^T \mathbb{E}[\ell^{\texttt{rew}}(h; z_t)|\mathcal{F}_{t-1}]$, we assume that the hypothesis class $\mathcal{H}$ is discrete and its cardinality $|\mathcal{H}|$ is at most N_1. In other words, there are at most N_1 candidate predictors within the hypothesis class $\mathcal{H}$. This assumption simplifies our notations and analysis. However, our analysis can also be extended to accommodate a hypothesis class with an infinite number of predictors, such as linear models by using some sequential complexity introduced in Rakhlin, Sridharan, and Tewari, 2015. Based on this assumption, Proposition 2.4.1 provides an upper bound for the convergence rate of the reweighted loss.

Proposition 2.4.1 (Generalization error bound for the reweighted loss)**.** *Suppose that $|\mathcal{H}| \leq N_1$. Let $\{z_1, z_2, ...z_T\}$ be a (non i.i.d.) sequence of random variables. Then, the following inequality holds:*

$$\mathbb{P}\left(\sup_{h\in\mathcal{H}}\left\{\left|\frac{1}{T}\sum_{t=1}^T (\mathbb{E}[\ell^{\texttt{rew}}(h; z_t)|\mathcal{F}_{t-1}] - \ell^{\texttt{rew}}(h; z_t))\right|\right\} \geq \epsilon\right) \leq 2N_1 \exp\left\{-\frac{2\tilde{p}^{2\mathbb{I}\{\tilde{p}>0\}} T\epsilon^2}{\omega_\ell^2(\hat{\mathcal{C}}, \mathcal{C})}\right\}.$$

Proposition 2.4.1 shows that $\frac{1}{T}\sum_{t=1}^T \ell^{\texttt{rew}}(h; z_t)$ converges to $\frac{1}{T}\sum_{t=1}^T \mathbb{E}[\ell^{\texttt{rew}}(h; z_t)|\mathcal{F}_{t-1}]$ at rate $\tilde{\mathcal{O}}(1/\sqrt{T})$.

2.4.2 MBAL-SPO with Hard Rejections

In this section, we develop excess risk bounds for the surrogate risk and the SPO risk, and present label complexity results, for MBAL-SPO with hard rejections. Our excess risk bounds for the surrogate risk hold for general feasible regions S. To develop risk bounds for the SPO risk, we consider two additional assumptions on S: *(i)* the case where S satisfies the strength property, and *(ii)* the case where S is polyhedral. The strength property, as defined in El Balghiti et al. (2022), is reviewed below in Definition 2.4.1.

Definition 2.4.1 (Strength Property for the Feasible Region S)**.** *The feasible region S satisfies the strength property with constant $\mu > 0$ if, for all $w \in S$ and $\hat{c} \in \mathbb{R}^d$, it holds that*

$$\hat{c}^T(w - w^*(\hat{c})) \geq \frac{\mu \cdot \nu_S(\hat{c})}{2}\|w - w^*(\hat{c})\|^2, \tag{2.5}$$

where ν_S is the distance to degeneracy function. We refer to μ as the strength parameter. □

The strength property can be interpreted as a variant of strong convexity that bounds the distance to the optimal solution based on the parameter μ as well as the distance to degeneracy $\nu_S(\hat{c})$. El Balghiti et al. (2022) demonstrate that the strength property holds when S is polyhedral or a strongly convex set. In addition, for some of the results herein, we make the following assumption concerning the surrogate loss function, which states the uniqueness of the surrogate risk minimizer and a relaxation of Hölder continuity.

Assumption 2.4.1 (Unique minimizer and Hölder-like property)**.** *There is a unique minimizer h^* of the surrogate risk, i.e., the set $\mathcal{H}_\ell^*$ is a singleton, and there exists a constant $\eta > 0$ such that the surrogate loss function ℓ satisfies*

$$|\mathbb{E}[\ell(\hat{c}, c) - \ell(h^*(x), c)|x]| \ \leq \ \eta \|\hat{c} - h^*(x)\|^2 \text{ for all } x \in \mathcal{X}, \text{ and } \hat{c} \in \hat{\mathcal{C}}.$$

It is easy to verify that the common squared loss satisfies Assumption 2.4.1 with $\eta = 1$ when the hypothesis class is well-specified. In Lemma 2.6.3 in Section 2.6, we further show that the SPO+ loss satisfies Assumption 2.4.1 under some noise conditions.

Theorem 2.4.1 is our main theorem concerning MBAL-SPO with hard rejections and with general surrogate losses satisfying Assumption 2.4.1. Theorem 2.4.1 presents bounds on the excess surrogate and SPO risks as well as the expected label complexity after T iterations.

Theorem 2.4.1 (General surrogate loss, hard rejection)**.** *Suppose that $|\mathcal{H}| \leq N_1$, that Assumptions 2.3.1 and 2.4.1 hold, and that Algorithm 1 sets $\tilde{p} \leftarrow 0$ and updates the set of predictors according to the optional update rule in Line 20 with $\vartheta \leftarrow \eta$. Furthermore in Algorithm 1, for a given $\delta \in (0, 1]$, let $r_0 \geq \omega_\ell(\hat{\mathcal{C}}, \mathcal{C})$, $r_t \leftarrow 2\omega_\ell(\hat{\mathcal{C}}, \mathcal{C})\sqrt{\frac{\ln(2TN_1/\delta)}{T}}$ for $t \geq 1$, $b_0 \leftarrow \max\{\phi(r_0), \sqrt{r_0/\eta}\}$, and $b_t \leftarrow \phi(2r_t + \frac{2\eta}{t}\sum_{i=0}^{t-1} b_i^2)$ for $t \geq 1$. Then, the following guarantees hold simultaneously with probability at least $1 - \delta$ for all $T \geq 1$:*

- *(a) The excess surrogate risk satisfies $R_\ell(h_T) - R_\ell^* \leq r_T + \frac{\eta}{T}\sum_{t=0}^{T-1} b_t^2$,*
- *(b) If the feasible region S satisfies the strength property with parameter $\mu > 0$, then the excess SPO risk satisfies*

$$R_{\text{SPO}}(h_T) - R_{\text{SPO}}^* \leq \inf_{\gamma_T \geq 2b_T} \left\{ \frac{2\rho(\mathcal{C})b_T}{\mu\gamma_T} + \Psi(\gamma_T)\omega_S(\mathcal{C}) \right\},$$

- *(c) If the feasible region S is polyhedral, then the excess SPO risk satisfies $R_{\text{SPO}}(h_T) - R_{\text{SPO}}^* \leq \Psi(2b_T)\omega_S(\mathcal{C})$,*
- *(d) The expectation of the number of labels acquired, $\mathbb{E}[n_T]$, deterministically satisfies $\mathbb{E}[n_T] \leq \sum_{t=1}^{T} \Psi(2b_{t-1}) + \delta T$.*

In the polyhedral case, Theorem 2.4.1 indicates that the excess SPO risk of Algorithm 2.4.1 converges to zero at rate $\mathcal{O}(\Psi(2b_T))$, and the expectation of the number of acquired labels grows at rate $\mathcal{O}\left(\sum_{t=1}^{T} \Psi(2b_t)\right)$ for small δ. (Usually, $\delta \ll \mathcal{O}(1/T)$.) Note that Theorem 2.4.1 is generic in that the excess risk and label complexity bounds depend on the functions ϕ and Ψ. In Section 2.6, we give the explicit forms of these functions in some special cases of interest.

Remark 2.4.1 (Updates of H_t)**.** *In Theorem 2.4.1, the set of predictors is updated according to Line 20 in Algorithm 1. This is a technical requirement for the convergence when setting*

$\tilde{p} = 0$. *This update process means that* $h_t \in H_{t-1} \subseteq H_{t-2} ... \subseteq H_0 = \mathcal{H}$. *By constructing these shrinking sets* H_t *of predictors, we are able to utilize the information from previous iterations. Particularly, Lemma 2.4.2 below shows that these shrinking sets* H_{t-1} *always contain the true optimal predictor* h^* *under certain conditions.* □

Remark 2.4.2 (Value of γ_T)**.** *In Theorem 2.4.1, to find the best value of the parameter* γ_T *in part* (b) *that minimizes the excess SPO risk for sets satisfying the strength property, we observe that the choice of* γ_T *depends on* Ψ, ϕ *and* T. *If* γ_T *satisfies that (1)* $\frac{b_T}{\gamma_T} \to 0$, *when* $r_T \to 0$, *and (2)* $\gamma_T \to 0$, *when* $T \to \infty$, *then the excess SPO risk will converge to zero. For example, we can set* $\gamma_T = (b_T)^\kappa$, *where* $\kappa \in (0, 1)$. □

Auxiliary Results for the Proof of Theorem 2.4.1. To achieve the risk bound in part *(a)* of Theorem 2.4.1, we decompose the excess surrogate risk into three parts. First, we denote the re-weighted surrogate risk for the features that are far away from degeneracy by $\ell_t^{\mathrm{f}}(h)$, defined by:

$$\ell_t^{\mathrm{f}}(h) := \mathbb{E}[\ell(h; z_t)\mathbb{I}\{\nu_S(h_{t-1}(x_t)) \geq b_{t-1}\}|\mathcal{F}_{t-1}] = \mathbb{E}[\ell(h; z_t)(1 - d_t^M)|\mathcal{F}_{t-1}],$$

where we use $\ell(h; z_t)$ to denote $\ell(h(x_t), c_t)$ and the expectation above is with respect to z_t. Since x_t and c_t are i.i.d. random variables, and only d_t^M depends on $\mathcal{F}_{t-1}$, $\ell_t^{\mathrm{f}}(h)$ can further be written as $\ell_t^{\mathrm{f}}(h) = \mathbb{E}[\ell(h(x_t), c_t)|d_t^M = 0]\mathbb{P}(d_t^M = 0|\mathcal{F}_{t-1})$. Note also that, since $\tilde{p} = 0$, the re-weighted loss function can be written as $\ell^{\mathtt{rew}}(h; z_t) = \ell(h(x_t), c_t)d_t^M = \ell(h(x_t), c_t)\mathbb{I}\{\nu_S(h_{t-1}(x_t)) < b_{t-1}\}$, for a given $h \in \mathcal{H}$. Next, for given $h \in \mathcal{H}$ and $h^* \in \mathcal{H}_\ell^*$, we denote the discrepancy between the conditional expectation and the realized excess re-weighted loss of predictor h at time t by $Z_h^{\mathtt{t}}$, i.e., $Z_h^{\mathtt{t}} := \mathbb{E}[\ell^{\mathtt{rew}}(h; z_t) - \ell^{\mathtt{rew}}(h^*; z_t)|\mathcal{F}_{t-1}] - (\ell^{\mathtt{rew}}(h; z_t) - \ell^{\mathtt{rew}}(h^*; z_t))$. Lemma 2.4.1 shows that the excess surrogate risk can be decomposed into three parts.

Lemma 2.4.1 (Decomposition of the excess surrogate risk)**.** *In the case of hard rejections, i.e.,* $\tilde{p} \leftarrow 0$ *in Algorithm 1, for any given* $h^* \in \mathcal{H}_\ell^*$ *and* $T \geq 1$, *the excess surrogate risk of any predictor* $h \in \mathcal{H}$ *can be decomposed as follows:*

$$R_\ell(h) - R_\ell(h^*) \;=\; \frac{1}{T}\sum_{t=1}^{T}\left(\ell_t^{\mathrm{f}}(h) - \ell_t^{\mathrm{f}}(h^*)\right) + \frac{1}{T}\sum_{t=1}^{T} Z_h^{\mathtt{t}} + \frac{1}{T}\sum_{t=1}^{T}\left(\ell^{\mathtt{rew}}(h; z_t) - \ell^{\mathtt{rew}}(h^*; z_t)\right).$$

The first part in Lemma 2.4.1 is the averaged excess surrogate risk for the hard rejected features at each iteration. Lemma 2.4.2 below further shows that $\left|\ell_t^{\mathrm{f}}(h) - \ell_t^{\mathrm{f}}(h^*)\right|$ is close to zero when $h \in H_{T-1}$.

Lemma 2.4.2. *Suppose that Assumptions 2.3.1 and 2.4.1 hold where* h^* *denotes the unique minimizer of the surrogate risk, and that Algorithm 1 sets* $\tilde{p} \leftarrow 0$ *and updates the set of predictors according to the optional update rule in Line 20 with* $\vartheta \leftarrow \eta$. *Furthermore, suppose that that* $r_0 \geq \omega_\ell(\hat{\mathcal{C}}, \mathcal{C})$, $r_t \geq \sup_{h\in\mathcal{H}}\left\{\left|\frac{1}{t}\sum_{i=1}^{t} Z_h^{\mathtt{i}}\right|\right\}$ *for* $t \geq 1$, $b_0 \leftarrow \max\{\phi(r_0), \sqrt{r_0/\eta}\}$, *and* $b_t \leftarrow \phi(2r_t + \frac{2\eta}{t}\sum_{i=0}^{t-1} b_i^2)$ *for* $t \geq 1$. *Then, for all* $t \geq 1$, *it holds that* (a) $h^* \in H_{t-1}$, *and* (b) $\sup_{h\in H_{t-1}}\left\{\left|\ell_t^{\mathrm{f}}(h) - \ell_t^{\mathrm{f}}(h^*)\right|\right\} \leq \eta b_{t-1}^2$.

With Lemma 2.4.2, we can appropriately bound the first average of terms in Lemma 2.4.1, involving the expected surrogate risk when far from degeneracy. Thus, Lemmas 2.4.1 and 2.4.2 enable us to prove the excess surrogate risk bound in part *(a)*. The proofs of the remaining parts follow by translating the excess surrogate risk bound to guarantees on the excess SPO risk and the label complexity.

2.4.3 Refined Bounds for SPO+ Under Separability

Next, we provide a smaller excess risk bound when using SPO+ as the surrogate loss, again in the case of hard rejections. The SPO+ loss function incorporates the structure of the downstream optimization problem and, intuitively, the excess SPO+ risk when far away from degeneracy will be close to zero when the distance $\text{Dist}_{\mathcal{H}_\ell^*}(h)$ is small and the distribution satisfies a separability condition, which we define below.

Assumption 2.4.2 (Strong separability condition)**.** *There exist constants $\varrho \in [0, 1)$ and $\tau \in (0, 1]$ such that, for all $h^* \in \mathcal{H}^*_{\text{SPO+}}$, with probability one over $(x, c) \sim \mathcal{D}$, it holds that:*

- *(1) $\|h^*(x) - c\| \leq \varrho \nu_S(h^*(x))$, and*
- *(2) $\nu_S(h^*(x)) \geq \tau \left(\sup_{h' \in \mathcal{H}^*_{\text{SPO+}}} \{\nu_S(h'(x))\}\right)$.*

The following proposition shows that the separability condition leads to zero SPO+ and SPO risk in the polyhedral case. Indeed, the SPO+ loss is a generalization of the hinge loss and the structured hinge loss in binary and multi-class classification problems and is expected to achieve zero loss when there is a predictor function $\bar{h}$ that strictly separates the cost vectors into different classes corresponding to the extreme points of S Elmachtoub and Grigas (2022). Assumption 2.4.2 and Proposition 2.4.2 formally define the notion of separability, wherein the distance between the prediction $\bar{h}(x)$ and the realized cost vector c, relative to the distance to degeneracy of $\bar{h}(x)$, is controlled.

Proposition 2.4.2 (Zero SPO+ risk in the polyhedral and separable case)**.** *Assume that there exists $\bar{h} \in \mathcal{H}$ and a constant $\varrho \in [0, 1)$ such that $\|\bar{h}(x) - c\| \leq \varrho \nu_S(\bar{h}(x))$ with probability one over $(x, c) \sim \mathcal{D}$. When the feasible region S is polyhedral, it holds that $R^*_{\text{SPO+}} = R^*_{\text{SPO}} = 0$ and $\bar{h}$ is a minimizer for both $R_{\text{SPO+}}$ and R_{SPO}.*

As compared to Theorem 2.4.1, Theorem 2.4.2 below presents improved surrogate risk convergence guarantees for SPO+ under separability in the polyhedral case.

Theorem 2.4.2 (SPO+ surrogate loss, hard rejection, polyhedral and separable case)**.** *Suppose that the feasible region S is polyhedral, that Assumptions 2.3.1 and 2.4.2 hold with $|\mathcal{H}| \leq N_1$, and that the surrogate loss function is SPO+. Suppose that Algorithm 1 sets $\tilde{p} \leftarrow 0$ and $H_t \leftarrow \mathcal{H}$ for all t. Furthermore in Algorithm 1, for a given $\delta \in (0, 1]$, let $r_0 \geq \omega_\ell(\hat{\mathcal{C}}, \mathcal{C})$, $r_t \leftarrow \omega_\ell(\hat{\mathcal{C}}, \mathcal{C})\sqrt{\frac{\ln(2TN_1/\delta)}{T}}$ for $t \geq 1$, $b_0 \leftarrow \max\{\phi(r_0), \rho(\hat{\mathcal{C}})\}$, and $b_t \leftarrow (1 + \frac{2}{\tau(1-\rho)})\phi(r_t)$ for*

$t \geq 1$. Then, the following guarantees hold simultaneously with probability at least $1-\delta$ for all $T \geq 1$:

- *(a) The excess SPO+ risk satisfies $R_{\text{SPO+}}(h_T) - R^*_{\text{SPO+}} = R_{\text{SPO+}}(h_T) \leq r_T$,*
- *(b) The excess SPO risk satisfies $R_{\text{SPO}}(h_T) - R^*_{\text{SPO}} = R_{\text{SPO}}(h_T) \leq \Psi(2b_T)\omega_S(\mathcal{C})$,*
- *(c) The expectation of the number of labels acquired, $\mathbb{E}[n_T]$, deterministically satisfies $\mathbb{E}[n_T] \leq \sum_{t=1}^{T} \Psi(2b_{t-1}) + \delta T$.*

Remark 2.4.3 (Benefits of SPO+ Under Separability)**.** *When using SPO+ in the separable case, the bound in part* (a) *of Theorem 2.4.2 is substantially improved as compared to Theorem 2.4.1. Intuitively, when an optimal predictor $h^*(x)$ is far away from degeneracy and $h_t(x)$ and $h^*(x)$ are close, then the excess SPO+ risk of $h_t(x)$ can be shown to be zero. As a result, the rejection criterion – which compares $\nu_S(h_t(x))$ to a quantity b_t that is related to the distance between h_t and h^* – is "safe" in the sense that whenever $h_t(x) \geq b_t$ we can demonstrate that $h_t(x)$ leads to a correct optimal decision with high probability. Thus, when using the SPO+ loss function, we can obtain a smaller excess SPO+ risk bound. Indeed, in Theorem 2.4.2, the value of r_t is determined by the i.i.d. covering number, which implies that this risk bound is the same as the risk bound of supervised learning that labels all the samples. Furthermore, another benefit of SPO+ under the separability assumption is that we do not need to update H_t at each iteration, which simplifies the computation substantially. Finally, Assumption 2.4.1 which assumes the minimizer h^* is unique is not needed.* □

Theorem 2.4.2 shows that the excess SPO+ risk converges to zero at rate $\tilde{\mathcal{O}}(1/\sqrt{T})$, which equals the typical learning rate for the excess SPO+ risk in supervised learning. As Algorithm 1 requires much fewer labels, this demonstrates the advantage of active learning. In fact, the main idea of the proof of Theorem 2.4.2 is to show that h_T actually, with high probability, achieves zero empirical SPO+ risk over all samples $(x_1, c_1), \ldots, (x_T, c_T)$ – including the cases where the label is not acquired. Indeed, in the separable case, the rejection criterion is "safe" and we are able to demonstrate that $\ell_{\text{SPO+}}(h_t(x_t), c_t) = 0$ when $\nu_S(h_{t-1}(x_t)) \geq b_{t-1}$. This of course implies that h_T is an empirical risk minimizer for SPO+ across T i.i.d. samples $(x_1, c_1), \ldots, (x_T, c_T)$ and we are able to conclude part *(a)*.

2.4.4 MBAL-SPO with Soft Rejections

In this section, we analyze the convergence and label complexity of MBAL-SPO with soft rejection. We return to the setting of a generic surrogate loss function ℓ. Compared to the hard rejection case in Theorem 2.4.1, this positive $\tilde{p}$ will lead to a larger label complexity than Theorem 2.4.1. On the other hand, when $\tilde{p}$ is positive, we do not have to construct the confidence set H_t of the predictors at each iteration. In other words, H_t can be set as $\mathcal{H}$, for all t as in Theorem 2.4.2. Thus, we do not have to consider t additional constraints when minimizing the empirical re-weighted risk, which will reduce the computational complexity significantly. Theorem 2.4.3 is our main theorem for the MBAL-SPO under a general surrogate

loss, which again provides upper bounds for the excess surrogate and SPO risk and label complexity of the algorithm.

Theorem 2.4.3 (General surrogate loss, soft rejection)**.** *Suppose that Assumption 2.3.1 holds with* $|\mathcal{H}| \leq N_1$*, and let* $\delta \in (0,1]$ *and* $\tilde{p} \in (0,1]$ *be given. In Algorithm 1, set* $H_t \leftarrow \mathcal{H}$ *for all* $t \geq 0$*,* $r_0 \geq \omega_\ell(\hat{\mathcal{C}}, \mathcal{C})$*,* $r_t \leftarrow \frac{4\omega_\ell(\hat{\mathcal{C}},\mathcal{C})}{\tilde{p}}\sqrt{\frac{\ln(2N_1/\delta)}{t}}$ *for* $t \geq 1$*,* $b_t \leftarrow 2\phi(r_t)$ *for* $t \geq 0$*. Then, the following guarantees hold simultaneously with probability at least* $1-\delta$ *for all* $T \geq 1$*:*

- *(a) The excess surrogate risk satisfies* $R_\ell(h_T) - R_\ell^* \leq r_T$*,*
- *(b) If the feasible region* S *satisfies the strength property with parameter* $\mu > 0$*, then the excess SPO risk satisfies*

$$R_{\mathrm{SPO}}(h_T) - R_{\mathrm{SPO}}^* \leq \inf_{\gamma_T \geq 2b_T}\left\{\frac{2\rho(\mathcal{C})b_T}{\mu\gamma_T} + \Psi(\gamma_T)\omega_S(\mathcal{C})\right\},$$

- *(c) If the feasible region* S *is polyhedral, then the excess SPO risk satisfies* $R_{\mathrm{SPO}}(h_T) - R_{\mathrm{SPO}}^* \leq \Psi(2b_T)\omega_S(\mathcal{C})$*,*
- *(d) The expectation of the number of labels acquired,* $\mathbb{E}[n_T]$*, deterministically satisfies* $\mathbb{E}[n_T] \leq \tilde{p}T + \sum_{t=1}^{T}\Psi(2b_{t-1}) + \delta T$*.*

Remark 2.4.4 (Value of b_t and $\tilde{p}$)**.** *In part* (d) *of Theorem 2.4.3,* $\mathbb{E}[n_T]$ *depends on both* $\tilde{p}T$ *and* $\sum_{t=1}^{T}\Psi(2b_t)$*. When the exploration probability* $\tilde{p}$ *is large,* $\tilde{p}T$ *in part* (d) *of Theorem 2.4.3 is large. On the other hand, in Theorem 2.4.3, the value of* b_t *depends on* r_t*, and* r_t *furthermore is in the order of* $O(1/\tilde{p})$*. It implies that when the exploration probability* $\tilde{p}$ *is small,* b_t *is large, and* $\sum_{t=1}^{T}\Psi(2b_t)$ *in part* (d) *of Theorem 2.4.3 is large. Hence, to minimize the label complexity, there is a trade-off when choosing the value of* $\tilde{p}$*. In Proposition 2.5.2, we will specify the value of* $\tilde{p}$ *and provide an upper bound for* $\mathbb{E}[n_T]$ *which is sublinear in* T*.* □

Although Theorem 2.4.3 does not require Assumptions 2.4.1 or 2.4.2, as will be shown later, to demonstrate the advantage of the supervised learning algorithm, we need to carefully select $\tilde{p}$, which will be elaborated in Section 2.5.2.

Setting Parameters in MBAL-SPO. To conclude this section, we discuss the issue of setting the parameters for MBAL-SPO in practice. Although Theorems 2.4.1 and 2.4.3 provide the theoretical settings for the parameters r_t and b_t in the MBAL-SPO algorithm, how to set the scale of these parameters is an important question in practice. The complexity of the hypothesis class, the noise level and the distribution of features all impact the settings of these parameters. When the noise level is larger, or the cost vector c is further away from the degeneracy, the scale of b_t for the algorithms should be larger. In addition, to set a proper scale of the parameters in practice, we need to consider the tradeoff between the budget of the labels (or the cost to acquire each label) and the efficiency of the learning process. A

reasonable practical approach is to set a "burn in" period of $\tilde{T}$ iterations where MBAL-SPO acquires all labels during the first $\tilde{T}$ iterations. One can then use the distribution of values $\nu_S(h_T(x_t))$ for all previous features x_t to inform the value of b_T. For example, we can set the scale of b_T as some order statistics of the past values $\nu_S(h_T(x_t))$ for $t \in \{1, \ldots, \tilde{T}\}$, e.g., the mean or other quantile depending on the practical cost of acquiring labels versus the rate at which feature vectors are collected. Then, the value of b_t for $t \geq T$ can be updated according to the value of b_T.

2.5 Risk Guarantees and Small Label Complexity Under Low Noise Conditions

To demonstrate the advantage of MBAL-SPO over supervised learning in Theorems 2.4.1 and 2.4.3, we need to analyze the functions ϕ and Ψ. In Section 2.6, we present some natural low-noise conditions such that we can provide concrete examples of ϕ under the SPO+ loss. In these examples, ϕ satisfies that $\phi(\epsilon) \sim \sqrt{\epsilon}$. In Section 2.6, we further show that Assumption 2.4.1 holds for SPO+ and derive the upper bound for η under some noise conditions. Given these results, in this section, we analyze the exact order of the label complexity and the risk bounds. These results demonstrate the advantages of MBAL-SPO.

2.5.1 Small Label Complexities

In this section, we analyze the order of the label complexity for both hard rejection and soft rejection. First, we characterize the noise as the level of near degeneracy in Assumption 2.5.1, which is similar in spirit to the low noise condition assumption in Hu, Kallus, and Mao (2022).

Assumption 2.5.1 (Near-degeneracy condition)**.** *There exist constants $b_0, \kappa > 0$ such that*

$$\Psi(b) = \mathbb{P}\left(\inf_{h^* \in \mathcal{H}^*} \{\nu_S(h^*(x))\} \leq b\right) \leq (b/b_0)^\kappa.$$

Assumption 2.5.1 controls the rate at which $\Psi(b)$ – which measures the probability mass of features with small distance to degeneracy – approaches 0 as b approaches 0. In other words, for small enough b so that $\frac{b}{b_0} < 1$, when the parameter κ is larger the probability near the degeneracy is smaller at a faster rate. When the above near-degeneracy condition in Assumption 2.5.1 holds and $\phi(\epsilon)$ satisfies that $\phi(\epsilon) \sim \mathcal{O}(\sqrt{\epsilon})$, we have the sublinear label complexity for the hard rejection in the polyhedral cases in Proposition 2.5.1.

Proposition 2.5.1 (Small label complexity for hard rejections)**.** *Suppose that Assumptions 2.3.1, 2.4.1, and 2.5.1 hold with $|\mathcal{H}| \leq N_1$. Suppose there exists a constant $C_\phi \in (0, \frac{1}{36\eta^2})$ such that Assumption 2.3.1.(2) holds with $\phi(\epsilon) = C_\phi \cdot \sqrt{\epsilon}$. Under the same setting of Algorithm 1 in Theorem 2.4.1, for a fixed $\delta \in (0, 1]$, the following guarantees hold simultaneously with probability at least $1 - \delta$ for all $T \geq 1$:*

- *The excess surrogate risk satisfies* $R_\ell(h_T) - R_\ell^* \leq \tilde{\mathcal{O}}(T^{-1/2})$.
- *The excess SPO risk satisfies* $R_{\text{SPO}}(h_T) - R_{\text{SPO}}^* \leq \tilde{\mathcal{O}}(T^{-\kappa/4})$.
- *The expectation of the number of labels acquired, conditional on the above guarantee on the excess surrogate risk, is at most* $\tilde{\mathcal{O}}\left(T^{1-\kappa/4}\right)$ *for* $\kappa \in (0, 4)$, *and* $\tilde{\mathcal{O}}(1)$ *for* $\kappa \in [4, \infty)$.

The last claim in Propositon 2.5.1 indicates that the label complexity is sublinear. Notice that, as compared to Theorem 2.4.1, for simplicity, we state the bound on the label complexity conditional on the excess SPO+ risk guarantee that holds with probability at least $1 - \delta$. When $\kappa > 4$, the label complexity is even finite. To compare this label complexity with supervised learning, we consider the excess SPO risk with respect to the number of labels n. Let $\bar{n} \leftarrow \mathbb{E}[n_T]$ be a fixed value. Under the same assumptions and similar proof procedures, we can show that the excess SPO risk of the supervised learning is at most $\tilde{\mathcal{O}}(\bar{n}^{-\kappa/2})$. In comparison, Proposition 2.5.1 indicates that the expected excess SPO risk of MBAL-SPO is at most $\tilde{\mathcal{O}}(\bar{n}^{-\frac{\kappa}{4-\kappa}})$. Thus, when $\kappa > 2$, MBAL-SPO acquires much fewer labels than the supervised learning to achieve the same level of SPO risk. This demonstrates the advantage of MBAL-SPO over supervised learning.

Remark 2.5.1 (Small label complexity under separability condition with SPO+ loss)**.** *Under the same setting as Theorem 2.4.2, obviously, we have that* $R_\ell(h_T) - R_\ell^* \leq \tilde{\mathcal{O}}(T^{-1/2})$. *If we further assume* $\phi(\epsilon) \sim \sqrt{\epsilon}$, *then following the same analysis in Proposition 2.5.1, we can obtain that* $R_{\text{SPO}}(h_T) - R_{\text{SPO}}^* \leq \tilde{\mathcal{O}}(T^{-\kappa/4})$ *and the expected number of labels is at most* $\tilde{\mathcal{O}}(T^{1-\kappa/4})$ *for* $\kappa \in (0, 4)$ *and* $\tilde{\mathcal{O}}(1)$ *for* $\kappa \in [4, \infty)$.

Similar to the case of MBAL-SPO with hard rejections, when Assumption 2.5.1 and the condition that $\phi(\epsilon) \sim \mathcal{O}(\sqrt{\epsilon})$ hold, we obtain sublinear label complexity of Algorithm 1 with soft rejections in Proposition 2.5.2.

Proposition 2.5.2 (Small label complexity for soft rejections)**.** *Suppose Assumptions 2.3.1 and 2.5.1 hold with* $|\mathcal{H}| \leq N_1$. *Suppose there exists a constant* $C_\phi > 0$ *such that Assumption 2.3.1.(2) holds with* $\phi(\epsilon) = C_\phi \cdot \sqrt{\epsilon}$. *Set* $\tilde{p} \leftarrow T^{-\frac{\kappa}{2(\kappa+2)}}$ *and* $H_t \leftarrow \mathcal{H}$ *for all* t, *and* b_t, r_t *the same values as Theorem 2.4.3. For a fixed* $\delta \in (0, 1]$, *the following guarantees hold simultaneously with probability at least* $1 - \delta$ *for all* $T \geq 1$*:*

- *The excess surrogate risk satisfies* $R_\ell(h_T) - R_\ell^* \leq \tilde{\mathcal{O}}\left(T^{-\frac{1}{2(\kappa+2)}}\right)$.
- *The excess SPO risk satisfies* $R_{\text{SPO}}(h_T) - R_{\text{SPO}}^* \leq \tilde{\mathcal{O}}\left(T^{-\frac{\kappa}{2(\kappa+2)}}\right)$.
- *The expectation of the number of labels acquired, conditional on the above guarantee on the excess surrogate risk, is at most* $\tilde{\mathcal{O}}\left(T^{1-\frac{\kappa}{2(\kappa+2)}}\right)$ *for* $\kappa > 0$.

In Proposition 2.5.2, the larger the parameter in near-degeneracy condition κ is, the smaller the label complexity will be. We observe that in Proposition 2.5.2, when $\tilde{p} = T^{-\frac{\kappa}{2(\kappa+2)}}$, the excess surrogate risk converges to zero at rate $\tilde{\mathcal{O}}(T^{-\frac{1}{\kappa+2}})$, which is slower than the typical learning rate of supervised learning, which is $\mathcal{O}(T^{-1/2})$. In the next section, we demonstrate that the excess surrogate risk can be reduced to $\tilde{\mathcal{O}}(T^{-1/2})$ under some further conditions.

2.5.2 Small Label Complexity with Soft Rejections.

In this section, we show that under certain conditions, the convergence rate of excess surrogate risk under soft rejection is $\tilde{\mathcal{O}}(T^{-1/2})$, which is the same as standard supervised learning (except for logarithmic factors). To achieve this rate, we allow $\tilde{p}$ to change dynamically, denoted as $\tilde{p}_t$. The soft rejection probability $\tilde{p}_t$ varies depending on the observed feature x_t at each iteration t. This adaptive approach ensures that the soft-rejection probability is not fixed to a small value, enabling the active learning algorithm to converge more quickly.

Particularly, we set $\tilde{p}_t = \max\{T^{-\frac{\kappa}{2(\kappa+2)}}, \alpha_1 \|h_T(x) - h^*(x)\|\}$, where α_1 is a constant. Proposition 2.5.3 shows that the excess surrogate risk of active learning, $R_\ell(h_t) - R_\ell(h^*)$ converges to zero at rate $\tilde{\mathcal{O}}(T^{-1/2})$, when $\tilde{p} > 0$.

Proposition 2.5.3. *Suppose that there exists a constant $C_\phi > 0$ such that Assumption 2.3.1 holds with $\phi(\epsilon) = C_\phi \cdot \sqrt{\epsilon}$. Suppose Assumption 2.5.1 holds with $\kappa = 1$. Suppose that the surrogate loss function $\ell(\cdot, c)$ is Lipschitz for any given $c \in \mathcal{C}$. Let $\tilde{p}_t \leftarrow \max\{T^{-\frac{\kappa}{2(\kappa+2)}}, \alpha_1 \|h_T(x) - h^*(x)\|\}$ for some constant $\alpha_1 > 0$. For some small $\delta \in (0, 1]$, consider Algorithm 1 under the same settings as Theorem 2.4.3. Then with probability at least $1 - \delta$, we have that $R_\ell(h_T) - R^*_\ell \leq \tilde{\mathcal{O}}(T^{-1/2})$ and $\mathbb{E}[n_t] \leq \tilde{\mathcal{O}}(T^{1-\frac{\min\{\kappa,1\}}{2(\kappa+2)}})$.*

Proposition 2.5.3 implies that for the excess risk of the surrogate function, our active learning algorithm achieves the same order as the supervised learning. However, compared to supervised learning, active learning algorithms acquire much fewer labels, which is at most $\tilde{\mathcal{O}}(T^{1-\frac{\min\{\kappa,1\}}{2(\kappa+2)}})$. In illustration, when the near-degeneracy condition holds with $\kappa = 1$, the label complexity of MBAL-SPO is $\tilde{\mathcal{O}}(T^{5/6})$ in Proposition 2.5.2. Therefore, the MBAL-SPO can achieve the same order of surrogate risk with a smaller number of acquired labels.

2.6 Examples of ϕ Functions and Upper Bound for η

The existence of non-trivial ϕ and Ψ depends on the distribution $\mathcal{D}$ and the feasible region S. In this section, we examine the case where we use the SPO+ loss as the surrogate loss function given the norm $\|\cdot\|$ as the ℓ_2 norm. We first present two special cases, polyhedral and strongly convex feasible regions, for which we can characterize the function ϕ. We then present sufficient conditions on the distribution $\mathcal{D}$ so that we can ultimately bound the label complexity. For simplicity, we use $\mathbb{P}(c|x)$ to denote the probability density function of c conditional on x. To study the pointwise error as needed in Assumption 2.3.1, we make a

recoverability assumption. Assumption 2.6.1 holds for linear hypothesis classes when the features have nonsingular covariance and for certain decision tree hypothesis classes when the density of features is bounded below by a positive constant (Hu, Kallus, and Mao, 2022).

Assumption 2.6.1 (Recoverability). *There exists $\varkappa > 0$ such that for all $h \in \mathcal{H}$, $h^* \in \mathcal{H}^*$, and almost all $x' \in \mathcal{X}$, it holds that*

$$\|h(x') - h^*(x')\|^2 \leq \varkappa \cdot \mathbb{E}\left[\|h(x) - h^*(x)\|^2\right].$$

Assumption 2.6.1 provides an upper bound of the pointwise error from the bound of the expected error. It implies that the order of pointwise error is no larger than the order of the expected error.

Polyhedral feasible region. First, we consider the case where the feasible region S is a polyhedron. Let $\mathcal{P}_{\text{cont, symm}}$ denote the class of joint distributions $\mathcal{D}$ such that $\mathbb{P}(\cdot|x)$ is continuous on $\mathbb{R}^d$ and is centrally symmetric with respect to its mean for all $x \in \mathcal{X}$. Following Theorem 2 in H. Liu and Grigas (2021), for given parameters $M \geq 1$ and $\alpha, \beta > 0$, let $\mathcal{P}_{M,\alpha,\beta}$ denote the set of all $\mathcal{D} \in \mathcal{P}_{\text{cont, symm}}$ such that for all $x \in \mathcal{X}$ and $\bar{c} = \mathbb{E}[c|x]$, there exists $\sigma \in [0, M]$ satisfying $\|\bar{c}\|_2 \leq \beta\sigma$ and $\mathbb{P}(c|x) \geq \alpha \cdot \mathcal{N}(\bar{c}, \sigma^2 I)$ for all $c \in \mathbb{R}^d$. Let D_S denote the diameter of S and define a "width constant" d_S associated with S by $d_S := \min_{v \in \mathbb{R}^d : \|v\|_2 = 1} \left\{\max_{w \in S} v^T w - \min_{w \in S} v^T w\right\}$. Notice that $d_S > 0$ whenever S has a non-empty interior.

Lemma 2.6.1 (Example of ϕ). *Given $\|\cdot\|$ as the ℓ_2 norm, suppose that Assumption 2.6.1 holds and the feasible region S is a bounded polyhedron. Define $\Xi_S := (1 + \frac{2\sqrt{3}D_S}{d_S})^{1-d}$. Suppose the hypothesis class $\mathcal{H}$ is well-specified, i.e., $h^*(x) = \mathbb{E}[c|x]$, for all $x \in \mathcal{X}$. When the distribution $\mathcal{D} \in \mathcal{P}_{M,\alpha,\beta}$, then it holds that for almost all $x \in \mathcal{X}$,*

$$R_{\text{SPO+}}(h) - R_{\text{SPO+}}(h^*) \leq \epsilon \Rightarrow \|h(x) - h^*(x)\|^2 \leq \varkappa \frac{8\sqrt{2\pi}\rho(\hat{\mathcal{C}})e^{\frac{3(1+\beta^2)}{2}}}{\alpha\Xi_S} \cdot \epsilon.$$

Lemma 2.6.1 indicates that $\phi(\epsilon) \leq \mathcal{O}(\sqrt{\epsilon})$.

Strongly-convex feasible region. Next, we consider the case where the feasible region S is a level-set of a strongly-convex and smooth function. In the spirit of Definition 4.1 in H. Liu and Grigas (2021), we consider two related classes of rotationally symmetric distributions with bounded conditional coefficient of variation. These distribution classes are formally defined in Definition 2.6.1 below, and include the multi-variate Gaussian, Laplace, and Cauchy distributions as special cases.

Definition 2.6.1 ($\mathcal{P}_{\beta_1,\beta_2}$ distribution). *We define $\mathcal{P}_{\text{rot symm}}$ as the class of joint distributions $\mathcal{D}$ with conditional rotational symmetry in the norm $\|\cdot\|$, namely for all $x \in \mathcal{X}$ and $\bar{c} = \mathbb{E}[c|x]$,*

there exists a function $q(\cdot) : [0, \infty) \to [0, \infty)$ *such that* $\mathbb{P}(c|x) = q(\|c - \bar{c}\|)$. *For constants* $\beta_1, \beta_2 \in (0, 1)$, *define*

$$\mathcal{P}_{\beta_1,\beta_2} := \left\{\mathcal{D} \in \mathcal{P}_{\text{rot symm}} : \textit{For any } c_1 \in \mathbb{R}^d, \mathbb{P}_{c|x}(0 \le c^T c_1 \le \beta_1 \|c_1\|\|c\|) \ge \beta_2\right\}.$$

□

Lemma 2.6.2 provides a bound for the pointwise error for strongly-convex feasible regions under the class of distributions in $\mathcal{P}_{\beta_1,\beta_2}$ specified in Definition 2.6.1.

Lemma 2.6.2 (Example of ϕ)**.** *Given* $\|\cdot\|$ *as the* ℓ_2 *norm, let* $f : \mathbb{R}^d \to \mathbb{R}$ *be a* μ_S*-strongly convex and* L_S*-smooth function for some* $L_S \ge \mu_S > 0$. *Suppose that the feasible region* S *is defined by* $S = \{w \in \mathbb{R}^d : f(w) \le r\}$ *for some constant* $r > f_{\min} := \min_w f(w)$. *Suppose that Assumption 2.6.1 holds and that the hypothesis class* $\mathcal{H}$ *is well-specified, i.e.,* $h^*(x) = \mathbb{E}[c|x]$, *for all* $x \in \mathcal{X}$. *When the distribution* $\mathcal{D} \in \mathcal{P}_{\beta_1,\beta_2}$, *then it holds that for almost all* $x \in \mathcal{X}$,

$$R_{\text{SPO+}}(h) - R_{\text{SPO+}}(h^*) \le \epsilon \Rightarrow$$
$$\|h(x) - h^*(x)\|^2 \le \frac{\varkappa \mu_S^2 r^{1/2}}{2^{1/2}\beta_2 L_S^{5/2}} \min\left\{\frac{2(1-\beta_1^2)}{\rho(\mathcal{C}, \hat{\mathcal{C}})}, \frac{\sqrt{17 + 8\beta_1} - 1 - 4\beta_1}{4\rho(\mathcal{C}, \hat{\mathcal{C}})}\right\}^{-1} \cdot \epsilon.$$

Lemma 2.6.2 implies that for the strongly-convex feasible region, if the distribution $\mathcal{D} \in \mathcal{P}_{\beta_1,\beta_2}$, we have $\phi(\epsilon) \le \mathcal{O}(\sqrt{\epsilon})$. Since Theorem 2.4.1 also requires Assumption 2.4.1 holds, in Lemma 2.6.3 below, we provide the conditions that Assumption 2.4.1 holds for the SPO+ loss.

Lemma 2.6.3 (Existence of η for $\ell_{\text{SPO+}}$)**.** *Given* $\|\cdot\|$ *as the* ℓ_2 *norm, let* $f : \mathbb{R}^d \to \mathbb{R}$ *be a* μ_S*-strongly convex and* L_S*-smooth function for some* $L_S \ge \mu_S > 0$. *Suppose that the feasible region* S *is defined by* $S = \{w \in \mathbb{R}^d : f(w) \le r\}$ *for some constant* $r > f_{\min} := \min_w f(w)$. *Suppose the hypothesis class* $\mathcal{H}$ *is well-specified, i.e.,* $h^*(x) = \mathbb{E}[c|x]$, *for all* $x \in \mathcal{X}$. *Suppose distribution* $\mathcal{D} \in \left\{\mathcal{D} \in \mathcal{P}_{\text{rot symm}} : \mathbb{P}_{c|x}(\|c\| \ge \beta) = 1, \textit{for all } x \in \mathcal{X}\right\}$, *for some positive* $\beta > 0$. *Then,* $\ell_{\text{SPO+}}(\cdot, c)$ *satisfies that for all* $x \in \mathcal{X}$, $c_1 \in \mathcal{C}$ *and* $h^* \in \mathcal{H}^*$,

$$|\mathbb{E}[\ell_{\text{SPO+}}(c_1, c) - \ell_{\text{SPO+}}(h^*(x), c)|x]| \le \frac{L_S^2 \rho(\mathcal{C})\sqrt{r - f_{\min}}}{\sqrt{2}\mu_S^{1.5}} \frac{4}{\beta} \|c_1 - h^*(x)\|^2.$$

Lemma 2.6.3 shows that when the feasible region is strongly convex and the hypothesis class is well-specified, and when the distribution of cost vectors is separated from the origin with probability 1, then η in Assumption 2.4.1 is finite for the SPO+ loss.

In conclusion of Section 2.6, it is worth noting that while our analysis in this section focused on the SPO+ loss function, similar results can be obtained for commonly used loss functions such as squared ℓ_2 norm loss. For example, under some noise conditions, we can also obtain $\phi(\epsilon) \sim \sqrt{\epsilon}$ and the upper bound for η under the squared ℓ_2 norm loss.

2.7 Numerical Experiments

In this section, we present the results of numerical experiments in which we empirically examine the performance of our proposed margin-based algorithm (Algorithm 1) under the SPO+ surrogate loss. We use the shortest path problem and personalized pricing problem as our exemplary problem classes. For both problems, we use (sub)gradient descent to minimize the SPO+ loss function in the MBAL-SPO algorithm. We set $\tilde{p} \leftarrow 10^{-5}$ and set $H_t \leftarrow \mathcal{H}$ according to Theorem 2.4.3. The norm $\|\cdot\|$ is set as the ℓ_2 norm. In both problems, to calculate the distance to the degeneracy, we use the result of Theorem 8 in El Balghiti et al. (2022), which was stated in Equation (2.4). We set the function ϕ as the square root function, which is used directly in the setting of the sequence $\{b_t\}$. The numerical experiments were conducted on a Windows 10 Pro for Workstations system, with an Intel(R) Xeon(R) Silver 4114 CPU @ 2.20GHz 20 cores.

2.7.1 Shortest Path Problem

We first present the numerical results for the shortest path problem. We consider a 3×3 (later also a 5×5) grid network, where the goal is to go from the southwest corner to the northeast corner, and the edges only go north or east. In this case, the feasible region S is composed of network flow constraints, and the cost vector c encodes the cost of each edge.

Data generation process. Let us now describe the process used to generate the synthetic experimental data. The dimension of the cost vector d is 12, corresponding to the number of edges in the 3×3 grid network. The number of features p is set to 5. The number of distinct paths is 6. Given a coefficient matrix $B \in \mathbb{R}^{d \times p}$, the training data set $\{(x_i, c_i)\}_{i=1}^{n}$ and the test data set $\{(\tilde{x}_i, \tilde{c}_i)\}_{i=1}^{n_{\text{test}}}$ are generated according to the following model.

1. First, we identify six vectors $\mu_j \in \mathbb{R}^p$, $j = 1, ..., 6$, such that the corresponding cost vector $B\mu_j$ is far from degeneracy, that is, the distance to the closest degenerate cost vector $\nu_S(B\mu_j)$ is greater than some threshold, and the optimal path under the cost vector $B\mu_j$ is the path j.

2. Each feature vector $x_i \in \mathbb{R}^p$ is generated from a mixed distribution of six multivariate Gaussian distributions with equal weights. Each multivariate Gaussian distribution follows $N(\mu_j, \sigma_m^2 I_p)$, where the variance σ_m^2 is set as $1/9$.

3. Then, the cost vector c_j is generated according to $c_j = \left[1 + (1 + b_j^T x_i/\sqrt{p})^{\text{deg}}\right] \epsilon_j$, for $j = 1, ..., d$, where b_j is the j^{th} row of the matrix B. The degree parameter deg is set as 1 in our setting and ϵ_j is a multiplicative noise term, which is generated independently from a uniform distribution $[1 - \bar{\epsilon}, 1 + \bar{\epsilon}]$. Here, $\bar{\epsilon}$ is called the noise level of the labels.

To determine the coefficient matrix B, we generate a random candidate matrix $\tilde{B}$ multiple times, whose entries follow the Bernoulli distribution (0.5), and pick the first B such that μ_j exists in Step 1 for each $j = 1, ..., 6$. The size of the test data set is 1000 sample points. In the context of our margin-based algorithm, we set $r_t = \sqrt{[d \times \ln(t) + \ln(1/\delta)]/t}$, where t is the iteration counter, $d = 12$ is the dimension of the cost vector, and δ is set as 10^{-7}.

According to Proposition 2.5.2, we set $b_t = 0.5\sqrt{r_t}$. The running time for one single run on a 3×3 grid to acquire 25 labels is about 10 minutes for the margin-based algorithm.

Figure 2.3 shows our results for this experiment. Excess SPO risks during the training process for MBAL-SPO and supervised learning are shown in the left plot of Fig. 2.3. The x-axis shows the number of labeled samples and the y-axis shows the log-scaled excess SPO risk on the test set. The results are from 25 trials, and the error bar in Figure 2.3 is an 85% confidence interval. We observe that as more samples are labeled, the margin-based algorithm performs better than supervised learning, as expected. Compared to supervised learning, the margin-based algorithm achieves a significantly lower excess SPO risk when the number of labeled samples is around 25.

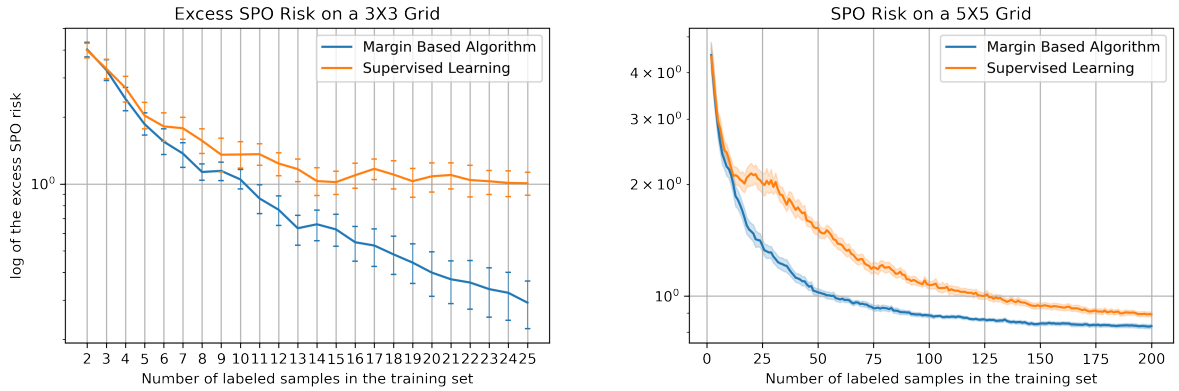


Figure 2.3: Risk on the test set during the training process in 3×3 grid, and 5×5 grid.

The margin-based algorithm has good scalability as long as the calculation of the distance to the degeneracy ν_S is fast, for example, in the case of relatively simple polyhedral sets. To further examine the performance of the margin-based algorithm on a larger-scale problem, we conduct a numerical experiment in a 5×5 grid network in the right plot of Figure 2.3, again shown with an 85 % confidence interval. We see that although both algorithms converge to the same optimal SPO risk level, the margin-based algorithm has a much faster learning rate than supervised learning and can achieve a lower SPO risk even after 200 labeled samples.

In Section 2.7.3, we further examine the impacts of the scale of parameters, r_t and b_t, on the number of labels and the SPO risk during the training process, which demonstrates the robustness of the SPO risk with respect to the scales of these parameters. In Section 2.7.4, we include more results in which we change the noise levels and variance of the features when generating the data. This verifies the advantages of our algorithms under various conditions.

2.7.2 Personalized Pricing Problem

In this section, we present numerical results for the personalized pricing problem. Suppose that we have three types of items, indexed by $j = 1, 2, 3$. We have three candidate prices for these three items, which are \$60, \$80, and \$90. Therefore, in total, we have $3^3 = 27$ possible combinations of prices. Suppose that the dimension of the features of the customers is $p = 6$. When a customer is selected to survey, their answers will reveal the purchase probability for all three items at all possible prices. These purchase probabilities are generated on the basis of an exponential function of the form $\mathcal{O}(e^{-p})$. We add additional price constraints between products, such that the first item has the highest price, and the third item has the lowest price. Please see the details in Section 2.7.5.

Because there are three items and three candidate prices, the dimension of the cost vector $d_j(p_i)$ is 9. Therefore, our predictor $h(x)$ is a mapping from the feature space $\mathcal{X} \subseteq \mathbb{R}^6$ to the label space $\mathcal{C} \subseteq [0, 1]^9$. We assume that the predictor is a linear function, so the coefficient of $h(x)$ is a $(6 + 1) \times 9$ matrix, including the intercept. Unlike the shortest path problem which can be solved efficiently, the personalized pricing problem is NP-hard in general due to the binary constraints. In our case, since the dimensions of products and prices are only three, we enumerate all the possible solutions to determine the prices with the highest revenue.

The test set performance is calculated on 1000 samples. In MBAL-SPO, we set $r_t = 250\sqrt{[d \times \ln(t) + \ln(1/\delta)]/t}$, where t is the iteration counter, d is the dimension of the cost vector, which is 9, and δ is set as 10^{-7}. According to Proposition 2.5.2, we set $b_t = 0.5 \times \sqrt{r_t}$. The scales of r_t and b_t are selected by the rules discussed at the end of Section 2.4.4. The excess SPO risks of MBAL-SPO and supervised learning on the test set as the number of acquired labels increases are shown in Figure 2.4. The results are from 25 simulations, and the error bars in Figure 2.4 represent an 85% confidence interval. Notice that the demand function is in an exponential form but our hypothesis class is linear, so the hypothesis class is misspecified. The results in Figure 2.4 show that MBAL-SPO achieves a smaller excess SPO risk than supervised learning even when the hypothesis class is misspecified.

2.7.3 Setting Parameters in the Algorithm

Here we discuss how to set the values of the parameters r_t and b_t in margin-based algorithm in practice. In general, these values depend on $\mathcal{D}$, the budget of the labeled samples, and the performance that we would like to achieve. Setting r_t and b_t, to be large numbers makes our active learning algorithm the same as supervised learning. Setting them as smaller numbers will make our algorithms less conservative and more sensitive to the first several samples.

To further illustrate the impact of the scale of these values, we run the following experiments by changing the scales of these parameters. In the margin-based algorithm, we set the value of r_t to $\sqrt{[d \times \ln(t) + \ln(1/\delta)]/t}$, where t is the number of samples, d is the dimension of cost vector, which is 12, and δ is set as 10^{-7}. According to Proposition 2.5.2, we set $b_t = \texttt{slackness} \times \sqrt{r_t}$, where `slackness` a parameter we will tune. The plot on the left of Figure 2.5 shows the ratio of labeled samples to total samples in the first 30 samples as the

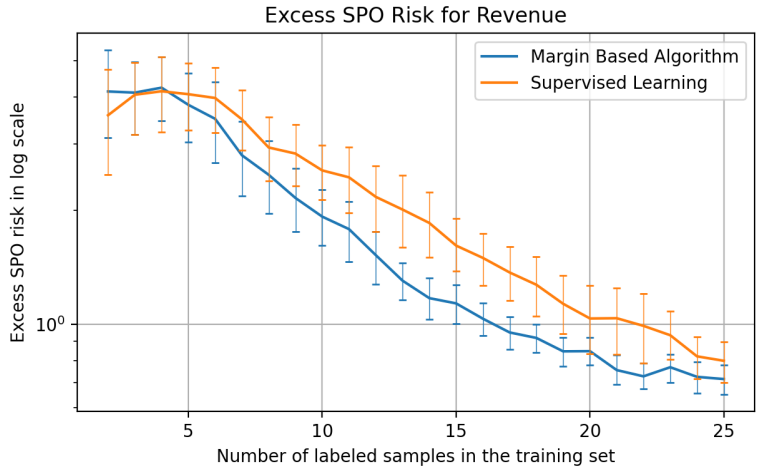


Figure 2.4: Excess test set risk during the training process in personalized pricing.

slackness value varies. We see that the larger slackness is, the more samples are labeled. The right plot in Figure 2.5 further shows the value of excess SPO risk as the value of slackness changes when the number of labeled samples is seven. It shows that the excess SPO risk is quite robust to the value of slackness. In other words, the value of slackness has little impact on the excess SPO risk given the same number of labeled samples, though the value of slackness affects the ratio of the labeled samples.

In practice, to find the set the scale for b_t and r_t, we can refer to the rules discussed at the end of Section 2.4.4, where we set a "burn in" period of $\tilde{T}$ iterations that acquires all labels during the first $\tilde{T}$ iterations. Then, we can use the distribution of values $\{\nu_S(h_T(x_t))\}_{t=1}^{\tilde{T}}$ to inform the value of b_T. For example, if we want to reduce the number of labels by 50%, compared to supervised learning, we can set the scale of b_T as the median of $\{\nu_S(h_T(x_t))\}_{t=1}^{\tilde{T}}$.

We also change the value of the minimum label probability $\tilde{p}$ in the soft rejection to see its impact on the performance. Figure 2.6 shows the percentage of labeled samples in the first 30 samples, and the excess SPO risk when the number of labeled samples is 10.

Figure 2.6 shows that the minimum label probability $\tilde{p}$ has no significant impact on the excess SPO risk. Intuitively, when $\tilde{p}$ is larger, the percentage of labeled samples is larger. In practice, we can set $\tilde{p}$ as a very small positive number that is close to zero.

2.7.4 Additional Results of Numerical Experiments.

To assess the performance of our active learning algorithm under different noise levels, we change the variance of features and the noise level of labels when generating the data and demonstrate the results in Figures 2.7 and 2.8.

Figures 2.7 and 2.8 show that when the variance of the features and the noise level of

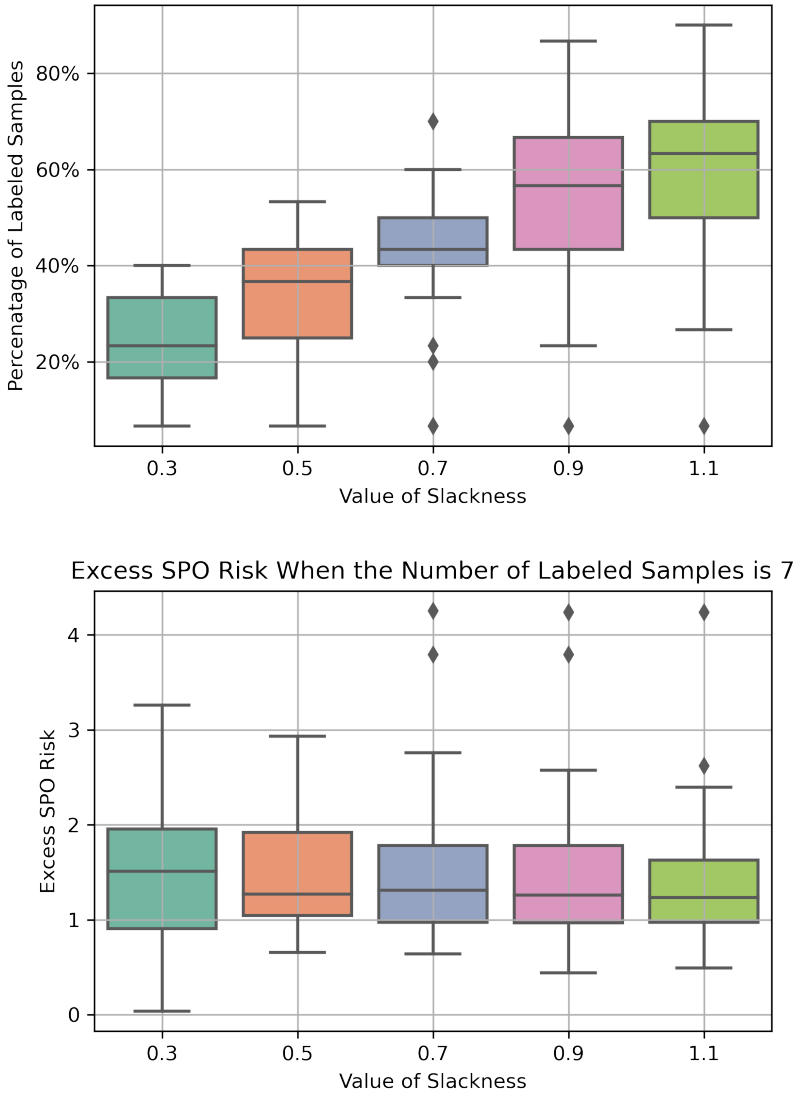


Figure 2.5: Performance under different settings of slackness in MBAL-SPO

the labels are small, both active learning and supervised learning have close performance. When the variance of features or the noise level of labels is large, our proposed active learning methods perform better than supervised learning.

Recall that the cost vector is gerenated according to $c_j = \left[1 + (1 + b_j^T x_i/\sqrt{p})^{\text{deg}}\right] \epsilon_j$. Next, we further show the result when changing the degree of the model. When the degree is not one, the true model is not contained in our hypothesis class. The results in Figure 2.9 show

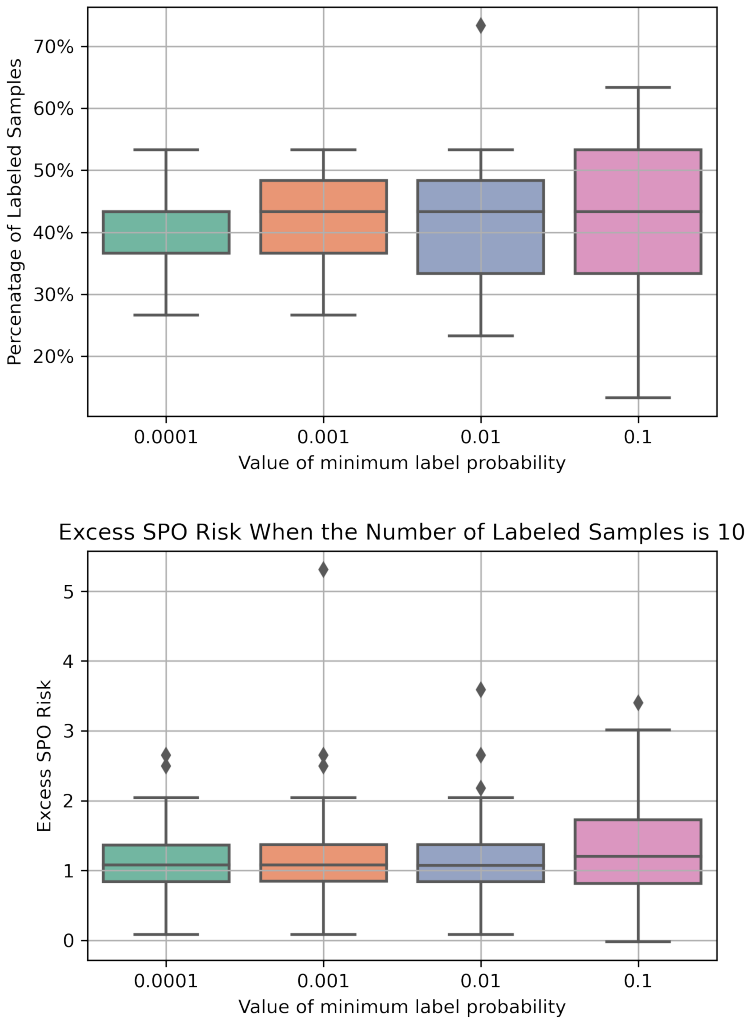


Figure 2.6: Performance under different settings of $\tilde{p}$

that when the model has a higher degree, the training process has a higher excess SPO risk at the beginning of the process.

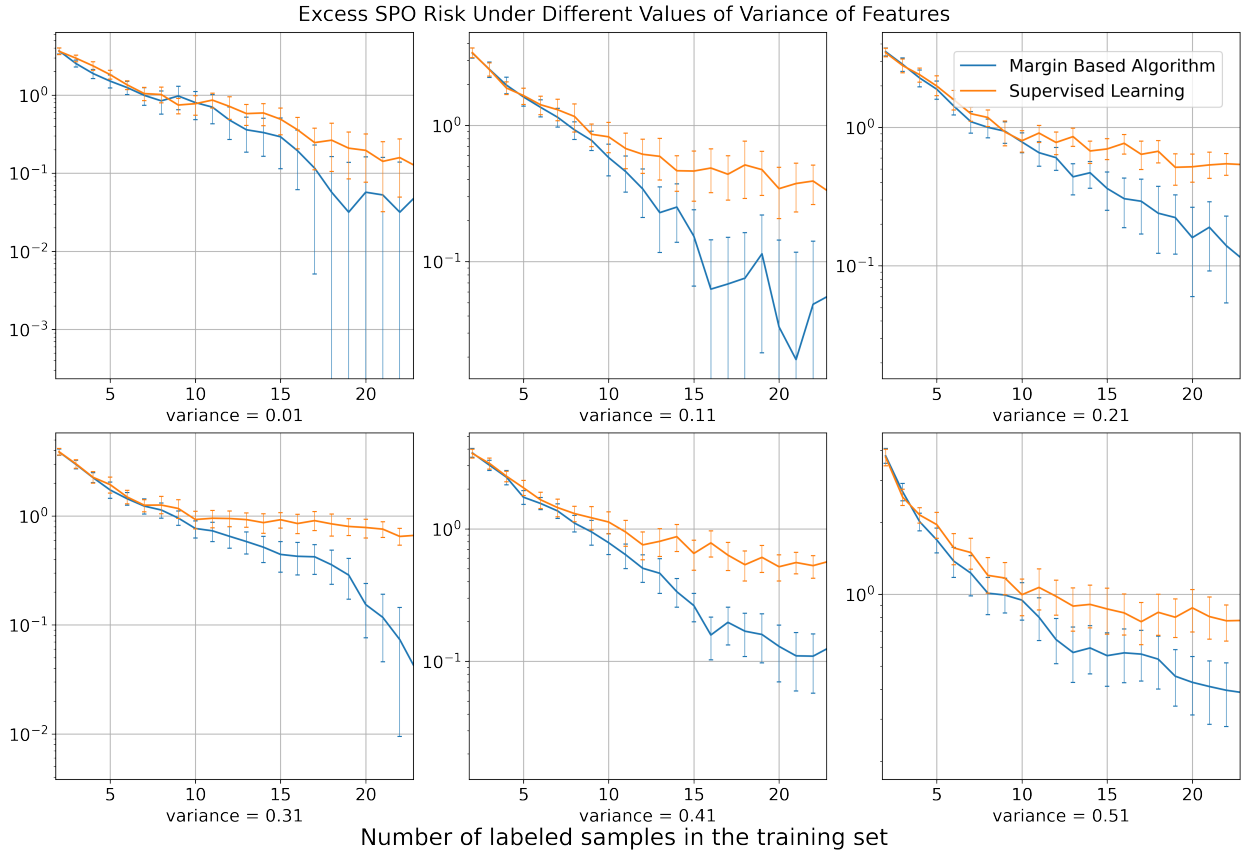


Figure 2.7: Excess SPO risk during the training process under different variance of features.

2.7.5 Data Generation for Personalized Pricing

In this section, we provide the parameter values for generating synthetic data in the personalized pricing experiment. Given a coefficient vector $B_j \in \mathbb{R}^5$ and $A_j \in \mathbb{R}^5$, the demand function for item j is generated as $d_j(p_i) = \epsilon e^{B_j^T X + A_j^T X p_i}$. Here, ϵ is a noise term drawn from a uniform distribution on $[1-\bar{\epsilon}, 1+\bar{\epsilon}]$. We set $\bar{\epsilon} = 0.1$. $A_j^T X$ can be viewed as the price elasticity. The customer feature vector is drawn from a mixed Gaussian distribution with seven different centers μ_k. The value of these centers μ_k, $k = 1, 2, ..., 7$ and the value of A_j and B_j, $j = 1, 2, 3$ are carefully chosen so that $h^*(X)$ is not a degenerate cost vector for any μ_k, $k = 1, 2, ..., 7$. Please find the value of these parameters at the end of this section. The variance of the feature for each Gaussian distribution is set as 0.01^2, which is on the same scale as the features.

We further have the following monotone constraints for the prices of these three items. Let the decision variable $w_{i,j}$ indicate whether price i is selected for item j. Then, the constraints

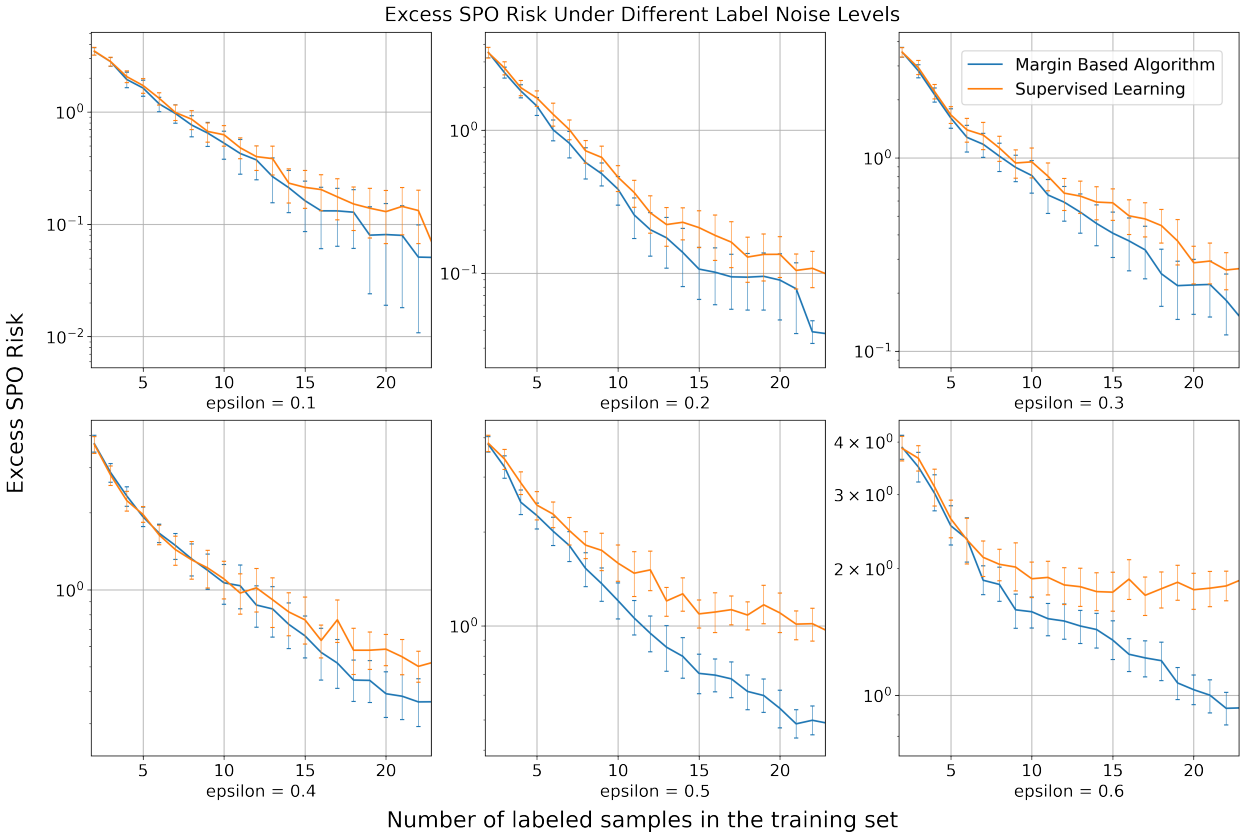


Figure 2.8: Excess SPO risk during the training process under different noise levels.

are as follows.

$$w_{1,j} + w_{2,j} + w_{3,j} \leq 1, \quad j = 1, 2, 3 \tag{2.5a}$$
$$w_{2,1} \leq w_{2,2} + w_{3,2} \tag{2.5b}$$
$$w_{3,1} \leq w_{3,2} \tag{2.5c}$$
$$w_{2,2} \leq w_{2,3} + w_{3,3} \tag{2.5d}$$
$$w_{3,2} \leq w_{3,3} \tag{2.5e}$$
$$w_{i,j} \in \{0, 1\}, \quad i, j = 1, 2, 3$$

(2.5a) requires each item can only select one price point. (2.5b) and (2.5c) require that the price of item 2 be no less than the price of item 1. (2.5d) and (2.5e) require that the price of item 3 be no less than the price of item 2.

Since the purchase probability is $d_j(p_i) = \epsilon \exp(B_j^T X + A_j^T X p_i)$, we need to specify the following parameters for generating the purchase probability given the feature X: A_j and

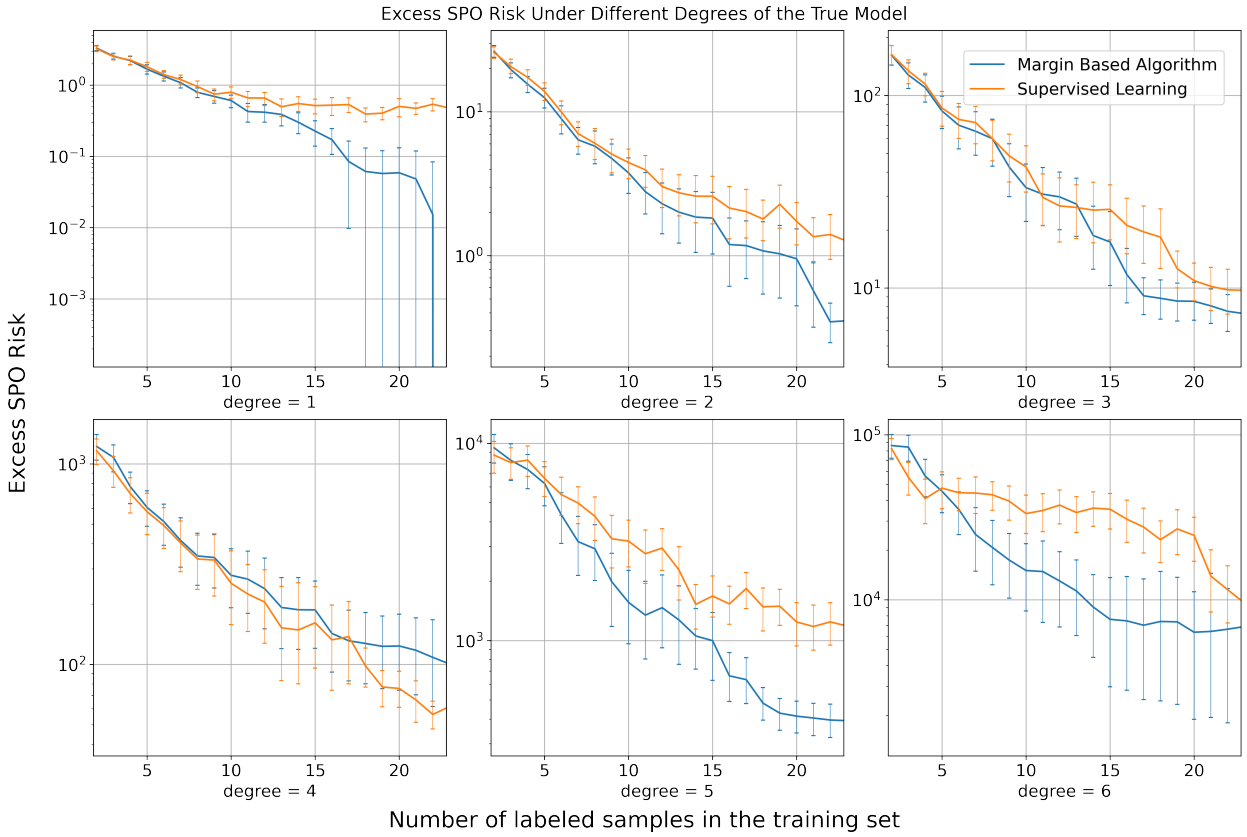


Figure 2.9: Excess SPO risk during the training process under different noise levels.

B_j, $j = 1, 2, 3$. The feature vector is from a mixed Gaussian distribution with seven centers. The optimal prices of three items for these seven centers are $(\$60, \$60, \$60)$, $(\$60, \$80, \$90)$, $(\$90, \$90, \$90)$, $(\$80, \$80, \$80)$, $(\$60, \$60, \$80)$, $(\$80, \$90, \$90)$, and $(\$60, \$60, \$90)$ respectively. To generate such centers, we consider the following values for X, A_j and B_j. Define $a_1 = -0.0202733, b_1 = -1.19155$, $a_2 = -0.0133531, b_2 = -1.45748$, $a_3 = -0.00540672, b_3 = -1.22819$. Then, we set

$$A_1 = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \\ 0 \\ 0 \end{bmatrix}, A_2 = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 1 \\ 0 \end{bmatrix}, A_3 = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 1 \end{bmatrix}, B_1 = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}, B_2 = \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}, B_3 = \begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \\ 0 \\ 0 \end{bmatrix}.$$

We set the centers of Gaussian distribution for the feature vectors as

$$\mu_1 = \begin{bmatrix} b_1 \\ b_1 \\ b_1 \\ a_1 \\ a_1 \\ a_1 \end{bmatrix}, \mu_2 = \begin{bmatrix} b_1 \\ b_2 \\ b_3 \\ a_1 \\ a_2 \\ a_3 \end{bmatrix}, \mu_3 = \begin{bmatrix} b_3 \\ b_3 \\ b_3 \\ a_3 \\ a_3 \\ a_3 \end{bmatrix}, \mu_4 = \begin{bmatrix} b_2 \\ b_2 \\ b_2 \\ a_2 \\ a_2 \\ a_2 \end{bmatrix}, \mu_5 = \begin{bmatrix} b_1 \\ b_1 \\ b_2 \\ a_1 \\ a_1 \\ a_2 \end{bmatrix}, \mu_6 = \begin{bmatrix} b_2 \\ b_3 \\ b_3 \\ a_2 \\ a_3 \\ a_3 \end{bmatrix}, \mu_7 = \begin{bmatrix} b_1 \\ b_1 \\ b_3 \\ a_1 \\ a_1 \\ a_3 \end{bmatrix}.$$

2.8 Conclusions and Future Directions

Our work develops the first active learning algorithms in the predict-then-optimize framework. We consider the SPO loss function and its tractable surrogate loss functions and provide a practical margin-based active learning algorithm (MBAL-SPO). We provide two versions of MBAL-SPO and develop excess risk guarantees for both versions. Furthermore, we provide upper bounds on the label complexity of both versions and show that the label complexity can be better than the supervised learning approach under some natural low-noise conditions. Our numerical experiments also demonstrate the practical value of our proposed algorithm. There are several intriguing future directions. Since directly minimizing the SPO loss function is challenging, one valuable direction is to design active learning algorithms in situations where we can minimize the SPO loss function approximately. While our work focuses on stream-based active learning in the predict-then-optimize framework, it is also worthwhile to consider pool-based active learning, where all feature vectors are revealed at once before training, in the future.

Chapter 3

Feature-Dependent Value of One Data Point

3.1 Introduction

Understanding customer preferences is a crucial component of operations management. For example, in assortment optimization, retailers strive to learn customer preferences in order to tailor the best possible assortment. Customer preferences can be indirectly gauged by analyzing their purchasing behaviors, but the most straightforward way is through active inquiry. A well-designed survey can provide valuable information about individual customer preferences, shedding light on the product utilities (ratings) for each customer. This information can be used to forecast future preferences. By leveraging these predictive models, retailers are empowered to make data-driven decisions and strategically create a product assortment that aligns with each customer's preferences.

Building such predictive models requires knowledge of the actual preference of customers, which are referred to as labels for the model. These true preferences are obtained through survey results at a cost, known as the label cost for each customer. The label cost includes the rewards given to each customer to fill out the survey, the labor cost to collect answers, and other expenses. As pointed out by Saar-Tsechansky, Melville, and Provost, 2009, "*without costly incentives, most consumers rarely provide this valuable feedback*". The terms "label cost" and "incentive" are interchangeably used to denote the cost incurred to uncover a customer's true preference. To reduce the label cost when building the prediction model, instead of providing fixed incentives to all customers for their feedback, the retailer can customize incentives for each customer. For instance, prominent online retailers like Amazon.com[1] may choose specific representative customers and offer them rewards as an incentive to complete the survey, in order to learn about customer preferences. Consequently, determining the appropriate personalized incentive for these representative customers, amidst potentially millions, becomes a crucial question when distributing surveys.

[1] `https://www.amazon.com/gp/help/customer/display.html?nodeId=G58ZQEGH2H25HAQR`

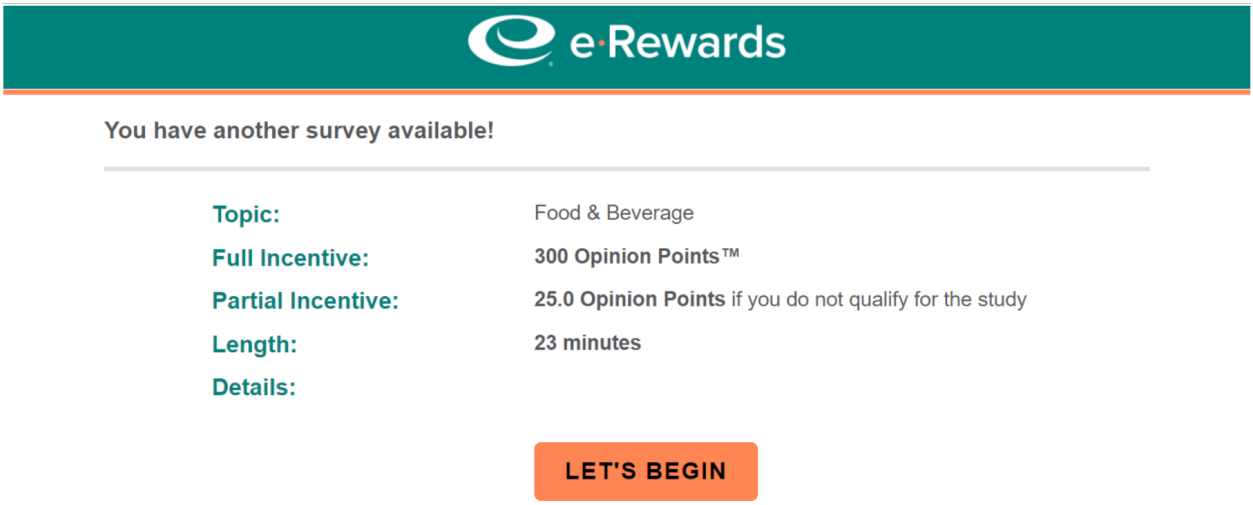


Figure 3.1: Personalized Rewards for completing the survey at e-Rewards platform

In our work, we focus on determining personalized incentives in a stream-based setting. In this process, each iteration starts with a customer's arrival, and the retailer observes the customer's characteristics, referred to as the "type" of this customer. Then the retailer offers some incentive to this customer for completing the survey. If the customer agrees to take the survey, she receives the incentive, and the retailer learns products' true utilities for her. However, if the customer declines the survey, no incentive is given. This process is referred to as the active label acquisition process with personalized incentives.

After the active label acquisition process, retailers build a prediction model for the utilities of all products based on the collected surveys. The evaluation of the prediction model is centered around our ultimate objective, which is to maximize the revenue generated from customized assortments. To quantify the performance of the current prediction model, instead of using the prediction error, we focus on the risk that is defined as the expected revenue loss when compared to the decisions made from the true model.

The label acquisition process has wide applications in the real world. For example, e-Rewards [2] is a research and survey platform that offers personalized incentives to customers for completing surveys. This platform has cooperated with various retailers to conduct market research.

For example, Figure 3.1 shows the screenshot of e-Rewards website after one customer completes a survey. It recommends one new available survey on food and beverages, which is worth 300 points for qualified customers. For the e-Reward platform, a critical question is how to determine the qualified customers and personalized incentives for each customer.

During the active label acquisition process, when determining the personalized incentive in the active label acquisition process, there exists a natural tradeoff. When the incentive

[2]https://www.e-rewards.com/en/

is insufficient, the customer has little probability of taking the survey and providing her preference, resulting in scant data. This lack of data may render the prediction model inaccurate and risky for assortment decisions. Conversely, if the incentive is too large, the cumulative label cost is large, which is unfavorable especially when the customer has little contribution to the revenue increase. Therefore, the long-term goal is to build a prediction model with a reduced risk, while the short-term goal is to minimize the label cost of each customer. To balance this tradeoff and determine the personalized incentive, a pivotal question to answer is: Given the current prediction model and a customer's type,

how much is the potential contribution to the ultimate goal, (i.e., revenue increase for the assortment optimization), of acquiring the true preference of ***this*** *particular data point?*

The answer to quantify the potential marginal contribution is termed the *value of one data point* in this chapter. Intuitively, when the value of one data point becomes smaller, we should offer a smaller incentive. However, evaluating the value of one data point is a challenging task because it involves two types of uncertainty. The first type of uncertainty is from the prediction error of the customer's preferences. The prediction error for each type depends on the quality and quantity of the current training set. For example, a higher concentration of similar types in the training set improves model accuracy for those types, reducing their value of one data point. The second type of uncertainty stems from the decision-making. Given the same uncertainty level of preferences, the confidence in distinguishing the optimal assortment from the sub-optimal assortment may vary. Consequently, the same prediction errors of the preferences may result in different possible risks. For example, if the model has enough confidence to distinguish the best assortment from the others, even though the prediction error of preferences is large, the value of one data point for this customer might be small.

Although the first type of uncertainty is addressed by most active learning (AL) algorithms to save the label cost, the second type of uncertainty, which involves regret due to the sub-optimal decisions (i.e., the possible revenue loss of sub-optimal assortment) is not addressed by the AL literature. However, it is essential to incorporate the second type of uncertainty when evaluating the value of one data point for one customer. For instance, if the current preference model, though not accurate, can already identify the best assortment for a given customer, improving the prediction accuracy for that customer's preference is unnecessary.

However, the second type of uncertainty is difficult to measure directly, due to the nonlinearity of the assortment optimization problem. To address this issue, we first study a problem with a linear objective, called the product selection problem whose primary goal is to provide the best personalized product recommendations for each customer to maximize expected customer satisfaction levels. Next, we utilize it to estimate the value of one data point in the assortment optimization problem. For both problems, we first consider a nonparametric model with finite discrete types of customers. Then we extend it to the cases allowing additional contextual information for each customer within one type.

Considering all these factors when offering personalized incentives, we propose an active label acquisition algorithm that determines personalized incentives based on the value of one data point. We demonstrate both theoretically and empirically that our personalized

incentive policy could achieve a much smaller comprehensive cost than the fixed incentive policies under various conditions. The contribution of our work is summarized as follows:

1. *Formulations*: Our work is the first to tackle active label acquisition in the context of assortment optimization and product selection. Compared to the traditional active learning algorithm that minimizes the prediction error, the goal of our active label acquisition algorithm is to minimize the risk that is defined on the objective of the decision-making problem (e.g., revenue).

2. *Concept:* To actively select customers to acquire their labels, we define a new concept called *value of one data point*, which captures the potential risk reduction of acquiring the label of one particular customer. We provide a theoretical feature-dependent upper bound for this term, which is useful in the active label acquisition process.

3. *Theory*: We derive non-asymptotic bounds for cumulative incentives, the risk of the model, and the comprehensive cost. By analyzing these bounds, we demonstrate theoretically that our personalized incentive policy could achieve a much smaller comprehensive cost than the fixed incentive policy under various settings. In particular, when the distribution of the utility vector satisfies some low-noise conditions, or the minimum incentive is positive, our algorithm can achieve a finite label cost.

4. *Insights*: Our theoretical results reveal some insights about the tradeoff between the risk of the model and the label cost. Firstly, when the minimum label cost is non-zero, (for example, there exists a minimum labor cost of collecting and analyzing a survey), the survey distribution process stops at some point. Secondly, this stop point of the survey distribution process does not necessarily prevent the risk of the prediction model from converging to zero. More specifically, when customer's preference distribution satisfies some low-noise conditions, the risk can still converge to zero. Thirdly, even if the minimum label cost is zero, the cumulative label cost does not necessarily go to infinity. Instead, it can converge to a finite value when the low-noise conditions are satisfied. Fourthly, if the optimal assortment is easier to distinguish from the suboptimal assortment, both the cumulative label cost and the risk of the model will get smaller.

5. *Numerical performance*: Using both real-world and synthetic data, our numerical experiments on the product selection and assortment optimization problems demonstrate the advantages of our personalized incentive policy over the fixed incentive policies. The results show that our personalized incentive policy requires much less label cost when achieving the same level of risk, compared to the fixed incentive policy.

In this chapter, we first define the value of one data point in Section 3.2. Then, we consider how to estimate the upper bounds for the value of one data point in the product selection and the personalized assortment optimization problem in Sections 3.3 and 3.4. We show that the upper bound for the value of one data point in the personalized assortment

optimization can be established on the analysis in the personalized product selection. These upper bounds provide significant insights to the importance of each customer's preference. To demonstrate the practical value of our upper bounds, we consider the incentive design problem during the customer survey process in Section 3.5. For both problems, we demonstrate that under some natural low-noise conditions, the total survey cost can be reduced significantly, while achieving the same levels of regret in Section 3.6. Next, in Section 3.7, we extend our algorithm to the case with additional contextual information within each type. In Section 3.10, we justify the practical performance of our personalized incentive policies using both real-world and synthetic data.

3.1.1 Literature Review

We summarize three streams of literature that are related to our research: *assortment optimization*, *active learning*, and *decision-focused learning* problems.

Assortment optimization. The assortment optimization problem studies how to determine the best set of products to display, in order to maximize the revenue, which plays an important role in economics, marketing, and operations management. One of the most popular and earliest choice models is multinomial logit choice (MNL) model. Under this model, when the parameters of the choice model are known, Talluri and Van Ryzin, 2004 propose an efficient revenue order method to determine the optimal assortment. When the capacity constraints of the assortment are considered, Qian Liu and Van Ryzin, 2008 provide a linear programming based approach to find the optimal assortment. Rusmevichientong, Shen, and Shmoys, 2010 further provide an efficient and simple algorithm to consider the capacity constraint. To further solve the large-scale assortment optimization problem efficiently, Bertsimas and Mišić, 2019 propose a mixed-integer optimization formulation and provide a specialized solution approach. For dynamic assortment problems under the MNL model with inventory constraints, Aouad, Levi, and Segev, 2018 propose a constant approximation algorithm. When the parameters of the MNL model need to be learned from data, Rusmevichientong, Shen, and Shmoys, 2010 propose an adaptive policy to maximize profit. In the online learning setting, different algorithms have been proposed to address the trade-off between exploration and exploitation and minimize the regret (e.g., Sauré and Zeevi, 2013; Cheung and Simchi-Levi, 2017; Agrawal et al., 2019; Agrawal et al., 2017; X. Chen and Yining Wang, 2017; Lei et al., 2022). Aouad, V. Farias, and Levi, 2021 study the assortment optimization problem under the consider-then-choose choice model. The assortment optimization problem under some general choice model, for example, the Markov chain choice model, has been addressed by J. B. Feldman and Topaloglu, 2017 and the capacity constrained problem was addressed by Désir, Goyal, Segev, et al., 2020 and S. Li et al., 2022. When customers can purchase multiple products at one time, Tulabandhula, Sinha, and Patidar, 2020; Lyu et al., 2021 propose different algorithms to find the optimal assortment. Désir, Goyal, Jagabathula, et al., 2021 further consider the assortment optimization problem under the mixture of Mallows choice model.

Active learning in customer survey. Following the ideas in the active learning, Zheng and Padmanabhan, 2006 and Saar-Tsechansky, Melville, and Provost, 2009 propose active label acquisition algorithms based on customers' features. However, all these literatures focus on the accuracy of the prediction model, instead of the cost of the downstream optimization problem. Krishnamurthy et al. (2017) and R. Gao and Saar-Tsechansky, 2020 consider cost-sensitive classification problems, where the misclassification cost depends on the true labels of the sample. Feng, 2020 and Yang and Feng, 2023 consider learning customer preferences by controlling the displayed products. All the above literature outputs the binary decisions, specifically, deciding whether to acquire labels. However, our work considers the personalized incentive, which is a continuous output. To determine the personalized incentive, our work evaluates the revenue contribution by estimating the potential revenue loss in the assortment optimization problem.

Decision-focused learning. Our work also contributes to the literature stream of predict-then-optimize, which incorporates the downstream optimization problem when training a prediction model. For example, Bertsimas and Kallus, 2020, Kao, Roy, and Yan, 2009, Elmachtoub and Grigas, 2022, Donti, Amos, and Kolter, 2017, Ho and Hanasusanto, 2019 and T. Zhu, Xie, and Sim, 2022 propose various frameworks to consider the downstream optimization problem. Particularly, Elmachtoub and Grigas, 2022 propose a smart predict-then-optimize (SPO) framework where the objective function in the optimization problem is linear and the uncertainty lies in the linear objective. Minimizing the SPO loss in supervised learning, which is nonconvex and non-Lipschitz, has been studied in several recent works. Elmachtoub and Grigas (2022) provides a surrogate loss function SPO+ and shows the consistency of this loss function. El Balghiti et al., 2019 considers the generalization error bounds of the SPO loss function. Ho-Nguyen and Kılınç-Karzan, 2022 and H. Liu and Grigas (2021) further consider the risk bounds of different surrogate loss functions in the SPO setting. M. Liu et al., 2023 further considers minimizing the SPO loss in the setting of active learning. However, to the best of our knowledge, in the literature, there is no efficient algorithm for training a prediction model when the objective function in the downstream problem is nonlinear. We are the first to solve the predict-then-optimize problem for nonlinear objectives.

3.2 Value of One Data Point

In this section, we introduce the definition of the value of one data point in the general setting, which is shared by both product selection and assortment optimization problems. This concept measures the importance of acquiring one customer's preference regarding the expected revenue of downstream decision-making problems. It depends on the feature of each customer and dynamically changes during the data collection process. Intuitively, it measures the expected revenue increase of including one specific customer in the training set before acquiring her preference.

Suppose there are d products and m types of customers. The sets of product and customer types are denoted by $[d]$ and $[m]$, respectively, i.e., $[d] := \{1, ..., d\}$ and $[m] := \{1, ..., m\}$. At each iteration t, the preference of customer t for all products is represented by a vector $\mathbf{y}_t \in \mathcal{Y} \subseteq \mathbb{R}^d$, which we call the label of customer t. In particular, the j^{th} entry of $\mathbf{y}_t$, denoted by $\mathbf{y}_t^j$, represents the utility of product j for customer t. The type of customer t is represented by $\xi_t \in [m]$. The categorization of these types can be determined based on various customer information, such as their membership status, whether they were referred by a friend, or their geographic location. In this section, we assume the utility of products for customers from different types is independent. A general case will be discussed later in Section 3.7.

Since our goal is to build a prediction model that outputs the preference vectors of customers, the preference vectors are also referred to as the labels. At the beginning of the data collection process, we assume the labels of customers are unknown. During the data collection process, we decide whether to acquire the label of each customer sequentially. We denote the binary space $\{0, 1\}$ by $\mathbb{B}$. The binary decision vector $w \in \mathbb{B}^d$ indicates which products are selected in the offered set. That is, if product j is selected, then the j^{th} entry of w is one, i.e., $w_j = 1$; otherwise, $w_j = 0$. At iteration t, we assume ξ_t, the type of customer t, is generated from a known discrete probability distribution $\mu(\xi_t)$. If there is any type $\xi \in [m]$ that has zero probability to occur, then it will not impact the risk of the predictor or the cumulative label cost, and we can ignore this type ξ. We use $\underline{\mu} > 0$ to denote the minimum positive probability of any type of customer's arrival; that is, $\underline{\mu} := \min_{\xi \in [m]}\{\mu(\xi)|\mu(\xi) > 0\}$. We denote the fixed but unknown joint distribution of $(\xi_t, \mathbf{y}_t)$ by $\mathcal{D}$. The unknown distribution of $\mathbf{y}$ conditional on the value of ξ is denoted by $\mathcal{D}_{\mathbf{y}|\xi}$.

Given the utility vector $\mathbf{y}$ and the decision vector w, the revenue function for the retailer is denoted by $G(w, \mathbf{y})$. The retailer's goal is to maximize the expected revenue $\mathbb{E}_{y \sim \mathcal{D}_{y|\xi}}[G(w, \mathbf{y})|\xi]$ given type ξ. Our study focuses on two specific operations management problems: personalized product selection and assortment optimization problems with the MNL choice model. For these two problems, we are able to find another function $g(w, \mathbf{y})$ that satisfies $\mathbb{E}[G(w, y)|\xi] = g(w, \mathbb{E}[y|\xi])$ for any $\xi \in [m]$ and any $w \in \mathbb{B}^d$, as shown in Sections 3.3 and 3.4. This transformation from G to g indicates that it suffices to predict $\mathbb{E}[y|\xi]$ to estimate the expected cost given our decisions for this assortment optimization.

Given type ξ, we use $w^*(\mathbb{E}[\mathbf{y}|\xi])$ to denote the optimal decision given the conditional distribution of the preference vector $\mathcal{D}_{\mathbf{y}|\xi}$, i.e.,

$$w^*(\mathbb{E}[\mathbf{y}|\xi]) = \arg\max_w g(w, \mathbb{E}[\mathbf{y}|\xi]) = \arg\max_w \mathbb{E}[G(w, \mathbf{y})|\xi].$$

When multiple optimal decisions exist, we can use any rules to break the tie. Therefore, to find the best decision for a customer with type ξ, it suffices to estimate the conditional expectation $\mathbb{E}[\mathbf{y}|\xi]$, which only requires a point-prediction model for $\mathbb{E}[\mathbf{y}|\xi]$. We use $h \in \mathcal{H} : [m] \to \mathcal{Y}$, to denote the predictor from type ξ to the preference vector $\mathbf{y}$, where $\mathcal{H}$ is the family of predictors. We assume the hypothesis class $\mathcal{H}$ is well-specified, i.e., there exists a true optimal predictor $h^* \in \mathcal{H}$, such that $h^*(\xi) = \mathbb{E}[y|\xi]$ for all $\xi \in [m]$.

Suppose the prediction of $\mathbb{E}[\mathbf{y}|\xi]$ is $\hat{\mathbf{y}}(\xi)$. To evaluate the prediction $\hat{y}$ by the cost of decision error, we introduce the regret of the prediction in Definition 3.2.1.

Definition 3.2.1 (Regret of prediction $\hat{\mathbf{y}}$.)**.** *The regret of prediction $\hat{y}$ for type ξ is defined as:*

$$\ell(\hat{\mathbf{y}}, \mathbb{E}[\mathbf{y}|\xi]) = g(w^*(\mathbb{E}[\mathbf{y}|\xi]), \mathbb{E}[\mathbf{y}|\xi]) - g(w^*(\hat{\mathbf{y}}), \mathbb{E}[\mathbf{y}|\xi]).$$

In Definition 3.2.1, the first term $g(w^*(\mathbb{E}[\mathbf{y}|\xi]), \mathbb{E}[\mathbf{y}|\xi])$ is the maximum expected revenue when $\mathbb{E}[\mathbf{y}|\xi]$ is known. The second term $g(w^*(\hat{\mathbf{y}}), \mathbb{E}[\mathbf{y}|\xi])$ is the expected revenue of the best decision $w^*(\hat{\mathbf{y}})$ based on the prediction $\hat{\mathbf{y}}$. Thus, the regret of prediction measures the excess expected revenue loss induced by the decision $w^*(\hat{\mathbf{y}})$. It is obvious that $\ell(\hat{\mathbf{y}}, \mathbb{E}[\mathbf{y}|\xi]) \geq 0$ and $\ell(\mathbb{E}[\mathbf{y}|\xi], \mathbb{E}[\mathbf{y}|\xi]) = 0$. Definition 3.2.1 generalizes the SPO loss defined in Elmachtoub and Grigas, 2022 to the nonlinear case. When g is a linear function, Definition 3.2.1 reduces to the SPO loss. By the property of SPO loss, in general, $\ell(\hat{\mathbf{y}}, \mathbb{E}[\mathbf{y}|\xi])$ is discontinuous and nonconvex with respect to $\hat{\mathbf{y}}$.

Note that $\ell(h(\xi), \mathbb{E}[\mathbf{y}|\xi])$ measures the regret of predictor h on a single type ξ. The expected regret over the entire distribution is

$$\mathsf{Regret}(h) := \mathbb{E}_\xi[\ell(h(\xi), \mathbb{E}[\mathbf{y}|\xi])],$$

where $\mathsf{Regret}(h) \geq 0$ and $\mathsf{Regret}(h^*) = 0$. Since $\mathsf{Regret}(h)$ considers the expected regret for a single purchase, we normalize it by the market size $\beta > 0$ and thereby $\beta \cdot \mathsf{Regret}(h)$ measures the total expected regret of the entire market, which is referred to as the risk of h.

During the label acquisition process, the customer with type ξ arrives with probability $\mu(\xi)$. Upon observing the type of a customer ξ_t, the algorithm determines the appropriate incentive to provide to customer t for participating in the survey.

To determine the proper incentives for each data point, we introduce the notion of value of one data point in Definition 3.2.2. Define $\mathcal{S}_{t-1}$ as the training set before offering incentives to customer t, i.e., $\mathcal{S}_{t-1}$ is the set of $(\xi_i, \mathbf{y}_i)$ for all customer i who took the survey before customer t. We also use $\mathcal{S}_{t-1}(\xi)$ to denote the subset of $\mathcal{S}_{t-1}$ on type ξ. The oracle from the training set $\mathcal{S}$ to the predictor h is denoted by f; that is, $h_t = f(\mathcal{S}_t)$. For example, h_t can be the empirical minimizer of the squared error loss in data set $\mathcal{S}_t$.

Definition 3.2.2 (Value of one data point)**.** *The value of one data point $V(\xi_t; \mathcal{S}_{t-1})$ for customer t with type ξ_t is defined as*

$$V(\xi_t; \mathcal{S}_{t-1}) := \beta \cdot \mathit{Regret}(h_{t-1}) - \beta \cdot \mathbb{E}_{y_t}\left[\mathit{Regret}\left(f\left(\mathcal{S}_{t-1} \bigcup \{(\xi_t, \mathbf{y}_t)\}\right)\right) \middle| \xi_t\right].$$

The second term in the definition of $V(\xi_t; \mathcal{S}_{t-1})$ is the expected risk that h_t can achieve when y_t follows the conditional distribution $\mathcal{D}_{y|\xi_t}$. The value of one data point measures the expected contribution to the risk reduction from customer t, given the current training set $\mathcal{S}_{t-1}$ and the customer type ξ_t.

In practice, estimating the value of one data point exactly is nontrivial because of the following reasons. First, the joint distribution of (ξ, y) is unknown so the prediction error for $h(\xi)$ is unknown. In general, the prediction error of $h(\xi)$ relies on the complexity of the model class. Secondly, a smaller prediction error does not imply a smaller regret. Thus,

even if we obtain the exact prediction error for $h(\xi)$, converting this prediction error to the regret on a single type is nontrivial. Thirdly, since we offer personalized incentives, different types of customers have different probabilities to take surveys, and consequently, the data in training set $\mathcal{S}_t$ is not i.i.d. Lastly, the predictor $h_t = f\left(\mathcal{S}_{t-1} \bigcup \{(\xi_t, \mathbf{y}_t)\}\right)$ in the second term is a random function, which involves the distribution of $\mathcal{D}_{y_t|\xi_t}$. It is difficult to estimate $\mathsf{Regret}(h_t)$ for all possible outcomes of h_t. Thus, estimating the value of one data point is challenging. Considering the aforementioned challenges in calculating the value of one data point, we attempt to find an upper bound for the personalized product selection and assortment optimization problems.

3.3 Value of One Data Point in Personalized Product Selection

In personalized product selection, customers have different evaluations on products. We aim to recommend z products to each customer based on the types of customers. The sum of the utilities of these recommended products is referred to as the satisfaction level. Our goal is to select the best recommendations to maximize the satisfaction level. Since we do not have knowledge of the utility for each product given the type of a customer, we build a model that predicts the utilities of customers and use this prediction model to recommend personalized products for each customer.

Personalized product selection is a widespread practice in the real world, exemplified by platforms like CookUnity[3]. CookUnity operates as an online meal delivery platform, tasked with matching customer orders to the local chefs in the neighborhood community. The objective is to deliver cooked meals that align with customer preferences. Given the inherent variability in customer tastes and the distinct cooking styles of different chefs, CookUnity needs to learn the taste and diet requirements of each customer to successfully pair them with the best chef. To achieve this, CookUnity can employ a strategy of surveying customers and collecting their feedback. By leveraging this feedback, CookUnity can infer ratings for each chef and build a prediction model that estimates the preferences of chefs for individual customers. In what follows, we formulate the product selection problem and provide an exact expression for $U(\xi_t; \mathcal{S}_{t-1})$.

3.3.1 Formulations for Product Selection Problem

Let us now formally introduce the personalized product selection problem. Recall that the decision vector $w \in \{0,1\}^d$ indicates which products are selected for the customer and random vector $\mathbf{y}_t$ denotes the true utilities for whole products for customer t. We assume that the utility for each product is finite and we denote its upper bound by $\eta_{\mathcal{Y}}$, i.e., $y_t^i \leq \eta_{\mathcal{Y}}$, $\forall i \in [d]$.

[3]https://www.cookunity.com/

Given the utility vector $\mathbf{y}_t$, the deterministic product selection problem can be written as:

$$\max_{w \in \mathbb{B}^d} \quad \mathbf{y}_t^T w \tag{P1}$$
$$\text{s.t. } w^T \mathbf{1} = z. \tag{3.1}$$

Since (P1) can be viewed as a knapsack problem with unit weights, the optimal solution to (P1) is the set of z products with top utilities. We define the optimization oracle that solves (P1) given the utility vector $\mathbf{y}_t$ by $w^*(\mathbf{y}) : \mathbb{R}^d \to \mathbb{B}^d$, where $w^*(\mathbf{y}) := \arg\max_{w \in \mathbb{B}^d} \mathbf{y}^T w$, subject to constraint (3.1).

In the personalized product selection, when observing customer type ξ_t, we select z products by maximizing her expected satisfaction:

$$\max_{w \in \mathbb{B}^d} \quad \mathbb{E}[\mathbf{y}_t^T w | \xi_t] \tag{P2}$$
$$\text{s.t. } w^T \mathbf{1} = z.$$

By the linearity of (P2), the objective function can be rewritten as $\mathbb{E}[\mathbf{y}_t^T|\xi_t]w$. This means that the revenue function g can be written as $g(w, \mathbf{y}) = \mathbf{y}^T w$ and the optimal solution to Problem (P2) is $w^*(\mathbb{E}[\mathbf{y}_t^T|\xi_t])$. Therefore, in the personalized product selection, the regret of the predicted vector $\hat{\mathbf{y}}$ for customer type ξ (defined in Definition 3.2.1) can be written as

$$\ell(\hat{\mathbf{y}}, \mathbb{E}[\mathbf{y}|\xi]) = \mathbb{E}[\mathbf{y}|\xi]^T w^*(\mathbb{E}[\mathbf{y}|\xi]) - \mathbb{E}[\mathbf{y}|\xi]^T w^*(\hat{\mathbf{y}}). \tag{3.2}$$

In (3.2), the first term $\mathbb{E}[\mathbf{y}|\xi]^T w^*(\mathbb{E}[\mathbf{y}|\xi])$ is the maximal satisfaction of the customer, assuming that we know the true utility vector $\mathbb{E}[\mathbf{y}|\xi]$ in hindsight. The second term is the actual satisfaction of that customer, supposing that we select products $w^*(\hat{\mathbf{y}})$ based on our estimation $\hat{\mathbf{y}}$. The regret in (3.2) is equivalent to the expected SPO loss defined in (Elmachtoub and Grigas, 2022).

Since the predictions of different customer types ξ are independent, the predictor $h(\xi_t)$ is simply the average of historical observations, i.e., $h_{t-1}(\xi) = \frac{1}{n_{t-1}(\xi)} \sum_{(\cdot,\mathbf{y}) \in \mathcal{S}_{t-1}(\xi)} \mathbf{y}$, where $n_{t-1}(\xi)$ is the cardinality of $\mathcal{S}_{t-1}(\xi)$.

3.3.2 Upper Bounds for Value of One Data Point

Due to the challenges listed in Section 3.2, estimating the value of one data point accurately is nontrivial. Thus, in this section, by utilizing the structure of the product selection problem, we provide an upper bound of value of one data point $V(\xi_t; \mathcal{S}_{t-1})$. Our proposed upper bound for the value of one data point satisfies the following two properties: (1) Provide useful insights on which sample is more important to acquire its label. (2) Be close to the true value under some natural conditions.

When deriving the upper bound $U(\xi_t; \mathcal{S}_{t-1})$, we consider the following three factors. The first is the prediction error for $h(\xi_t)$, which is independent across different types. The second factor is $\mu(\xi_t)$. When $\mu(\xi_t)$ is smaller, the impact of this type ξ_t will become smaller, and thus

the value of one data point becomes smaller. The third factor is the location of $h(\xi_t)$, which is characterized by the distance to the degeneracy introduced in Elmachtoub and Grigas, 2022. The set of degenerate cost vector predictions is represented by $\mathcal{Y}^o := \{\mathbf{y} \in \mathbb{R}^d : \min_w : g(w, \mathbf{y})$ has multiple optimal solutions$\}$.

Definition 3.3.1 (Distance to degeneracy in general cases)**.** *The distance to degeneracy of the prediction $\hat{\mathbf{y}}$ is $\nu_S(\hat{\mathbf{y}}) := \inf_{\mathbf{y}\in\mathcal{Y}^o}\{\|\mathbf{y} - \hat{\mathbf{y}}\|\}$ where $\|\cdot\|$ is the ℓ_2 norm.*

This definition extends Definition 2.3.1 to the general nonlinear cases. The distance to degeneracy $\nu_S(h(\xi))$ captures the proximity of $h(\xi)$ to the closest utility vector that can result in multiple optimal decisions. When the distance to degeneracy is large, it is easier to distinguish the optimal decision from sub-optimal decisions. Consequently, the value of one data point is expected to be small. In the product selection and assortment optimization problems, we can use Equation (2.4) to calculate the distance to degeneracy. Suppose the set $\mathcal{S}^o := \{w_j^o : j = 1, ..., K\}$ is the set of all the possible combinations of z products. According to Equation (2.4), the distance to degeneracy of any vector $\mathbf{y} \in \mathbb{R}^d$ satisfies the following equation:

$$\nu_S(\mathbf{y}) = \min_{w\in\mathcal{S}^o : w\neq w^*(\mathbf{y})} \left\{ \frac{\mathbf{y}^T(w - w^*(\mathbf{y}))}{\|w - w^*(\mathbf{y})\|} \right\}. \tag{3.3}$$

Based on the notion of distance to degeneracy, Theorem 3.3.1 provides a tighter upper bound for the value of one data point.

Theorem 3.3.1 (Upper bound for the value of one data point)**.** *Given customer type $\xi \in [m]$, suppose $\|h_t(\xi) - \mathbb{E}[\mathbf{y}|\xi]\| \leq \rho_t(\xi)$ for some $\rho_t(\xi) > 0$. Then, the value of one data point for type ξ is no larger than $U_M(\xi, h_{t-1}(\xi), \rho_{t-1}(\xi))$, i.e.,*

$$V(\xi; \mathcal{S}_{t-1}) \leq U_M(\xi, h_{t-1}(\xi), \rho_{t-1}(\xi)),$$

where $U_M(\xi, \mathbf{y}, \rho) := \sqrt{2\min\{z, d - z\}}\beta\mu(\xi)\rho \cdot \mathbb{I}\{\nu_S(\mathbf{y}) \leq \rho\}$.

Although the upper bound in Theorem 3.3.1 is derived in theory, this upper bound provides useful practical insights on the importance of each data point. These insights can be summarized into the following three points:

1. **Insights from $\mathbb{I}\{\nu_S(h(\xi_t) \leq \rho_t(\xi_t)\}$.** When $\nu_S(h(\xi_t)) > \rho_t(\xi_t)$, the distance to degeneracy of $h(\xi_t)$ is larger than the prediction error. As demonstrated in the proof, in this case, the prediction model has enough confidence in identifying the optimal decision based on the prediction $h(\xi_t)$, resulting in zero regret. Consequently, acquiring the label of ξ_t does not make any improvement at ξ_t. Thus, the value of one data point should be zero.

2. **Insights from $\rho_t(\xi_t)$.** The term $\rho_t(\xi_t)$ indicates that the value of one data point is proportional to the prediction error. Intuitively, when the training set contains a lot of observations of type ξ_t, then the prediction error for this type of customer $\rho_t(\xi_t)$ gets smaller. It implies the importance of this type gets smaller, and we should collect more observations of other types.

3. **Insights from $\mu(\xi_t)$.** The term $\mu(\xi_t)$ indicates that the value of one data point is proportional to the test set density. Intuitively, when type ξ_t appears more in the future, this type becomes more important, and we should acquire more information about this type of customer.

During the data collection process, as the prediction error converges to zero, $\sqrt{2\min\{z, d-z\}}$ $\mu(\xi_t)\rho_t(\xi_t)$ also approaches zero. Thus, in Theorem 3.3.1, the upper bound converges to zero due to two factors: $\rho_t(\xi_t)$ can be smaller than $\nu_S(h(\xi_t)$, and the prediction error $\rho_t(\xi_t)$ converges to zero.

Next, to demonstrate that our upper bounds for the value of one data point are close to the true values, Theorem 3.3.2 provides its corresponding lower bound, which has the same order as our upper bounds.

Theorem 3.3.2 (Matching lower bounds for the value of one data point)**.** *For personalized product selection problem, there exists a distribution $\mathcal{D}$ and constant $K > 0$, such that for any t,*

$$V(\xi_t; \mathcal{S}_{t-1}) \geq K\beta \cdot \mu(\xi_t)\rho_t(\xi_t)\mathbb{I}\{\nu_S(h_{t-1}(\xi_t)) \leq \rho_{t-1}(\xi_t)\}\,,$$

where $\rho_t(\xi) := \|h_t(\xi) - \mathbb{E}[\mathbf{y}|\xi]\|$.

Theorem 3.3.2 demonstrates that under some noise distributions, our estimated upper bounds for the value of one data point have the same order of t as the true value of one data point. In the next section, we utilize this upper bound to derive an upper bound for the value of one data point in the personalized assortment optimization problem.

3.4 Value of One Data Point in Assortment Optimization

In this section, we consider the upper bound for the value of one data point in the context of assortment optimization. In assortment optimization, the objective is to select at most z products to display, while also accounting for the possibility of customers choosing not to make a purchase. The ultimate goal is to maximize the expected revenue, which is impacted by the individual prices of each product.

3.4.1 Formulations for Assortment Optimization

In the assortment optimization, $\mathbf{y}_t$ represents the utility vector of customer t, and each entry, y_t^i, represents the utility of product i for customer t. The utility for the no-purchase option is denoted by y_t^0, which is assumed to be fixed for all customers. Customers will purchase the product (or no-purchase option) with the largest utility in the assortment. Suppose the results of each survey reveal the true utility vector y_t for all products. We denote the set of recommended products (i.e., assortment) by Z, and define set $[\bar{d}] := \{0\} \cup [d]$. Since the utility vector is random, the purchase probability of product i given customer type ξ is

$$\mathbb{P}(y_t^i \geq y_t^j, \forall j \in [\bar{d}], j \neq i|\xi). \tag{3.4}$$

In this section, we assume the customer follows the MNL choice model (Ben-Akiva and Lerman, 1985), so the utility for product i can be written as $y_t^i = \bar{y}_t^i + \epsilon_i$, where $\bar{y}_t^i$ represents the expected utility, i.e., $\bar{y}_t^i = \mathbb{E}[y_t^i]$, and ϵ_i is the noise. The noise ϵ_i is assumed to follow a Gumbel distribution with variance $\frac{\pi^2}{6}\sigma^2$ and mean 0, where the parameter σ controls the noise level. Recall that w denotes the decision vector. By the property of the Gumbel distribution, given $\mathbf{y}$, the purchase probability (3.4) can be written as

$$\mathbb{P}(y_t^i > y_t^j, \forall j \in [\bar{d}], j \neq i|\xi) = \frac{e^{\bar{y}_t^i/\sigma}}{\sum_{j=0}^{d} e^{\bar{y}_t^j/\sigma}}.$$

Without the loss of generality, we assume the mean of the no-purchase option is 0, i.e., $\bar{y}_t^0 = 0$ for all t. Then, the cardinality-constrained assortment optimization problem under the MNL model can be written as:

$$\begin{aligned} \max_{w\in\mathbb{B}^d, u\in\mathbb{R}^d} \quad & \frac{\sum_{i\in[d]} u_i p_i w_i}{1 + u^T w} \\ s.t. \quad & w^T \mathbf{1} \leq z, \\ & u_i = e^{\bar{y}_t^i/\sigma}, \quad \forall i \in [d], \end{aligned} \tag{P3}$$

where $u_i = e^{\bar{y}_t^i/\sigma}$ is an intermediate variable for convenience.

We denote the objective function of (P3) by $g(w, \bar{\mathbf{y}})$. Following the definition of regret in Definition 3.2.1, the regret for a prediction $\hat{\mathbf{y}}$ in the assortment optimization problem can be written as:

$$\ell_p(\hat{\mathbf{y}}, \bar{\mathbf{y}}) = g(w^*(\bar{\mathbf{y}}), \bar{\mathbf{y}}) - g(w^*(\hat{\mathbf{y}}), \bar{\mathbf{y}}). \tag{3.5}$$

The regret of the prediction in (3.5) measures the revenue loss based on the current prediction of the utility vector. Following (3.5), we can obtain the expression of the expected regret $\mathsf{Regret}(h)$ for a predictor h and the value of one data point in the setting of assortment optimization. Next, we consider how to estimate an upper bound for the value of one data point.

3.4.2 Upper bound for Value of One Data Point in Assortment Optimization

In contrast to (P1) whose objective function is linear, the objective function in (P3) is nonlinear. This introduces additional challenges when estimating the upper bound of the value of one data point $U(\xi_t; \mathcal{S}_{t-1})$. To address the challenge from the non-linearity of (P3), we first convert the assortment optimization problem into a similar form of the product selection problem. For convenience, we denote $g(\mathbf{y}, w)$ by $g_a(u, w)$ in the assortment optimization problem, where $u \in \mathbb{R}^d$ is a vector composed of u_i, $i \in [\bar{d}]$.

Suppose $g_a^*(u)$ is the maximum revenue of (P3). Indicated by Rusmevichientong, Shen, and Shmoys, 2010, the optimal solution to the capacitated assortment optimization problem is the set of the products with top z highest positive value of $u_i(p_i - g_a^*(u))$. In addition, $g_a^*(u)$ satisfies that:

$$g_a^*(u) = \max_w : w_i u_i (p_i - g_a^*(u)).$$

Thus, given the optimal revenue $g_a^*(u)$ and u, the best assortment in (P3) can be obtained by solving the following product selection problem:

$$\begin{aligned} \max_w : \quad & w_i u_i (p_i - g_a^*(u)) \\ s.t. \quad & w^T \mathbf{1} \leq z \,. \end{aligned} \tag{3.6}$$

In (3.6), the coefficient $u_i(p_i - g_a(u))$ can be viewed as the utility in the product selection problem. Thus, according to the upper bound of the value of one data point in Theorem 3.3.1, to estimate the revenue loss in the assortment, it suffices to control the estimation error of the coefficient $u_i(p_i - g_a(u))$ in (3.6), which is shown in Lemma 3.4.1 below.

Lemma 3.4.1. *If $|u_i - u_i'| \leq \varepsilon$ for any $i \in [d]$, then the estimation error of the coefficient in* (3.6) *satisfies*

$$|u_i(p_i - g_a^*(u)) - u_i'(p_i - g_a^*(u'))| \leq (z\eta_p e^{\eta_{\mathcal{Y}}/\sigma} + \eta_p)\varepsilon.$$

Lemma 3.4.1 implies that the estimation error of the coefficients in (3.6) can be bounded by the estimation error of u. Recall that $y_t^i \leq \eta_{\mathcal{Y}}$, $\forall i \in [d]$. Then, we have that $|u_i - u_i'| \leq e^{\eta_{\mathcal{Y}}/\sigma}|y_t^i - y_t^{i'}|$, $\forall i \in [d]$, which means the estimation error of u can be further bounded by the estimation error of expected utility vector $\bar{\mathbf{y}}$. Suppose that $\rho(\xi)$ is the confidence interval for the estimation of y_t^i given type ξ, $\forall i \in [d]$. Then we can conclude that

$$|u_i(p_i - g_a^*(u)) - u_i'(p_i - g_a^*(u'))| \leq (z\eta_p e^{\eta_{\mathcal{Y}}/\sigma} + \eta_p)e^{\eta_{\mathcal{Y}}/\sigma}|y_t^i - y_t^{i'}| = (z\eta_p e^{\eta_{\mathcal{Y}}/\sigma} + \eta_p)e^{\eta_{\mathcal{Y}}/\sigma}\rho(\xi).$$

Define a constant $\varkappa := (z\eta_p e^{\eta_{\mathcal{Y}}/\sigma} + \eta_p)e^{\eta_{\mathcal{Y}}/\sigma}$. We can derive an upper bound for the value of one data point for the assortment optimization in Theorem 3.4.1.

Theorem 3.4.1 (Upper bound for the value of one data point in the assortment optimization)**.** *Suppose Assumption 3.5.1 holds. Assume the current predictor for ξ_t is $\hat{\mathbf{y}}$, and the prediction error bound for $h(\xi_t)$ is $\rho_t(\xi_t)$. The value of one data point can be upper bounded by*

$$V(\xi_t; \mathcal{S}_{t-1}) \leq U_M^A(\xi_t, \hat{\mathbf{y}}, \rho_t(\xi_t))\,,$$

where $U_M^A(\xi, \mathbf{y}, \rho) := \min\left\{\varkappa\sqrt{2\min\{z, d-z\}}\mu(\xi)\beta\rho\mathbb{I}\{\nu_S(\mathbf{y}) \leq \rho\}, \eta_p\right\}$.

In Theorem 3.4.1, the constant $\varkappa$ converts the prediction error of the y^i to the estimation error of the coefficient $u_i(p_i - g_a(u))$. The second upper bound η_p in Theorem 3.4.1 is because the maximum revenue loss from one customer is at most η_p. Theorem 3.4.1 demonstrates that when the prediction error for the expected utility $\rho_t(\xi_t)$ is smaller than the distance to degeneracy of $\hat{y}$, then the optimal assortment is identified with high probability and the value of one data point will drop to zero.

3.5 Active Label Acquisition Algorithms

In this section, we illustrate how to use the upper bound of the value of one data point to determine personalized incentives during the customer survey process. Intuitively, we offer more incentives to the customer with higher value of one data point. This procedure of determining incentives by the value of one data point is illustrated in Figure 3.2.

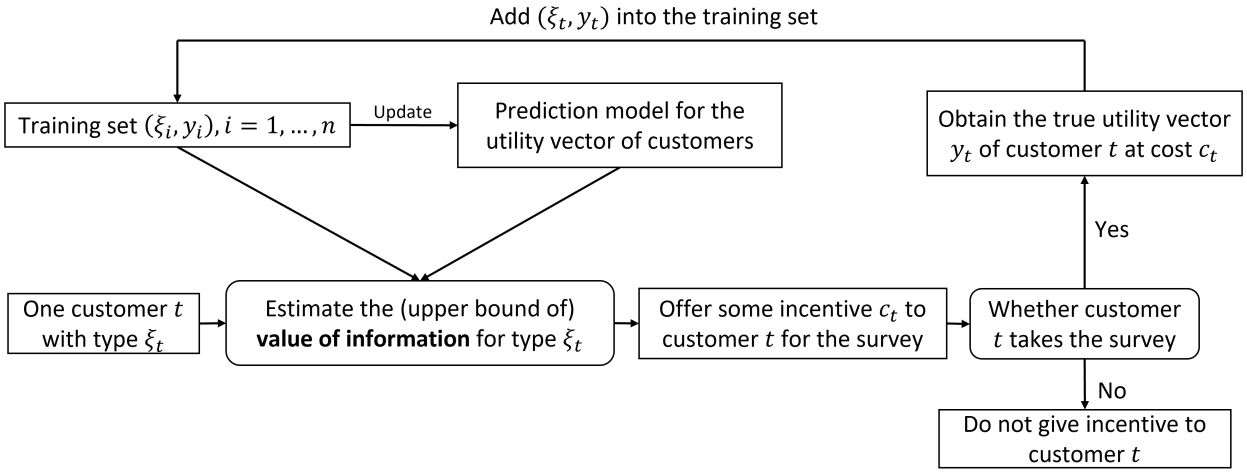


Figure 3.2: Value of one data point in active label acquisition

In Figure 3.2, at time t, when a new customer with type ξ_t arrives, the retailer estimates the value of one data point of this customer. This estimation is performed using the current prediction model and training set. Subsequently, based on the estimated value of one data

point, the retailer offers a personalized incentive to this customer. The customer then has a certain probability of taking the survey, which only depends on the offered incentive. If the customer takes the survey, the retailer can acquire the true preference $\mathbf{y}_t$ and add a new data point $(\xi_t, \mathbf{y}_t)$ into the training set $\mathcal{S}_t$. We propose a general active label acquisition algorithm that determines personalized incentives sequentially in Algorithm 2.

Recall that there exists a tradeoff between the label cost and the risk associated with the prediction model when determining the incentives. To capture this tradeoff, we introduce the concept of the *comprehensive cost* for the retailer as the weighted sum of the incentives (i.e., the cumulative label cost) and the expected regret of the prediction model after the data acquisition period. We use c_t to denote the incentive for customer t and h_t to denote the prediction model used at time t.

Definition 3.5.1 (Comprehensive cost)**.** *The comprehensive cost at time T is defined as*

$$\mathscr{C}(\mathbf{c}_T, h_T) := \sum_{t=1}^{T} c_t \cdot \mathbb{I}\{\textit{Customer } t \textit{ accepts the offer for survey}|c_t\} + \beta \cdot \mathsf{Regret}(h_T).$$

We assume c_t is within some range $\{0\} \cup [c_{\mathsf{min}}, c_{\mathsf{max}}]$. When $c_t = 0$, it means the retailer does not provide any incentive for this customer or gather feedback from the customer. c_{min} refers to the minimum incentive we need to spend on each surveyed customer. For example, it may include the labor cost for collecting and analyzing answers. c_{max} refers to the maximum incentive that can be allocated to one customer. The first term of the comprehensive cost is the cumulative label cost, which is a non-decreasing term. The second term, $\beta \cdot \mathsf{Regret}(h_T)$ is the risk of the current predictor, which is supposed to be decreasing as more customers are surveyed. The market size β can be viewed as a hypoparameter that controls the tradeoff between the cumulative label cost and the regret of the final prediction model. In Section 3.5.1, we provide the connection between β and the Lagrange multiplier when ensuring the final regret level is smaller than some threshold.

When customer t does not accept the offer, the training set S_{t-1} remains the same and we have that $h_t = h_{t-1}$. Our objective is to minimize the expectation of the comprehensive cost $\mathscr{C}(\mathbf{c}_T, h_T)$ at time T by judiciously offering personalized incentives.

We provide an algorithm to determine the personalized incentive based on an upper bound of the value of one data point in Algorithm 2. It follows the idea in Figure 3.2 to determine the incentive according to the upper bound of the value of one data point. There are two scenarios. When this upper bound is no less than c_{min}, the incentive c_t is set as any value between $[c_{\mathsf{min}}, c_{\mathsf{max}}]$. Otherwise, when its upper bound is less than c_{min}, the value of one data point for this customer has been smaller than the minimum incentive c_{min}. It implies that offering incentives between $[c_{\mathsf{min}}, c_{\mathsf{max}}]$ to this customer will increase the comprehensive cost. In this case, we should not offer any incentives.

We use $p(c) \in [0, 1]$ to denote the probability of taking the survey given the incentive c. We make the following assumption on the lowest probability of taking the survey, without assuming any specific function form of $p(c)$. It is worth noting that the structure of $p(c)$ is not needed for our active label acquisition algorithm.

Algorithm 2 Active Label Acquisition in Product Selection and Assortment Optimization problems

Input: An oracle $U(\xi; \mathcal{S})$ that calculates the upper bound of the value of one data point, the maximum unit label cost c_{max} and minimum unit label cost c_{min}, the probability of taking the survey $p(c)$.
Initialization: Training set $\mathcal{S}_0 \leftarrow \emptyset$; Arbitrarily pick predictor $h_0 \in \mathcal{H}$.
for t from $1, 2, ..., T$ **do**
  Customer t arrives with type ξ_t, where ξ_t follows the probability density $\mu(\xi)$.
  Calculate the upper bound $U(\xi_t; \mathcal{S}_{t-1})$ of the value of one data point.
  if $U(\xi_t; \mathcal{S}_{t-1}) < c_{\mathsf{min}}$ **then**
    Do not survey customer t and offer 0 incentive.
    $\mathcal{S}_t \leftarrow \mathcal{S}_{t-1}$, $h_t \leftarrow h_{t-1}$.
  else
    Offer customer t with any incentive c_t within $[c_{\mathsf{min}}, c_{\mathsf{max}}]$ for the survey.
    if Customer t decides to take the survey **then**
      The retailer gets the true preference vector $\mathbf{y}_t$ of all products at cost c_t.
      Update $\mathcal{S}_t \leftarrow \mathcal{S}_{t-1} \cup \{(\xi_t, y_t)\}$.
      Update $h_t \leftarrow f(\mathcal{S}_t)$.
    else
      $\mathcal{S}_t \leftarrow \mathcal{S}_{t-1}$, $h_t \leftarrow h_{t-1}$.
    end if
  end if
end for

Assumption 3.5.1. *The probability for the customer to take the survey $p(c)$ is an increasing function of the offered incentive c. Furthermore, $p(c_{min}) \geq p_{min}$ for some $p_{min} > 0$.*

In Assumption 3.5.1, we assume that the probability of taking the survey is an increasing function of the incentive, without assuming other structural information such as concavity or continuity. The lower bound of $p(c_{\mathsf{min}})$ is a mild assumption that applies to most incentive models in practice.

Under Assumption 3.5.1, Theorem 3.5.1 provides a general theoretical upper bound for the comprehensive cost of Algorithm 2. We denote the risk bound of f at iteration T for the naive supervised learning by $R_{\mathsf{s}}(T)$. That is, if we passively acquire the labels before iteration T and $h_T = f(\{(\xi_t, y_t)\}_{t=1}^T)$, then $\beta\mathsf{Regret}(h_T)$ is bounded by $\beta R_{\mathsf{s}}(T)$. For example, when the revenue function g is linear, H. Liu and Grigas, 2021 shows that $R_{\mathsf{s}}(T) \leq \tilde{\mathcal{O}}(T^{-1/2})$ under some conditions, where $\tilde{\mathcal{O}}(\cdot)$ represents the asymptotic order that ignores the logarithm dependence.

Theorem 3.5.1 (Weak Guarantees for Algorithm 2)**.** *Suppose Assumption 3.5.1 holds and $U(\xi_t; \mathcal{S}_{t-1})$ is a universal upper bound for all $\xi_t \in [m]$, which satisfies that $U(\xi_t; \mathcal{S}_{t-1}) \geq$*

$\beta\mathsf{Regret}(h_{t-1})$ and $U(\xi_t;\mathcal{S}_{t-1}) \geq U(\xi_{t+1};\mathcal{S}_t)$ for all $t \geq 1$. In Algorithm 2, at iteration T, we have the following guarantees for the two terms in the comprehensive cost:

(1) With probability at least $1 - e^{-p_{min}T/8}$, the risk of h_T satisfies that

$$\beta\mathsf{Regret}(h_T) \leq \beta R_{\mathsf{s}}(p_{min}T/2) + c_{min}.$$

(2) The incentive $c_t = 0$ for all $t \geq \underline{t}$, where $\underline{t} := \inf\{t \geq 1 : U(\xi_t;\mathcal{S}_{t-1}) \leq c_{min}\}$. The cumulative label cost by time t is at most $c_{max}\min\{\underline{t}, T\}$.

Remark 3.5.1 (Value of $U(\xi_t;\mathcal{S}_{t-1})$)**.** *The conditions in Theorem 3.5.1 imply that $U(\xi_t;\mathcal{S}_{t-1})$ can be any non-increasing sequence that is larger than $\mathsf{Regret}(h_{t-1})$ and independent of ξ. However, when choosing the value of $U(\xi_t;\mathcal{S}_{t-1})$, we should expect that $U(\xi_t;\mathcal{S}_{t-1})$ converges to zero. Otherwise, if $U(\xi_t;\mathcal{S}_{t-1})$ is always larger than c_{min}, $\underline{t}$ in Theorem* 3.5.1.(2) *does not exist, which means the bound in Theorem* 3.5.1.(2) *becomes the naive $c_{max}T$.*

Theorem *3.5.1.(1)* demonstrates that the risk of h_T is at most $R_{\mathsf{s}}(p_{\mathsf{min}}T/2) + c_{\mathsf{min}}$ with high probability. The first term $R_{\mathsf{s}}(p_{\mathsf{min}}T/2)$ is in the same order as supervised learning while the second term c_{min} is an upper bound for the final risk of the predictor h_T when T tends to infinity. Note that the risk of h_T does not necessarily converge to zero due to the minimum label cost. Theorem *3.5.1.(2)* further implies that when $U(\xi_t;\mathcal{S}_{t-1})$ goes below c_{min}, we will stop surveying customers and offer zero incentives. Intuitively, considering the positive minimum incentive, we do not have to push the risk of the predictors to zero. When $t \geq \underline{t}$, it is not worth continuing labeling customers, because the benefit of risk reduction is less than the future label cost. If $c_{\mathsf{min}} = 0$, $\underline{t}$ does not exist and the upper bound for the cumulative label cost in Theorem *3.5.1.(2)* becomes the naive Tc_{max}.

The results of Theorem 3.5.1 do not depend on the structure of the assortment optimization or the product selection problem. The bounds in Theorem 3.5.1 can be improved by utilizing the upper bounds we derived for the value of one data point. By tailoring $U(\xi_t;\mathcal{S}_{t-1})$ for different contexts ξ_t, we further reduce the bound for the cumulative label cost. These results demonstrate that the active label acquisition algorithm can achieve a smaller comprehensive cost than simple supervised learning under some natural noise conditions.

Remark 3.5.2 (Tailoring incentives when $p(c)$ is known)**.** *In Algorithm 2, we assume $p(c)$ is unknown, and set the incentive as any value between $[c_{min}, c_{max}]$ when we decide to survey that customer. This allows our theoretical guarantees hold generally under mild assumptions of customer behaviors. When knowing the exact form of $p(c)$, we can set incentives by minimizing the comprehensive cost. However, knowing the exact form of $p(c)$ does not improve the order of the theoretical guarantees of our algorithm. This discussion is provided in detail in Section 3.5.2.*

3.5.1 Illustration for Comprehensive Cost during the Active Label Acquisition

In this section, we illustrate the comprehensive cost by transforming the active label acquisition algorithm into a constraint optimization problem. When considering a specific regret threshold $\overline{R} > 0$ for the predictor, the goal of the active label acquisition can be formulated as follows.

$$\begin{aligned} \min: \quad & \sum_{t=1}^{T} c_t \cdot \mathbb{I}\{\text{Customer } t \text{ accepts the survey offer}|c_t\} \\ s.t. \quad & \mathsf{Regret}(h_T) \leq \overline{R}. \end{aligned}$$

In this formulation, we want to minimize the total label cost up to time T, while ensuring that the expected regret $\mathsf{Regret}(h_T)$ remains below the specified threshold $\overline{R}$. In this constraint, $\mathsf{Regret}(h_T)$ is the risk of the current predictor, which is supposed to be decreasing as more customers are surveyed.

Remark 3.5.3 (Illustration of the comprehensive cost)**.** *We explain our objective from the other perspective. When considering a specific regret threshold $\overline{R} > 0$ for the predictor, the minimization of the objective function $\mathscr{C}(\mathbf{c}_T, h_T) - \beta\overline{R}$ can be regarded as a Lagrangian function in the following problem: minimize $\sum_{t=1}^{T} c_t \cdot \mathbb{I}\{$Customer t accepts the survey offer$|c_t\}$, subject to the constraint $\mathsf{Regret}(h_T) \leq \overline{R}$. In other words, the minimization of the comprehensive cost can also be interpreted as minimizing the total label cost up to time T, while ensuring that the expected regret $\mathsf{Regret}(h_T)$ remains below the specified threshold $\overline{R}$. In this context, the market size β can be seen as the Lagrange multiplier, representing the "shadow price" of the regret requirement.*

To decompose the variation of the comprehensive cost into each customer, let us consider the difference of the comprehensive cost between time step $t-1$ and t. The difference is

$$c_t \cdot \mathbb{I}\{\text{Customer } t \text{ accepts the offer for survey}|c_t\} + \beta \cdot [\mathsf{Regret}(h_t) - \mathsf{Regret}(h_{t-1})]. \quad (3.7)$$

We use $p(c) \in [0,1]$ to denote the probability of taking the survey given the incentive c. When customer t does not accept the offer, we have that $h_t = h_{t-1}$, and the second term in (3.7) becomes zero. Thus, the expectation of (3.7) can be written as $p(c_t)\,(c_t - V(\xi_t; \mathcal{S}_{t-1}))$. To minimize the expectation of the comprehensive cost at iteration t, we greedily minimize this expectation of difference. We can either offer zero incentive to ignore this customer, or offer some incentive between $[c_{\mathsf{min}}, c_{\mathsf{max}}]$ that minimizes the expectation of the difference. This minimizer is denoted by $c^*(V(\xi_t; \mathcal{S}_{t-1}), p)$, i.e., we have

$$c^*(V(\xi_t; \mathcal{S}_{t-1}), p) := \arg \min_{c \in [c_{\mathsf{min}}, c_{\mathsf{max}}]} \{p(c)\,(c - V(\xi_t; \mathcal{S}_{t-1}))\}\,.$$

Calculating $c^*(V(\xi_t; \mathcal{S}_{t-1}), p)$ requires the knowledge of $p(c)$ and the value of one data point $V(\xi_t; \mathcal{S}_{t-1})$. In Proposition 3.5.1, we demonstrate that if $p(c)$ is a concave function, then the optimal incentive based on the value of one data point is less than the optimal incentive based on the upper bound of the value of one data point.

Proposition 3.5.1 (Upper bound for the optimal incentive)**.** *Suppose $p(c)$ is a twice-differentiable increasing concave function, then for any upper bound $U(\xi_t; \mathcal{S}_{t-1})$ of $V(\xi_t; \mathcal{S}_{t-1})$, we have $c^*(V(\xi_t; \mathcal{S}_{t-1}), p) \leq c^*(U(\xi_t; \mathcal{S}_{t-1}), p)$.*

In Proposition 3.5.1, the conditions on $p(c)$ are satisfied by a large class of functions. For example, $p(c)$ can be a logarithmic or linear function. Proposition 3.5.1 implies that, we can use $c^*(U(\xi_t; \mathcal{S}_{t-1}), p)$ as an upper bound for $c^*(V(\xi_t; \mathcal{S}_{t-1}), p)$.

In the original algorithm, when we decide to survey a customer, we can offer any incentive between $[c_{\mathsf{min}}, c_{\mathsf{max}}]$. If the exact form of $p(c)$ is known, we can use $c^*(U(\xi_t; \mathcal{S}_{t-1}), p)$ as the optimal incentive. All the guarantees about Algorithm 2 throughout the chapter also apply to both cases. These guarantees do not rely on the specific structure of $p(c)$; they only necessitate its lower bound, as stipulated in Assumption 3.5.1. In Section 3.5.2, we provide some examples of $p(c)$ and discuss the solutions of $c^*(V(\xi_t; \mathcal{S}_{t-1}), p)$ when $V(\xi_t; \mathcal{S}_{t-1})$ and function $p(c)$ are known.

3.5.2 Impact of $p(c)$ on Incentives

In this section, we present some examples of the function $p(c)$, and consider setting the incentive as $c^*(U(\xi_T; \mathcal{S}_{T-1}), p)$. Note that since $c^*(U(\xi_T; \mathcal{S}_{T-1}), p)$ is a smaller and more careful incentive than $U(\xi_T; \mathcal{S}_{T-1})$, the theoretical guarantees in the main body for $U(\xi_T; \mathcal{S}_{T-1})$ still apply to $c^*(U(\xi_T; \mathcal{S}_{T-1}), p)$. In this section, we provide some insights on the impact of $p(c)$ on $c^*(U(\xi_T; \mathcal{S}_{T-1}), p)$. Recall that

$$c^*(U(\xi_T; \mathcal{S}_{T-1}), p) = \arg \min_{c \in \{0\} \cup [c_{\mathsf{min}}, c_{\mathsf{max}}]} : \{p(c)\,(c - U(\xi_T; \mathcal{S}_{T-1}))\}\,.$$

For simplicity of the expression, we use $c^*(U, p)$ to denote $c^*(U(\xi_T; \mathcal{S}_{T-1}), p)$ in this section. Recall that Proposition 3.5.1 indicates that $c^*(U, p)$ is an upper bound for $c^*(V, p)$, where U is the upper bound of V. To obtain insights, for simplicity, we ignore the bounded constraint $c \in \{0\} \cup [c_{\mathsf{min}}, c_{\mathsf{max}}]$. Then, we have the following insights in Proposition 3.5.2.

Proposition 3.5.2. *Suppose $p_0(c)$ is a twice-differentiable increasing concave function. Then, we have the following insights on the relations between $c^*(U, p)$ and $p(c)$.*

1. *$c^*(U, p)$ will be smaller if $p(c)$ is shifted higher, i.e., suppose that $p(c) = p_0(c) + k_1$ for some $k_1 > 0$. Then, $c^*(U, p)$ is a decreasing function of k_1.*

2. *$c^*(U, p)$ is independent of the scale of $p(c)$, i.e., suppose that $p(c) = p_0(c) \cdot k_1$ for some $k_1 > 0$. Then, $c^*(U, p)$ is the same for all $k_1 > 0$.*

Proposition 3.5.2 indicates if the location of $p(c)$ is higher, the incentive should be smaller. On the other hand, the incentive is independent of the scale of $p(c)$. Next, we consider one specific example of $p(c)$, where $p(c) = k_1 + k_2 c$ for some $k_1, k_2 > 0$. In other words, $p(c)$ is a linear function that satisfies Proposition 3.5.1 and Assumption 3.5.1. Then, we have the following properties.

Proposition 3.5.3. *Suppose $p(c) = k_1 + k_2 c$ for some $k_1, k_2 > 0$. Then, we have $c^*(U, p) = \frac{U}{2} - \frac{k_1}{2k_2}$.*

Proposition 3.5.3 indicates that for the fixed interception k_1, when the slope is larger, the incentive should be larger. In other words, if customers are more sensitive to the incentives, then we tend to offer higher incentives. Proposition 3.5.3 further reveals that $c^*(U, p)$ is like a calibration of U according to the structure of $p(c)$. This calibration does not change the order of $c^*(U, p)$ from U. Thus, considering the structure of $p(c)$ has little impact on the order of the incentives, which means that the performance of the active label acquisition algorithm mostly depends on the U rather than the structure of $p(c)$.

3.6 Theoretical Guarantees for Active Label Acquisition

In this section, we provide the theoretical performance guarantees for our active label acquisition algorithm in the contexts of personalized product design and assortment optimization problem. By analyzing its performance guarantees, we demonstrate that our algorithm can achieve a much smaller comprehensive cost than simple supervised learning under some low-noise conditions.

3.6.1 Personalized Product Selection

In this section, we consider the active label acquisition in the context of the personalized product selection.

In order to obtain a more precise bound for the comprehensive cost compared to Theorem 3.5.1, we delve deeper into the density of the distribution of $\mathbb{E}[\mathbf{y}|\xi]$ near the degeneracy. For this purpose, we adopt the definition of the near-degeneracy function Ψ in Definition 2.3.2:

$$\Psi(\rho) := \mathbb{P}\big(\nu_S(\mathbb{E}[\mathbf{y}|\xi]) \leq \rho\big)\,.$$

The near-degeneracy function $\Psi(\cdot)$ quantifies the probability that the distance to degeneracy of $\mathbb{E}[\mathbf{y}|\xi]$ is less than ρ, given that ξ follows the distribution $\mu(\xi)$. Intuitively, when Ψ is smaller, the density allocated near the degeneracy becomes smaller. This suggests that identifying the optimal decisions becomes easier, resulting in a smaller cumulative label cost.

Theorem 3.6.1 provides an upper bound for the first part of the comprehensive cost.

Theorem 3.6.1 (Bound for the risk)**.** *Suppose Assumption 3.5.1 holds. In Algorithm 2, at each iteration, given $\delta \in (0,1)$, set $U(\xi_t; \mathcal{S}_{t-1})$ as $U_M(\xi_t, h(\xi_t), \rho_t(\xi_t))$, where $\rho_t(\xi_t) = 2\eta_{\mathcal{Y}}\sqrt{\frac{d\ln(t/\delta)}{n(\xi_t)}}$. For any $\delta \in (0,1)$, with probability at least $1 - mTe^{-\frac{\underline{\mu} p_{min} T}{8}} - mTe^{-T\ln(\frac{1}{1-p_{min}\underline{\mu}})} - \frac{\delta}{T^2}$, we have the risk of h_T is at most $\beta\mathcal{R}(T, c_{min})$, where*

$$\mathcal{R}(T, c_{min}) := \varphi(T) + \frac{c_{min}}{\beta\underline{\mu}}\Psi\Big(\frac{\sqrt{2}c_{min}}{\beta\underline{\mu}\sqrt{\min\{z, d-z\}}}\Big),$$

and $\varphi(T)$ is defined as

$$\varphi(T) := \eta_{\mathcal{Y}} \min\left\{ \sum_{\xi\in[m]} 4\mu(\xi)\sqrt{\min\{z, d-z\}}\sqrt{\frac{d\ln(T/\delta)}{p_{min}\mu(\xi)T}}, \sqrt{d}\sum_{\xi\in[m]} \Psi\Big(4\eta_{\mathcal{Y}}\sqrt{\frac{2d\ln(T/\delta)}{p_{min}\mu(\xi)T}}\Big)\right\}.$$

Theorem 3.6.1 provides an upper bound for the risk of h_T, and this upper bound $\mathcal{R}(T, c_{\mathsf{min}})$ has two parts. The first part $\varphi(T)$ represents the upper bound for the non-asymptotic convergence rate, while the other part is an upper bound for the risk of the predictor when T tends to infinity. The function $\varphi(T)$ is the minimum value of two upper bounds, where the first term is at most $\tilde{\mathcal{O}}(T^{-1/2})$, which is in the same order as the typical supervised learning. The second term in $\varphi(T)$, which depends on Ψ, can achieve a smaller rate than $\tilde{\mathcal{O}}(T^{-1/2})$ under certain conditions, which will be shown later in Proposition 3.6.1.

The upper bound $\beta\mathcal{R}(T, c_{\mathsf{min}})$ in Theorem 3.6.1 is smaller than Theorem *3.5.1.(1)* in two ways. First, by considering the near-degeneracy function, the convergence rate $\varphi(T)$ can be smaller than $\tilde{\mathcal{O}}(T^{-1/2})$. Secondly, the final convergence result of $\mathcal{R}(T, c_{\mathsf{min}})$ can be smaller than c_{min}/β of Theorem *3.5.1.(1)*. Recall that Theorem 3.5.1 indicates that when the minimum incentive is positive, i.e., $c_{\mathsf{min}} > 0$, we will stop surveying customers when T is larger than some threshold, and the final risk of the predictor h_T will stop at some value that is at most c_{min}. Theorem 3.6.1 further demonstrates that even when $c_{\mathsf{min}} > 0$, as long as $\Psi\Big(\frac{\sqrt{2}c_{\mathsf{min}}}{\beta\underline{\mu}\sqrt{\min\{z,d-z\}}}\Big) = 0$, the risk of h_T will converge to zero. Intuitively, it means that when $\mathbb{E}[\mathbf{y}|\xi]$ is allocated far from the degeneracy and the minimum incentive is small, we can distinguish the optimal products for all type ξ efficiently and the risk of h_T will converge to zero. Theorem 3.6.1 shows that when the near degeneracy function Ψ is smaller (i.e., when it is easier to distinguish the optimal decisions from the sub-optimal decisions), both the convergence rate and the final convergence result of the risk get smaller.

Next, we analyze the cumulative label cost, and provide a smaller bound than Theorem *3.5.1.(2)*. In particular, we want to show that the cumulative label cost is sublinear in T, even when $c_{\mathsf{min}} = 0$.

Theorem 3.6.2 (Bound for the cumulative label cost)**.** *Under the same setting of Theorem 3.6.1, after considering T customers, we have the following bounds for the expectation of the cumulative label cost:*

(1) For any $c_{min} \geq 0$, the expectation of the cumulative label cost is at most

$$c_q + 2\delta + \min\left\{8\sqrt{\min\{z, d-z\}}\eta_{\mathcal{Y}}\sqrt{\frac{dT\ln(T/\delta)}{p_{min}\underline{\mu}}}, \sqrt{d}\eta_{\mathcal{Y}}\sum_{t=1}^{T}\Psi\Big(4\eta_{\mathcal{Y}}\sqrt{\frac{2d\ln(t/\delta)}{p_{min}\underline{\mu}t}}\Big)\right\},$$

where c_q is a constant that is independent of T.

(2) If $c_{min} > 0$, then the cumulative label cost is finite.

Theorem 3.6.2 provides two upper bounds for the cumulative label cost up to time t when $c_{\min} \geq 0$. The first term in the minimization implies that, without considering the near-degeneracy condition, the cumulative label cost is at most $\tilde{\mathcal{O}}(T^{1/2})$. The second term in the minimization provides another upper bound that depends on the near-degeneracy function Ψ. The second term can be much smaller than the first term when function Ψ is small. If the minimum incentive is positive, the proof of Theorem 3.6.2 shows that we will stop exploiting the customers with type ξ, when the number of labeled samples $n(\xi) \geq \underline{t}_\xi := \inf\{t : \beta\sqrt{2\min\{z, d-z\}}\mu(\xi)\rho_t(\xi) \leq c_{\min}\}$. Thus, the cumulative label cost is finite and smaller than a constant which is independent of T. This constant is provided in the proof of Theorem 3.6.2, which is an increasing function of β. It implies that a larger market size results in a larger cumulative label cost but ends up with a smaller regret.

Combining the insights from Theorem 3.6.1, we conclude that when $c_{\min}$ becomes larger, the cumulative label cost gets smaller but the final risk gets larger. Thus, when minimizing the comprehensive cost $\mathscr{C}(\mathbf{c}, g)$, the value of minimum incentive $c_{\min}$ controls the tradeoff between the final risk and the cumulative label cost. When $c_{\min}$ becomes larger, i.e., we have to spend more time or cost to collect and analyze the results of surveys, and we will end up with a larger final risk with a smaller number of surveys. Theorem 3.6.3 further shows that this tradeoff in Algorithm 2 results in a smaller comprehensive cost than the simple supervised learning that offers fixed incentives to all customers.

Theorem 3.6.3 (Comparison with supervised learning)**.** *Under the same setting of Theorem 3.6.1, when achieving the same level of risk, we have the following guarantees for the cumulative label cost:*

(1) If $c_{min} = c_{max}$, then at any time T, the comprehensive cost $\mathscr{C}(\mathbf{c}, g)$ of Algorithm 2 is no more than the cumulative label cost of the supervised learning algorithm that offers fixed incentives c_{min} to all customers.

(2) If $c_{min} = 0$ and $p(c)$ is a twice-differentiable increasing concave function, then at any time T, the expected comprehensive cost $\mathscr{C}(\mathbf{c}, g)$ of Algorithm 2 is no more than the expected comprehensive cost of the supervised learning algorithm that offers c_{max} to all customers.

(3) If $c_{min} > 0$, then there exists a time point $T_s > 0$: when the time $T > T_s$, the comprehensive cost $\mathscr{C}(\mathbf{c}, g)$ of Algorithm 2 is no more than that of the supervised learning algorithm that offers a fixed incentive between $[c_{min}, c_{max}]$ to all customers.

In Theorem 3.6.3.(1), the personalized incentive problem reduces to the customer selection problem when $c_{\min} = c_{\max}$. Theorem 3.6.3.(1) shows that the comprehensive cost of our active label acquisition algorithm is always smaller than the algorithm that selects all customers. In Theorem 3.6.3.(2), compared to the supervised learning algorithm that offers $c_{\max}$ incentives to all customers, Algorithm 2 selects some customers to offer a smaller or zero incentive. This may increase the risk of the final prediction model because it will reduce the probability

for customers to take the survey and thereby reduce the number of samples in the training set. However, Theorem 3.6.3.(2) shows that this increase of risk is smaller than the saved label cost, so Algorithm 2 has a smaller comprehensive cost. Theorem 3.6.3.(3) further shows that when $c_{\mathsf{min}} > 0$, Algorithm 2 can always achieve a smaller comprehensive cost than the supervised learning that offers any fixed incentive between $[c_{\mathsf{min}}, c_{\mathsf{max}}]$, when the number of customers considered is larger than some number T_s.

3.6.2 Small Label Complexity Under the Low-noise Condition

Theorem 3.6.2 shows that when $c_{\mathsf{min}} > 0$, the comprehensive cost is finite. Actually, when $c_{\mathsf{min}} = 0$, the comprehensive cost can still be finite when the near-degeneracy function Ψ satisfies some conditions, which is defined as the low-noise condition in this section.

We characterize the noise level by the degree of near-degeneracy function in Assumption 2.5.1, with parameter κ. Since $\rho \leq \eta_{\mathcal{Y}}$, in Assumption 2.5.1, a larger κ implies a smaller Ψ. By substituting the near-degeneracy function $\Psi(\cdot)$ into Theorems 3.6.1 and 3.6.2 with the upper bound in Assumption 2.5.1, we have the following upper bounds for the comprehensive cost.

Proposition 3.6.1 (Bounds under the low-noise condition)**.** *Under the same setting of Theorem 3.6.2, suppose that Assumption 2.5.1 holds with parameter κ, then we have the following guarantees for Algorithm 2 after T iterations:*

1. *The expectation of cumulative label cost after considering T customers is no more than $\tilde{\mathcal{O}}(T^{1-\kappa/2})$, when $\kappa \leq 2$, and $\tilde{\mathcal{O}}(1)$, when $\kappa > 2$.*
2. *The risk of the predictor h_T is at most $\tilde{\mathcal{O}}(T^{-\kappa/2}) + \frac{c_{min}}{\underline{\mu}} \Psi\Big(\frac{\sqrt{2} c_{min}}{\beta \underline{\mu} \sqrt{\min\{z, d-z\}}}\Big)$.*

Proposition 3.6.1 demonstrates that when $\kappa > 2$, the comprehensive cost of Algorithm 2 is also finite. As a result, when $\kappa > 2$ and the number of iterations T is large, the comprehensive cost of Algorithm 2 is ultimately smaller than the comprehensive cost of the supervised learning that offers fixed positive incentives to all customers.

3.6.3 Active Label Acquisition in Personalized Assortment Optimization

Next, following the analysis in the product selection problem, we have the following bounds for both the risk and cumulative label cost for our active label acquisition algorithm in the assortment optimization problem.

Theorem 3.6.4 (Guarantees for the assortment optimization problem)**.** *Suppose Assumption 3.5.1 holds. Set $\rho_t(\xi_t) = 2\eta_{\mathcal{Y}} \sqrt{\frac{d \ln(t/\delta)}{n(\xi_t)}}$. Using $U_M^A(\xi_t, \hat{\mathbf{y}}, \rho_t(\xi_t))$ in Algorithm 2, for the assortment optimization problem, we have the following guarantees:*

(1) After T iterations, for any $\delta \in (0,1)$, with probability at least $1 - mTe^{-\frac{\underline{\mu} p_{min} T}{8}} - mTe^{-T\ln(\frac{1}{1-p_{min}\underline{\mu}})} - \frac{\delta}{T^2}$, we have the risk $R(h_T)$ is at most $\beta\mathcal{R}(T, c_{min})$, where

$$\mathcal{R}(T, c_{min}) := \varphi(T) + \frac{c_{min}}{\beta\underline{\mu}}\Psi\Big(\frac{\sqrt{2}c_{min}}{\varkappa\beta\underline{\mu}\sqrt{\min\{z, d-z\}}}\Big),$$

and function $\varphi(T)$ is defined as

$$\varphi(T) := \min\Bigg\{\eta_{\mathcal{Y}}\varkappa\sum_{\xi\in[m]} 4\mu(\xi)\sqrt{\min\{z,d-z\}}\sqrt{\frac{d\ln(T/\delta)}{p_{min}\mu(\xi)T}},$$
$$\sqrt{d}\eta_p\sum_{\xi\in[m]}\Psi\Big(4\eta_{\mathcal{Y}}\sqrt{\frac{2d\ln(T/\delta)}{p_{min}\mu(\xi)T}}\Big)\Bigg\}.$$

(2) After T iterations, the expectation of the cumulative label cost is at most

$$c_q + 2\delta + \min\left\{8\sqrt{\min\{z,d-z\}}\eta_{\mathcal{Y}}\kappa\sqrt{\frac{dT\ln(T/\delta)}{p_{min}\underline{\mu}}}, \sqrt{d}\eta_p\sum_{t=1}^{T}\Psi\Big(4\eta_{\mathcal{Y}}\sqrt{\frac{2d\ln(T/\delta)}{p_{min}\underline{\mu}t}}\Big)\right\},$$

where c_q is a constant that is independent of T. If $c_{min} > 0$, then the cumulative label cost is finite.

(3) Theorem 3.6.3 and Proposition 3.6.1 still hold for the assortment optimization problem.

Theorem 3.6.4 provides upper bounds for the risk of the prediction model and the cumulative label cost, which have the same order as the product selection problem. It demonstrates that our active label acquisition algorithm can achieve a much smaller comprehensive cost than the supervised learning algorithm when function Ψ is small.

The results and algorithm in Theorem 3.6.4 can be extended to the general random utility maximization (RUM) choice model when the retailer recommends only one product in the assortment. When the choice probability satisfies the Lipschitz continuity, the above analysis still holds, which will be discussed in Section 3.8.

3.7 Extension to Active Label Acquisition with Contextual Information

In previous sections, we have assumed that customers are categorized into finite types. The recommended (or selected) products are the same for the same type of customers. In this section, we consider an extension of this setting, where we can observe the contextual information in addition to the customer type. This contextual information can help us

customize the selected products for the customers within one type. Given this additional contextual information, we study how to determine the incentive for each customer.

At each iteration t, we observe not only the type of customer t, but also a feature vector x_t. The feature space of x_t is denoted by $\mathcal{X}^{\xi_t}$, which is continuous, bounded and dependent on type ξ_t. For example, if a customer type ξ indicates that she has a membership, then $\mathcal{X}^{\xi}$ may include the membership information. If a customer type ξ indicates that she is referred by a friend, then $\mathcal{X}^{\xi}$ may include her friend information as well.

The feature vector x for type ξ comes from a fixed and known distribution whose probability density at x is denoted by $\mu_\xi(x)$. Furthermore, given the side information (ξ, x), the true model for the utility vector y is assumed to be $\mathbf{y} = h^*_\xi(x) + \epsilon$, where $\epsilon \in \mathbb{R}^d$ is the noise term with zero mean, and $h^*_\xi : \mathcal{X}^\xi \to \mathbb{R}^d$ is the true prediction for type ξ. Since h^*_ξ is unknown, we need to estimate the predictor from a hypothesis class $\mathcal{H}_\xi$, where we assume the hypothesis class is well-specified, i.e., $h^*_\xi \in \mathcal{H}_\xi$.

At each iteration, customers of the same type share the same incentive, and thus they have the same probability of being included in the training set. It implies that the samples in the training set of type ξ_t, $\mathcal{S}_{t-1}(\xi_t)$, are i.i.d. This i.i.d. property enables us to use any supervised learning oracles to calculate the predictor for type ξ, which is denoted by $h_{t-1,\xi}$. For example, given the training set for type ξ_t, $\mathcal{S}_{t-1}(\xi_t)$, we can minimize the empirical squared loss of the predictors to obtain h_{t-1,ξ_t}.

One challenge of calculating the upper bound for the value of one data point lies in the fact that the predictions of different features are correlated. Thus, when calculating the upper bound of the value of one data point for type ξ, we need to consider all the features within the space $\mathcal{X}^\xi$. Suppose the prediction error for any feature $x \in \mathcal{X}^\xi$ is at most ρ, i.e., $\sup_{x\in\mathcal{X}^\xi}\{\|h_{t-1,\xi}(x) - h^*_\xi(x)\|\} \le \rho$. Then, by taking the expectation of the upper bound in Theorem 3.3.1 over $x \sim \mu_\xi(x)$, we obtain that the upper bound for type ξ can be written as

$$\int_{x\in\mathcal{X}^\xi} \mu_\xi(x)\sqrt{2\min\{z, d-z\}}\beta\rho\mathbb{I}\left\{\nu_S(h_{t-1,\xi}(x)) \le \rho\right\} dx.$$

The upper bound for ρ can be obtained by the theoretical guarantees in the supervised learning, since the samples contained in the training set $\mathcal{S}_{t-1}(\xi_t)$ are i.i.d.. We assume that the prediction error (in terms of the sup norm) of $h_{t-1,\xi}$ shrinks in a certain order, which is stated in Assumption 3.7.1. Let $\mathbb{Z}^+$ denote the set of non-negative integers $\{0, 1, 2, ...\}$.

Assumption 3.7.1 (Prediction errors for supervised learning.)**.** *There exists a function* $\Phi(n, \xi, \delta) : \mathbb{Z}^+ \times [d] \times (0, 1) \to \mathbb{R}$ *such that, when customers with type* ξ *have* n *i.i.d. observations of* (x_i, ξ, y_i) *in the training set* $\mathcal{S}_{t-1}(\xi)$*, then* $\sup_{x\in\mathcal{X}^\xi}\{\|h_{t-1,\xi}(x) - h^*_\xi(x)\|\} \le \Phi(n, \xi, \delta)$ *with probability at least* $1 - \delta$*.*

This function $\Phi(n, \xi, \delta)$ exists for a broad range of hypothesis classes in supervised learning. In Section 3.9, we consider $\mathcal{H}^\xi$ as a class of linear functions, or decision trees, and provide some conditions for $\Phi(n, \xi, \delta)$ to be at most $\tilde{\mathcal{O}}(\sqrt{\ln(1/\delta)/n})$. Based on this function, Algorithm 3 shows how to calculate the upper bound for the value of one data point for type ξ_t customer.

Algorithm 3 Calculate incentive $U(\xi_t; \mathcal{S}_{t-1})$ with contextual information

Input: Type of customer t, ξ_t; training set at iteration $t-1$, $\mathcal{S}_{t-1}$; contextual information for customer t, x_t; confidence level $\delta \in (0,1)$.
$\rho \leftarrow \Phi(|\mathcal{S}_{t-1}(\xi_t)|, \xi_t, \delta)$.
Return $U(\xi_t; \mathcal{S}_{t-1}) \leftarrow \mu(\xi)\sqrt{2\min\{z, d-z\}}\beta\rho\mathbb{P}_{x_t\sim\mathcal{D}_\xi}(\nu_S(h_{t-1}(\xi, x_t)) \le \rho)$.

To characterize the near-degeneracy condition within each type ξ, we define a near-degeneracy function for type ξ as $\Psi_\xi(\rho) := \mathbb{P}_{x\sim\mu_\xi(x)}\big(\nu_S(h^*_\xi(x)) \le \rho\big)$. Based on this function Ψ_ξ, Theorem 3.7.1 demonstrates the bounds for both parts in the comprehensive cost.

Theorem 3.7.1 (Guarantees for the product selection with contextual information)**.** *Suppose Assumptions 3.5.1 and 3.7.1 hold. Given $\delta \in (0,1)$, define $\rho_t := \max_{\xi\in[m]}\{\Phi(\lfloor 0.5 p_{\min} t\underline{\mu}\rfloor, \xi, \delta)\}$. When using Algorithm 3 to calculate the incentive, then for the active label acquisition Algorithm 2 of the product selection problem, we have the following guarantees:*

(1) After T iterations, suppose the output predictor is h_T. Then, with probability at least $1-\delta$, we have the risk $R(h_T) - R(h^*)$ is at most $\beta\mathcal{R}(T, c_{\min})$, where

$$\mathcal{R}(T, c_{\min}) := \varphi(T) + \frac{c_{\min}}{\beta},$$

where $\varphi(t)$ is defined as

$$\varphi(T) := \min\left\{\sqrt{2\min\{z, d-z\}}\rho_T, \sqrt{d}\eta_{\mathcal{Y}}\sum_{\xi\in[m]}\mu(\xi)\Psi_\xi(2\rho_T)\right\}.$$

(2) After considering T customers, the expectation of the cumulative label cost is at most

$$\min\left\{\sum_{t=1}^{T}\beta\sqrt{2\min\{z, d-z\}}\rho_t, \sqrt{d}\eta_{\mathcal{Y}}\sum_{t=1}^{T}\Psi(2\rho_t)\right\}.$$

If $c_{\min} > 0$, the cumulative label cost is finite.

(3) Theorem 3.6.3 still holds. When $\Phi(n, \xi, \delta)$ is at most $\tilde{\mathcal{O}}(\sqrt{\ln(1/\delta)/n})$, Proposition 3.6.1 holds as well.

Theorem 3.7.1.(1) demonstrates that convergence rate $\varphi(T)$ can be bounded in two distinct ways. The first type of upper bound in $\varphi(T)$ depends on ρ_t, whose order is at most $\tilde{\mathcal{O}}(T^{-1/2})$ when the condition $\Phi(n, \xi, \delta) \le \tilde{\mathcal{O}}(\sqrt{\ln(1/\delta)/n})$ holds. The second type of upper bound depends on Ψ_ξ, which can be much smaller $\tilde{\mathcal{O}}(T^{-1/2})$ when Ψ_ξ is small. Notably, compared to Theorem 3.6.4 where $\varphi(T)$ depends on Ψ, Theorem 3.7.1.(1) delves into the distribution of $h^*_\xi(x)$ for each type and provides a tighter bound that depends on Ψ_ξ. It implies if $h^*_\xi(x)$ is

allocated further away from the degeneracy than $\mathbb{E}[y|\xi]$, incorporating contextual information can further improve the convergence rate $\varphi(T)$. Theorems 3.7.1.(2) and 3.7.1.(3) are similar to the results in Theorem 3.6.4.

Similar to Section 3.4, the results in Theorem 3.7.1 can be extended to the personalized assortment optimization problem with the MNL model, if we multiply ρ_t in Theorem 3.7.1 by $\varkappa$ and replace the maximum satisfaction level $\sqrt{d}\eta_{\mathcal{Y}}$ with the maximum revenue loss $\sqrt{d}\eta_p$. Moreover, the results can further be extended to the assortment optimization problem with the general RUM choice model and one capacity. Please see the details in Section 3.8.

3.8 Extension to the General RUM Choice Model

In this section, we demonstrate that the results and algorithm in Theorem 3.6.4 can be extended to the general random utility maximization (RUM) choice model when only one product is recommended in the assortment.

We consider the case where the choice of customers follows the random utility maximization (RUM) model, and thus customers pick the recommended product when its utility is larger than the utility of the no-purchase option. We use $\bar{\mathbf{y}}$ to denote the expectation of the utility vector.

Then, we use $\phi(i;\bar{\mathbf{y}})$ to denote the purchase probability of product i when the expected utility vector is $\bar{\mathbf{y}}$ and product i is recommended. By this definition, we assume the purchase probability $\phi(i;\bar{\mathbf{y}})$ for product i is a function of the expected utility $\bar{\mathbf{y}}$ when only product i is in the assortment.

Thus, when the designer would like to maximize the revenue from only one product, the problem can be re-written in (P4):

$$\begin{aligned} \max_{w\in\mathbb{B}^d} \quad & \sum_{i\in[d]} w_i p_i u_i && \text{(P4)} \\ s.t. \quad & w^T\mathbf{1} = 1 \\ & u_i = \phi(i;\bar{\mathbf{y}}), \quad \forall i\in[d] && (3.8) \end{aligned}$$

Objective (P4) is linear and constraints (3.8) are nonlinear.

Since the retailer only recommends one product in the assortment, obviously, to maximize the expected revenue, the best assortment is the product with the highest $p_i\phi(i;\bar{\mathbf{y}})$.

We assume that the choice probability has the following property.

Assumption 3.8.1. *There exists a constant $\eta_b > 0$, such that for product i, the choice probability function $\phi(i;\bar{\mathbf{y}})$ satisfies that $\phi(i;\bar{\mathbf{y}}_1) - \phi(i;\bar{\mathbf{y}}_2) \le \eta_b|\bar{y}_1^i - \bar{y}_2^i|$.*

Assumption 3.8.1 holds for the MNL model, NL model, probit model and some other RUM models. Assumption 3.8.1 means that we can construct a confidence region based on the confidence region of y_t^i. Then, we can show that the value of one data point for type ξ_t is

no more than

$$V(\xi_t, \mathcal{S}_{t-1}) \leq \min\left\{\eta_p\eta_b\sqrt{2\min\{z, d-z\}}\mu(\xi_t)\rho_t\mathbb{I}\{\nu_S(\hat{\mathbf{y}}) \leq \rho_t\}, 1\right\} \tag{3.9}$$

.

Thus, we have the following Theorem 3.8.1 for Algorithm 2 in the purchase probability maximization problem with RUM choice model.

Theorem 3.8.1. *Suppose that $\eta_p \geq 1$ and Assumption 3.8.1 holds. If we replace the parameter $\varkappa$ in Theorem 3.6.4 with $\eta_p\eta_b$, then the arguments in the modified theorem hold for the purchase probability maximization problem with RUM choice model.*

When we have additional contextual information within one type, similar to Theorem 3.7.1 in Section 3.7, the results in Theorem 3.8.1 can also be extended to the general RUM model, by using the similar proofs and multiplying ρ_t in 3.7.1.(2) by $\eta_b\eta_p$.

3.9 Examples of Function $\Phi(n, \xi, \delta)$

In this section, we provide an example of $\Phi(n, \xi, \delta)$ for the linear class. Let us consider one type of customer, which is denoted by type ξ. Suppose the dimension of $\mathcal{X}^\xi$ is m_ξ. Suppose there exists a preference matrix $\Theta^* \in \mathbb{R}^{m_\xi \times d}$ for customers, such that for one type ξ, the true model is $\mathbf{y}_t = \Theta^* x_t + \epsilon_t$, where $\epsilon_t \in \mathbb{R}^d$ is the noise term with zero mean. Then, we make the following assumption on the distribution of $\mu_\xi(x)$.

Assumption 3.9.1 (Isometry condition)**.** *There exists $\underline{\lambda} > 0$, such that the distribution of x, $\mu_\xi(x)$ satisfies that $\lambda_{min}(\mathbb{E}[xx^T]) \geq 2\underline{\lambda} > 0$.*

We use $\lambda_{\min}(\cdot)$ and $\lambda_{\max}(\cdot)$ to denote the minimum and maximum eigenvalue of a matrix. We assume the absolute value of each entry in noise vector $\epsilon_t = \mathbf{y}_t - \Theta^* x_t$ is less than σ_ϵ. Intuitively, σ_ϵ controls the noise level. We define $\mathbf{\Lambda}_t := \sum_{(x,\cdot)\in W_t} xx^T$, and $\mathbf{X}_t^T := [x_1, x_2, ..., x_{n_t}]$. Thus, we have that $\mathbf{\Lambda}_t = \mathbf{X}_t^T\mathbf{X}_t$.

Suppose that the feature space has a bounded set, in other words, $\forall x \in \mathcal{X}, \|x\|_2 \leq \eta_{\mathcal{X}}$. Then, we have that $\lambda_{\max}(XX^T) \leq \eta_{\mathcal{X}}^2$, $\forall x \in \mathcal{X}$. When the distribution of x satisfies Assumption 3.9.1, by the matrix Chernoff's inequality (Theorem 1.1) in Tropp, 2012, we have that with probability at least $1 - \sqrt{2}m_\xi e^{-\underline{\lambda}n_t/(2\eta_{\mathcal{X}}^2)}$,

$$\lambda_{\min}(\mathbf{\Lambda}_t) \geq \frac{n_t 2\underline{\lambda}}{2} = n_t\underline{\lambda}.$$

Next, Lemma 3.9.1 provides an example of $\Phi(n, \xi, \delta)$ in this case.

Lemma 3.9.1 (Example of $\Phi(n, \xi, \delta)$)**.** *Suppose Assumption 3.9.1 holds. Then, for any $\delta \in (0, 1)$, $\Phi(n, \xi, \delta)$ satisfies*

$$\Phi(n, \xi, \delta) \leq \tilde{\mathcal{O}}\left(\sqrt{n\ln(\frac{1}{\delta})}\right).$$

3.10 Numerical Experiments

In this section, we evaluate the performance of our proposed active label acquisition algorithms in two settings. First, we consider the product selection problem and use real-world campus survey data. Second, we evaluate the assortment optimization applications with MNL choice model. The results from both settings verify that our algorithms can significantly reduce the comprehensive cost.

3.10.1 Product Selection Problem

In this experiment, we use a real-world survey dataset that consists of student responses to the survey on ideal student life [4]. The dataset covers various aspects such as student interests and activities. Our objective is to leverage these survey results to predict students' potential interests and provide recommendations about student groups for each individual.

The surveyed students are from 21 different departments, which are used as the type to predict their interests. There are six interest groups: *art and culture*, *science and technology*, *social welfare and diversity*, *entrepreneurship*, *sports*, and *others*. The goal is to recommend two of the six groups to each student, based on their department information, to maximize their satisfaction level. To illustrate the application of this setting, imagine that we are the organizers of the student interest group, and we would like to provide some guides for freshman students based on their department information. Although freshmen may have limited knowledge about the experiences of each group, we can collect survey results from students in higher grades. To encourage students to fill out surveys, we need to provide some incentives. Here, we use the active label acquisition algorithm to decide the personalized incentives for each student. The incentive we offer is assumed to be within $[c_{\mathsf{min}}, c_{\mathsf{max}}] = [\$15, \$30]$. We can also provide zero incentive for students and not collect their responses if their ratings are not informative. The probability for one student to fill out the survey is assumed to be as follows:

$$p(c) = \frac{c - c_{\mathsf{min}}}{c_{\mathsf{max}} - c_{\mathsf{min}}} \times 0.6 + 0.2\,.$$

This is a linear function of c between two points $(c_{\mathsf{min}}, 0.2)$ and $(c_{\mathsf{max}}, 0.8)$. The rating of each group from one student is an integer from 0 to 17. Please refer to Appendix B.5 for the details on how to obtain the rating of all groups from the answer of one survey. We assume that when recommending two groups to one student, the satisfaction level of this student is the sum of his ratings of two groups.

There are 2958 rows in the dataset. We randomly select 30% of them as the test set, and the rest as the training set. At each iteration of the survey distribution process, we randomly select one student from the training set to offer some incentive. We either provide some positive incentive between \$15 and \$30 for this student or provide zero incentive to skip her response. We run 20 independent trials for each setting.

[4] `https://www.kaggle.com/datasets/shivamb/ideal-student-life-survey`

Figure 3.3 shows the comprehensive cost with the 95% confidence intervals during the survey distribution process. The x-axis represents the number of students considered, which is denoted by T in the algorithm. The two plots in Figure 3.3 represent market sizes β of 1000 and 500, respectively. The first plot shows that our personalized incentive policy always yields a lower comprehensive cost compared to the fixed incentive policy, which offers \$15, \$20, \$25, or \$30 to all students, particularly when the number of students is large. By comparing these two plots, we observe that as the market size β increases from 500 to 1000, the comprehensive cost of the personalized incentives becomes closer to the fixed incentive policy with \$15. In the second plot, as T increases, our personalized incentive policy eventually achieves a smaller comprehensive cost compared to the fixed incentive policy.

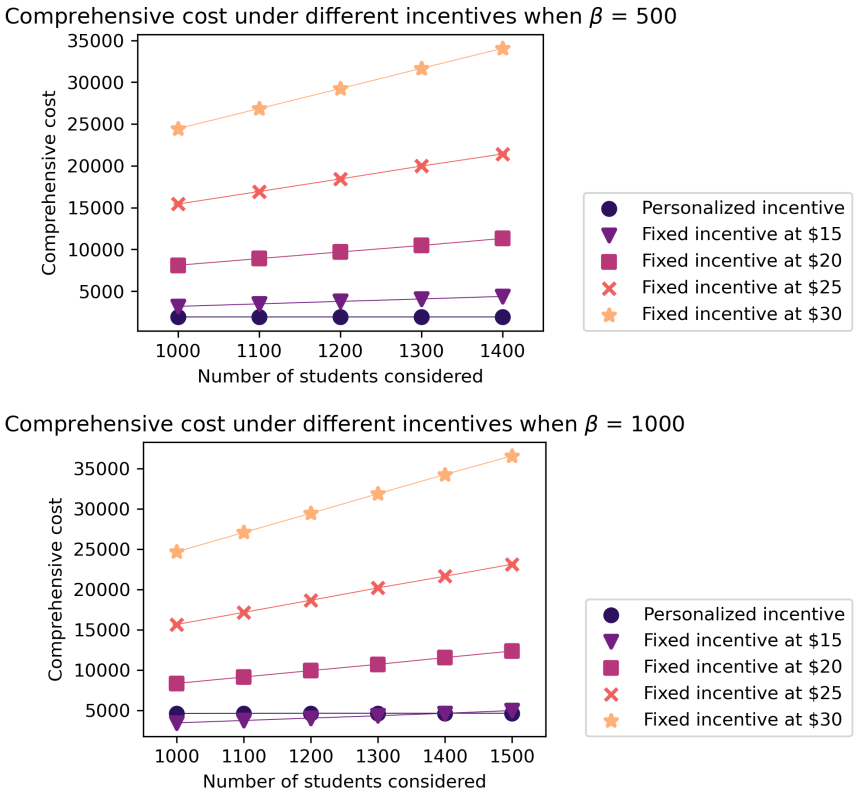


Figure 3.3: Comprehensive cost of the active label acquisition algorithms with different market sizes β.

Figure 3.4 further displays the two parts of the comprehensive cost during the survey process. It shows that as the market size increases (from $\beta = 500$ to $\beta = 1000$), the expected regret in the test risk will converge to a smaller value, and the total survey cost becomes larger. This observation aligns with the insights regarding the tradeoff between the survey cost and the final risk.

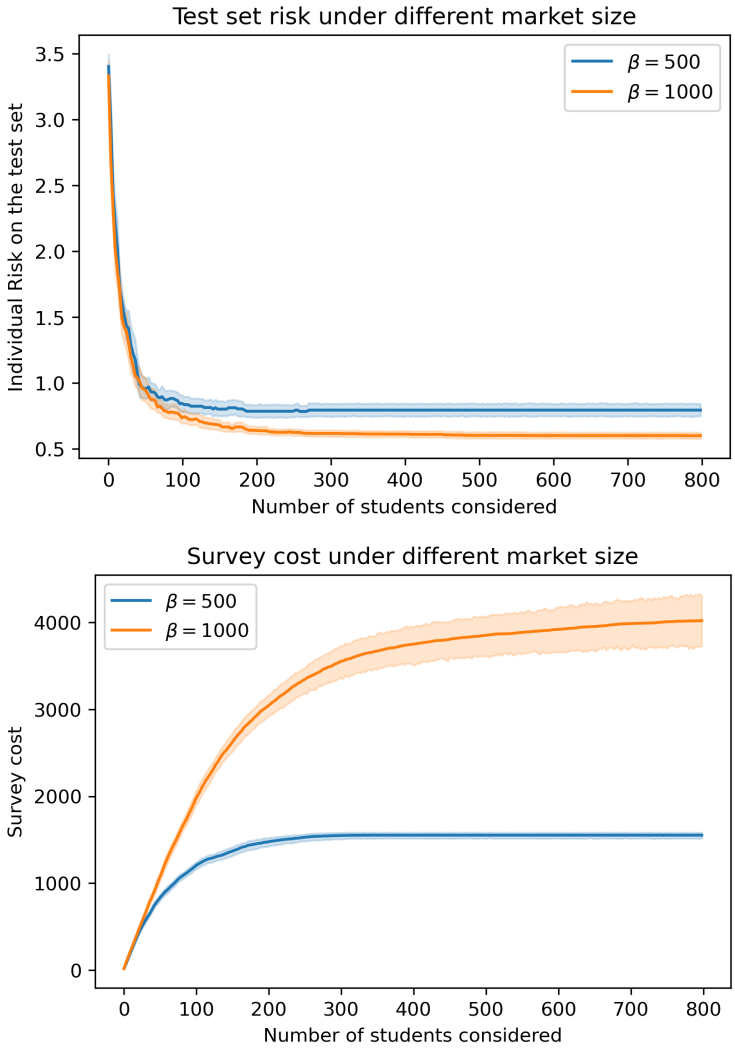


Figure 3.4: Risk and the total survey cost with different market sizes β.

3.10.2 Assortment Optimization

In the second experiment, we consider the assortment optimization problem using synthetic data. We assume there are 5 different types of customers. Each type of customer has an 8-dimensional feature vector. The goal is to select at most 5 products in the assortment out of 10 products to maximize the revenue. The utility vector for each type of customer is generated as follows. For each type, we randomly generate a coefficient matrix Θ^* whose dimension is 10×8. Specifically, the value of $\Theta^* = \frac{1}{8000}\sum_{i=1}^{8} \zeta_i^s \zeta_i^{pT}$, where $\zeta_i^s \in \mathbb{R}^{10}$ and $\zeta_i^p \in \mathbb{R}^8$ are random vectors whose values follow uniform distribution between $[1, 10]$ and $[10, 20]$. For each dimension of the feature, the value follows a uniform distribution between

$[0, 10]$. Given feature x, the average utilities for the 10 products are $\Theta^* x$. The realized utilities for one customer is $\Theta^* x$ plus the noise, where the noise follows the Gumbel distribution with standard deviation $\sigma = 1$. The revenue for product i is $5\epsilon_i + 2000$, when ϵ_i follows a normal distribution. For the distribution of each type of customers, we assume the probability of encountering type ξ is $\mu(\xi) = \xi/15$, for $\xi \in [5]$. The market sizes are set to be $500, 750$, and 1000. For each customer, we can either offer some incentive between \$20 and \$40 or zero incentive. The probability of taking the survey is a linear function that goes through two points $(c_{\mathsf{min}}, 0.3)$ and $(c_{\mathsf{max}}, 0.9)$. The test set is of size 3000.

We implement Algorithm 3 to provide personalized incentives to each type of customer. When updating the incentives in Algorithm 3, for each customer with side information (ξ, x), given the estimated $\hat{\Theta}$, we calculate $\mathbb{P}(\nu_S(\hat{\Theta}x) \leq \rho)$, which requires the estimation of the distribution of $\nu_S(\hat{\Theta}x)$. This distribution is estimated using the features of the entire training set, which contains 7000 features. To speed up the algorithm, we update distribution of $\nu_S(\hat{\Theta}x)$, only at iteration t, when $t = 2^{\tilde{n}}$, for some $\tilde{n} \in \mathbb{Z}^+$. For each setting, we run 20 independent random trials. The comprehensive costs during the survey distribution process with the market size 500 and 1000 are shown in Figure 3.5.

Figure 3.5 illustrates that for both market sizes, as T approaches 1500, our personalized policy incurs lower comprehensive costs compared to any fixed incentive policy. Furthermore, in Figure 3.6, we examine the relationship between risk and cumulative label cost. The left plot presents the minimum cumulative label costs required to achieve the same risk level, averaged over 20 trials with 95% confidence intervals when $\beta = 1000$. The x-axis represents the risk of the predictor, while the y-axis represents the corresponding cumulative label cost. As the risk decreases (towards the left side of the x-axis), more label cost is needed. Thus, the cumulative label cost decreases with the risk of the predictor. The results demonstrate that our personalized incentive policy requires significantly less label cost than the fixed incentive policy to achieve the same risk level. This observation aligns with the alternative formulation of the comprehensive cost introduced in Remark 3.5.3.

The right plot in Figure 3.6 further shows the average risk of the predictor when spending the same amount of incentives with $\beta = 750$. The x-axis represents the cumulative incentives and the y-axis represents the corresponding risk at that incentive, averaged over 20 trials. It shows that compared to the fixed incentive policy, our personalized incentive policy achieves a much smaller risk when spending the same amount of label cost.

Table 3.1 further examines the advantage of our personalized incentive policy over the fixed incentive policy. It shows that when achieving the same excess risk level (\$5000), our personalized incentive policy can reduce the label cost by over 70%, compared to the fixed incentive at \$20, \$30 and \$40. Our personalized incentive policy also reduces the size of the training set by over 80%.

Next, we examine the impacts of market sizes β and probability $\mu(\xi)$ on the cumulative label cost. Similar to the insights from the product selection problem, the left plot in Figure 3.7 shows that when the market size increases, we need to collect more surveys to achieve a smaller individual risk, which yields a larger survey cost. Recall that there are five types of customers, and each type of customers receives different incentives. To examine the

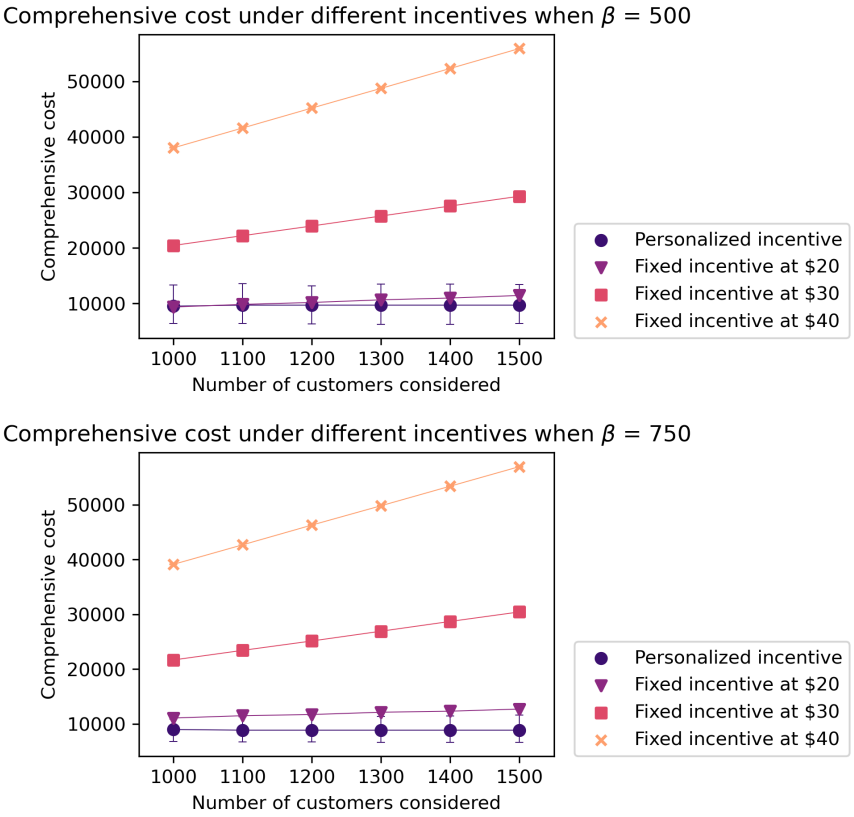


Figure 3.5: Comprehensive cost of the active label acquisition algorithms with different market sizes β.

distribution of incentives across these customer types, the right plot in Figure 3.7 shows the cumulative incentives for each type as T increases. It shows that the majority of incentives are allocated to type 4, while type 0 receives the least incentives. This is because type 4 has the largest probability to occur, while type 0 has the lowest probability to occur. This observation is consistent with the insights that a higher value of $\mu(\xi)$ corresponds to a higher incentive, as indicated by Theorem 3.3.1.

In summary, by using both real-world and synthetic data, the experiments on active label acquisition in the setting of product selection and assortment optimization problems demonstrate the advantages of our proposed personalized incentive algorithms.

3.11 Conclusion

We propose a new concept to quantify the marginal contribution to the revenue increase when including a new data point into the training set. This concept is called value of

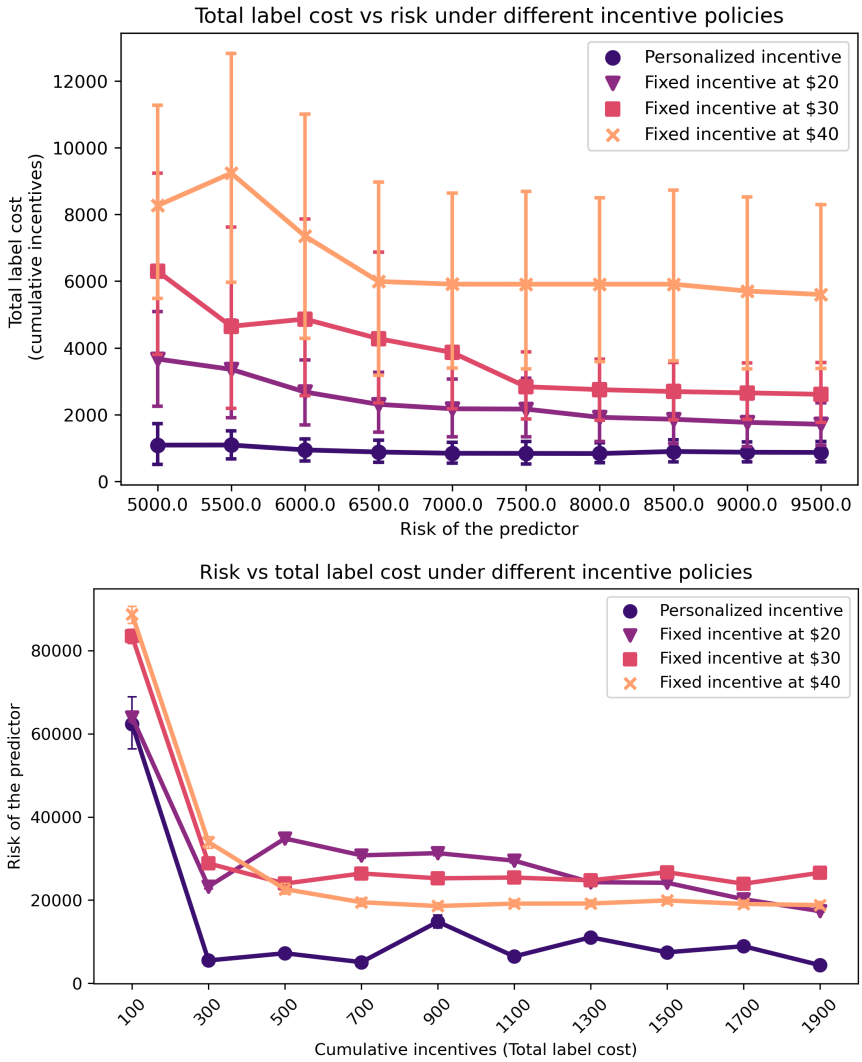


Figure 3.6: Comparison between risk and cumulative label cost.

one data point. We provide an upper bound for estimating the value of one data point, which provides a lot of insights for the label acquisition. We further utilize this upper bound to determine personalized incentives for customers to disclose their preferences in the assortment optimization problem. For both the product selection and assortment optimization problems, our active label acquisition algorithm can achieve a smaller comprehensive cost than supervised learning with fixed incentives under some regularity conditions. When considering additional contextual information for each type of customers, we further provide guarantees for the general prediction models. Our numerical experiments on both synthetic and real-world datasets show that our active label acquisition algorithms can achieve a much smaller comprehensive cost and require much less label cost to achieve the same level of

	Personalized incentive	Fixed $20	Fixed $30	Fixed $40
Required label cost	**1088**	3668 (-70%)	6295 (-79%)	8262 (-87%)
Required number of surveyed customers (Size of training set)	**30**	184 (-84%)	210 (-86%)	206 (-85%)

Table 3.1: Comparison for different incentive policies when achieving the excess risk level of $5000

regret.

There are several interesting future research directions. First, it would be intriguing to study the personalized incentive under more general settings of operations management problems, such as the pricing problem and the assortment optimization under other choice models. Secondly, the incentive function $p(c)$ is unknown in practice. It is interesting to incorporate the estimation of $p(c)$ in the surveying process. For example, retailers can explore the structure of $p(c)$ at different incentives at the beginning of survey distribution. Lastly, if customers can strategically decide whether to take surveys to maximize their utility, retailers need to consider the personalized incentive in a game setting, where offering a high incentive at the beginning may change the structure of the incentive function in the future.

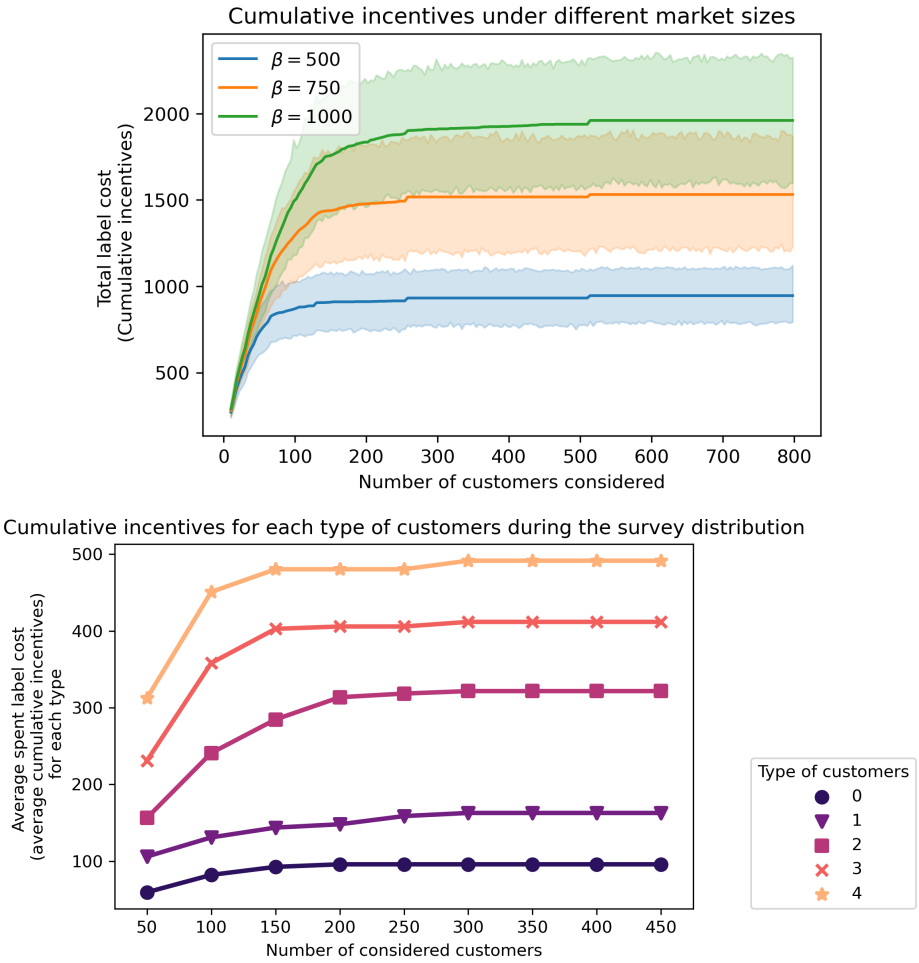


Figure 3.7: Impacts of market sizes β and probability $\mu(\xi)$ on the cumulative label cost.

Chapter 4

Pricing from Click Transition Data

4.1 Introduction

Traditionally, in brick-and-mortar stores, to learn the preference toward products, retailers have to infer from the purchase behavior of customers. These purchase data are usually limited and influenced by the availability of products. The transition of retail business from brick-and-mortar stores to online digital platforms has provided retailers with detailed clickstream data from individual customers. The clickstream data reveals the search and selection process of customers before they either make a purchase or leave the system. This process provides additional information for understanding customer's preferences. By utilizing these click data efficiently, online retailers are able to make better decisions, such as pricing. The finding from the real-world data (in Section 4.5) further demonstrates that incorporating the random click data can significantly enhance the prediction accuracy of customer purchase behaviors and increase the revenue of pricing decisions.

However, how to utilize clickstream data efficiently has been a challenge for online retailers. The reasons are fourfold. First, modeling obstacles arise from the curse of dimensionality associated with tracking click trajectories. The click model should be able to capture the dependence of current browsing products, because the click behavior of customers is heavily influenced by the product they are currently browsing. For instance, if customers like the product they are currently browsing, they are more likely to click on similar products next. Conversely, if customers dislike the current product, this may discourage them from clicking on similar products and even increase the likelihood of leaving the website without making a purchase. Existing literature often focuses on the click-through rate of each single product, but these dynamically changing click behaviors are largely overlooked. For the studies on click trajectory models, such as the cascade model (Craswell et al., 2008; X. Gao et al., 2022), and the click chain model (Guo et al., 2009), they often assume a fixed sequence of click behaviors and thus are unable to capture the random back-and-forth click transition behaviors of customers. To address this first challenge, motivated by the connection between click transitions of customers and state transitions within a Markov chain, we propose a joint

click and pricing model, called *Markovian Dynamic Attraction Click (MDAC)* model. In MDAC model, each product is treated as a state in a Markov chain, where the transition probability between each pair of states follows the typical attraction choice model (Luce, 1959). The attractiveness of one product relative to another is characterized by a term called *relative attraction value*. This term captures the dependence of click transitions on the currently browsed product. Such dependence has been modeled in various ways in prior literature, as exemplified by Besbes, Gur, and Zeevi, 2016; Amaldoss and He, 2018. In our study, by incorporating the click transition behavior using the MDAC model, we reduce the prediction error of purchase behavior by approximately 15% in the test set and increase the expected revenue by around 18% in the real-world dataset.

Given our click model, the second challenge is the scalability issue of learning customer's behaviors. The relative attraction values between all products form the attraction matrix. When the entire product set is large, the attraction matrix between all products has a very large dimension. With limited click and purchase data from each customer, it is difficult to efficiently estimate this entire attraction matrix. To address this issue, we utilize the similarity between products and explore similar click transition patterns. We characterize these similarities by utilizing the low-rank structure of the attraction matrix, which is further verified by using a real-world click dataset from JD.com, 2020. By incorporating the low-rank structure, we can significantly reduce the prediction error for the purchase behavior when using the same amount of click data.

The third challenge of utilizing clickstream data stems from the dynamic availability of products. In practice, given a large number of products, not all of them are available for purchase due to some exogenous factors. These factors include stockouts, limited editions, discontinuation, customer-set filters, and geographical restrictions. For instance, Carmax, an online used car platform, buys cars from customers and sells them online. The availability of each car model is constantly changing, depending on the inflow and outflow of their inventory. Additionally, when browsing cars to purchase, customers may set some specific constraints, such as price, location, or functionality. These constraints will reduce the number of products shown to customers. Some e-commerce sites may gray out or disable the link to the unavailable products, which means that clickability is associated with purchasability. Thus, due to the availability constraint, customers may only consider a limited number of products before making a purchase or leaving the system. Given that online retailers cannot control the availability of products, to prepare for any possible outcome of availability in the future, it is crucial for retailers to understand customers' click and purchase behaviors under any product availability, necessitating the recovery of the entire attraction matrix.

However, estimating the entire attraction matrix is challenging because different customers are presented with different available products, leading to varied click transition patterns. This variation makes it difficult to combine the click data from different customers. To address this issue, based on our MDAC model, we further propose efficient estimation algorithms and pricing strategies that smartly integrate the click data across varying availabilities. By utilizing limited but dynamically changing available product sets, we can effectively recover the entire attraction matrix and thus optimize the pricing decisions with a small regret.

The fourth challenge is how to efficiently infer pricing decisions from the click data. Online retailers seek to adjust their pricing decisions timely as new customer data becomes available. Thus, it is critical to design a computationally efficient and tractable framework to integrate the pricing decisions into the learning process of the click model for online retailers. Given that the pricing decision influences the click behaviors of customers, which in turn affects future data collection and model estimation, the simple greedy pricing policy is usually suboptimal for the joint pricing and learning problem. The design of the optimal pricing policy typically involves the balance between exploration and exploitation, which incurs a huge computational complexity, for example, J. Broder and Rusmevichientong, 2012. However, in our MDAC model, we surprisingly discover that the greedy pricing policy yields a low regret with a small computational complexity. This unexpected conclusion is attributed to the inherent exploration facilitated by the random click behaviors.

4.1.1 Contributions

The contributions of our study are summarized as follows:

1. *Formulation*: Motivated by the similarity between click transitions and state transitions in MDAC, this work is, to the best of our knowledge, the first attempt to use a Markov chain-based model to capture user click trajectories. Compared to other click models, our MDAC model incorporates high-dimensional click transition behaviors and the dynamic availability of products. Our proposed model is supported by various theoretical justifications, and its empirical performance is also validated using real-world data.

2. *Theoretical Contributions:*

   a) *Estimation error bound under the low-rank structure*: We propose efficient algorithms to estimate the parameters in MDAC by using both click data and sales data. To address the scalability issue of estimation, we utilize the similarities between products and exploit the low-rank structure of the attraction matrix, which is verified in a real-world dataset. For the estimation of this low-rank matrix, we introduce a penalty based on the nuclear norm of the attraction matrix. We demonstrate that leveraging the low-rank structure can reduce the estimation error bound for the attraction matrix to $\tilde{\mathcal{O}}(\sqrt{r\ln(n)})$, where n and r denote the number of products and the rank of the attraction matrix, respectively. This error bound is smaller than the results $\tilde{\mathcal{O}}(\sqrt{rn})$ in Kallus and Udell, 2020.

   To address the estimation problem under dynamic availabilities of products, we propose an efficient algorithm to integrate the estimation results from different availabilities. It enables us to recover the entire attraction matrix and has a small estimation error bound under the dynamic availabilities of products.

   b) *Offline Pricing Policy*: We analyze the relations between the optimal prices and the attraction matrix. Surprisingly, we found that a higher click-through rate does

not necessarily lead to a higher optimal price for a particular product. Instead, whether the optimal price goes up or down depends on the notion of *optimal stationary revenue*, which we introduce and define in this chapter.

c) *Greedy Online Pricing Strategies and Regret Bounds*: We consider the joint estimation and pricing problem, where our current pricing decisions will influence the click behaviors of customers and the future estimation of the click model. Taking advantage of the structure of our click model, we provide an exploration-free algorithm where we can greedily set optimal prices to maximize revenue from each customer. We further derive an upper bound for the regret of our online algorithm under the low-rank structure of the attraction matrix, which leads to a regret reduction from $\tilde{\mathcal{O}}(n\sqrt{T})$ to $\tilde{\mathcal{O}}(\sqrt{nrT})$. We additionally derive regret bounds for our greedy online pricing algorithms under dynamic availabilities.

3. *Numerical Performance:*

a) Our investigation of the real-world dataset shows the value and importance of leveraging the click data. Firstly, our results show that using both the click data and purchase data in the MDAC will have a much smaller prediction error for the purchase behavior, compared to the estimation methods that only uses the purchase data. Secondly, our results reveal that the estimation using both click and purchase data can lead to better pricing decisions, and hence yields higher revenue, compared to the estimation that only uses the purchase data.

b) Through extensive numerical experiments on both synthetic and real datasets, we verify the assumptions regarding click behaviors. First, we verified the assumption that customers' currently clicked product can represent their state in the Markov chain. We also verified the low-rank structure. Subsequently, we tested our online algorithm using transaction and click data from JD.COM, a leading online retailer in China. This demonstrates the efficacy and practical value of our greedy pricing policy.

The remainder of the chapter is organized as follows. In Section 4.2, we introduce our click model, MDAC model. In Section 4.3, we propose algorithms to estimate the attraction matrix in the MDAC under dynamic availability of products and characterize the estimation error bound of the attraction matrix with a low-rank structure. In Section 4.4, we study the pricing problem in the MDAC, where we first analyze the properties of optimal prices in the static setting, and then in the online setting, we provide the exploration-free online algorithm. In Section 4.5, we conduct numerical experiments to show superior performances of our proposed algorithm.

4.1.2 Literature Review

Our work contributes to a few streams of literature: *click models*, *joint pricing models*, *online pricing algorithms*, and *low-rank matrix estimation*.

Click models. Characterizing which products customers will click or buy is a fundamental problem in web analysis and online marketing. The click models can be naturally viewed as a discrete choice model. However, the multinomial logit (MNL) model (in (Luce, 1959; McFadden et al., 1973)) or Consider-Then-Choose (CTC) models (in (Qing Liu and Arora, 2011)) cannot capture the sequential search behavior of customers. Thus, various click models have been proposed, for example, the Click-based MNL model (Aouad, J. Feldman, et al., 2019), Engageability model (Besbes, Gur, and Zeevi, 2016), Position-Based model, Cascade Click model, Dependent Click model and Click Chain model. We refer the reader to Chuklin, Markov, and Rijke, 2015 for a detailed review. The cascade click model (Craswell et al., 2008; Z. A. Zhu et al., 2010) is one of the most popular click models. X. Gao et al., 2022 further propose a general cascade click model and analyze the optimal pricing policy and online algorithms. However, several strong assumptions in the cascade click model might restrict its flexibility. For example, the cascade click model assumes that customers click the products according to a fixed sequence of products one-by-one, and will not return to the previously clicked products. However, in reality, customers usually click between products back and forth, to compare products and collect information. Therefore, the no-revisit assumption or the fixed sequence assumption may not hold in reality. In contrast, in our proposed click models, customers can randomly click on products back and forth.

Pricing models based on Markov chain. In the literature, some papers study pricing models based on a Markov chain without click behaviors. For example, in Goutam, Goyal, and Soret, 2019 and Dong, Simsek, and Topaloglu, 2019, they study the purchase behavior of customers by a Markov chain, where state transitions are characterized by a transition matrix. These models are generalizations of the Markov chain choice model (MCCM), a discrete choice model proposed in Blanchet, Gallego, and Goyal, 2016. The flexibility of the MCCM makes it more powerful in describing the complex behaviors of customers (See Berbeglia, Garassino, and Vulcano, 2021; J. B. Feldman and Topaloglu, 2017, Goutam, Goyal, and Soret, 2019; Dong, Simsek, and Topaloglu, 2019 for examples.) However, at the same time, it brings higher complexities to the estimation problem. To estimate parameters by the purchase data, Şimşek and Topaloglu, 2018 use an expectation-maximization algorithm to estimate the transition matrix and Fu and Ge, 2021 adopt the subgradient descent method. S. Li et al., 2022 and Gallego and W. Lu, 2021 consider learning the preferences of customers and deciding assortments at the same time to minimize the regret. These estimation methods only use sales data, while overlooking the click trajectories of customers. In addition to the learning problem, Kleywegt and Shao, 2022 considers the joint pricing and assortment optimization problem under a similar generalized Markov chain choice model.

It should be noted that the choice models in the above literature are not suitable for modeling click behavior. This is because, in these models, the state transition matrix is fixed and independent of the assortment. However, the click transition probability should depend on the availability of products. Indeed, Dong, Simsek, and Topaloglu, 2019 notes that state transitions in the above models are only conceptual transitions and are not the

real transitions of customers. Thus, in our study, apart from the above literature, we propose a novel click model based on a Markov chain that specifically captures customers' browsing and purchase behaviors.

Online pricing. Online pricing is a natural setting when the demand function of customers is unknown and needs to be learned on the fly. Various studies have proposed different algorithms to minimize the regret, which refers to the cumulative revenue loss compared to the optimal pricing policy. Suppose T is the length of the period. J. Broder and Rusmevichientong, 2012 shows a lower bound being $\Omega(\sqrt{T})$ in the general parametric choice model and provides a pricing policy matching this bound. Besbes and Zeevi, 2009 study both parametric and non-parametric cases, and provide the lower bounds of regret in each case. When the inventory is finite, the lower bound of regret can be reduced to $\Omega(\log T)$, which is shown in Boer and Zwart, 2015. Following this work, Qiang and Bayati, 2016 shows that the regret of a greedy iterative pricing policy (i.e., setting the optimal prices based on the current least squared estimation of parameters) can achieve the corresponding lower bound, when the retailers have the extra demand covariates information. Ban and Keskin, 2020 consider the personalized pricing problem with high-dimensional features and infinite inventories, and propose algorithms that achieve the near-optimal regret. When the demand function is time-varying, Keskin and Zeevi, 2017 propose an algorithm to update the estimation of demand functions and change pricing strategies under some "variation" budget of prices, whose regret matches the lower bound.

In addition to the single product pricing problem, determining the prices for multiple items jointly is critical for online platforms when considering the relations between products, for example, the substitution effect. Dong, Simsek, and Topaloglu, 2019 propose algorithms to find the optimal prices under the GMCCM, and study the equilibrium of optimal prices when there are multiple competitors. To solve pricing and assortment jointly and efficiently, various choice models have been proposed and different efficient algorithms have been studied, for example, Paul, J. Feldman, and Davis, 2018, Jagabathula and Rusmevichientong, 2017, Alptekinoğlu and Semple, 2016, Yanqiao Wang and Shen, 2017, Ferreira and Mower, 2022, P. Gao et al., 2021. N. Chen et al., 2021 propose a model-free assortment pricing algorithm using the historical transaction data. When the parameters of the choice model are unknown, the online learning version of the joint pricing and assortment problem has also been studied recently. X. Gao et al., 2022 consider a general click cascade model, and use the idea of the upper confidence bound (UCB) to find the optimal pricing and ranking in the online setting. Miao and X. Chao, 2021 consider the MNL model, and solve the joint pricing and assortment optimization problem using the idea of Thompson sampling.

Low-rank models. Low-rank matrix structure has made its appearance in many different domains such as assortment optimization (Kallus and Udell, 2020), causal inference (Athey et al., 2021; V. F. Farias, A. A. Li, and Peng, 2021; Agarwal et al., 2021), and recommender systems (Jannach et al., 2010). Facing the exploration-exploitation dilemma, recent studies

have formulated rank-one bandits (Katariya et al., 2017), bilinear bandits with low-rank structure (Jun et al., 2019), and low-rank bandits (Yangyi Lu, Meisami, and Tewari, 2021). Z. Zhu et al., 2021 consider estimating the Markov transition matrix under low-rank structures in the offline setting, while our work considers the low-rank matrix estimation in the online setting with pricing decisions. In our work, to estimate the attraction matrix in the MDAC with a low-rank structure, we extend the offline estimation algorithm in Kallus and Udell, 2020 from exponential cases to linear cases and derive a much smaller error bound. Based on this novel error bound, we further develop a small regret bound using the proposed exploration-free online algorithm.

4.2 Click Model with Purchase Behavior

In this section, we incorporate click data into pricing decisions by introducing our Markovian Dynamic Attraction Click model (MDAC model). This model describes customers' both browsing and purchase behavior based on a Markov chain.

In practice, due to stockouts and other exogenous factors, not all products are available for clicking. Therefore, before introducing our MDAC model, we clarify product availability in Definition 4.2.1. In this chapter, to simplify the notation, we assume that the set of clickable products is the same as the set of purchasable products. For instance, products that are out of stock and thus not available for purchase are not displayed on the e-commerce platform. If certain products are clickable but not purchasable, we can introduce a separate set for purchasable products and our analysis would remain valid.

Definition 4.2.1 (Availability of products)**.** *If a product is clickable and purchasable, then we say that this product is available.*

The availability of products may vary among customers, resulting in distinct browsing behaviors and purchase decisions. Given one customer, we assume that the availability of products remains the same until she leaves the website or purchases one product.

Our Markovian Dynamic Attraction Click model is based on a Markov chain, where the states of the Markov chain represent different products and the no-purchase alternative. The product set contains n products, which is denoted by $[n] = \{1, 2, ..., n\}$. We use index 0 to denote the no-purchase option of the customer and $[\bar{n}]$ to denote the set $[n] \cup \{0\}$. We use $S \subseteq [n]$ to denote the set of available products. We denote the number of available products by $|S|$. For simplicity, we define $\bar{S} := S \cup \{0\}$ and $|\bar{S}| := |S| + 1$.

The browsing and purchase behaviors of customers in MDAC are as follows. An arriving customer initially clicks on product i with arrival probability λ_i for $i \in [n]$, where $\sum_{i\in[n]} \lambda_i = 1$. After a customer clicks on product i, she has three options: Purchase this product immediately, click on other products, or leave the platform without any purchase. The probability of purchasing this product immediately is defined as the instant purchase probability $\mu(i; \mathbf{p})$, where $\mathbf{p} = \{p_i\}_{i=1}^n$ is a vector of prices of all products. The instant purchase probability

$\mu(i;\mathbf{p})$ is between $[0,1]$. Since customers can only purchase the available products, we have $\mu(i;\mathbf{p}) = 0$ for all $i \in [n]\backslash S$.

If this customer does not purchase product i immediately, she transitions to the state of other available products or the no-purchase state. Since the click behaviors of customers depend on the availability of the products, to characterize the probability of click transitions, given the available product set S, we define a click transition matrix $\Theta^S \in \mathbb{R}^{(|\bar{S}|)\times(|\bar{S}|)}$. Each entry Θ^S_{ij} where $i,j \in \bar{S}$, denotes the probability that a customer clicks on product j right after product i, conditional on the customer not purchasing product i. Thus, the probability of clicking on product j right after clicking on product i is $(1-\mu(i;\mathbf{p}))\Theta^S_{ij}$. Particularly, $(1-\mu(i;\mathbf{p}))\Theta^S_{i0}$ denotes the probability that the customer leaves the system after clicking product i without any purchase. The summation of transition probabilities equal 1; that is, $\sum_{j\in\bar{S}} \Theta^S_{ij} = 1$, $\forall i \in S$.

Given the available product set S, the click transition matrix Θ^S is specified as follows:

$$\Theta^S_{ij} = \frac{\rho_{ij}}{\sum_{k\in\bar{S}} \rho_{ik}}, \quad \forall i,j \in \bar{S}. \tag{4.1}$$

Here, ρ_{ij} is called the *relative attraction value* of state j with respect to state i, for any i,$j \in [\bar{n}]$. Intuitively, ρ_{ij} characterizes the attractiveness of state j after clicking on product i. When ρ_{ij} gets larger, the probability of transitioning to state j from state i gets larger. The click transition probability Θ^S_{ij} in (4.1) can be viewed as the attraction choice model (Luce, 1959) given the attractiveness ρ_{ij}. The MNL choice model is a special case of the attraction choice model.

The relative attraction value ρ_{ij} is a universal parameter that does not depend on product availability or prices. Since click transition probability in (4.1) is scale-independent of ρ_{ij}, without loss of generality, we assume that $\sum_{j\in[\bar{n}]} \rho_{ij} = 1$. We define the attraction matrix $\boldsymbol{\rho} \in \mathbb{R}^{(n+1)\times(n+1)}$, where each entry is the value of ρ_{ij}. Since the no-purchase state is an absorbing state, we have that $\rho_{00} = 1$ and $\rho_{0j} = 0$ for $j \neq 0$.

Overall, in the MDAC model, there are three sets of parameters: arrival probability $\boldsymbol{\lambda}$, attraction matrix $\boldsymbol{\rho}$, and the instant purchase function $\mu(i;\mathbf{p})$. In MDAC, customers continue to transition until they purchase the product in the current state or transition to a no-purchase state. Thus, all these three sets of parameters jointly impact the purchase behavior and transition path of one customer.

Our MDAC model is capable of characterizing three important properties for browsing and purchase behaviors of online customers: *relevance between products*, *limited products availability*, and *joint pricing effect*. These three properties are elaborated as follows.

Relevance between products. In MDAC model, the click transition probability Θ^S_{ij} depends on the current browsing product i. This dependence is the main feature that separates our click models from other click models, including the cascade click model or the click chain model. In practice, the next product to be clicked on often depends on what product is currently being browsed. In the example of a used car website, if a customer clicks

on a very good deal for a specific car brand, she is more likely to click on cars of the same brand next. Conversely, if she clicks on a bad deal for a brand, she may explore other brands or leave the website without making any purchase. Thus, the click transition probability is dependent on the current browsing product. Previous literature, such as Besbes, Gur, and Zeevi, 2016; Amaldoss and He, 2018, has modeled this dynamic dependence in various ways. In our MDAC model, then ρ_{ij} is close to 1, product j is very likely to be clicked next after product i is clicked. It is worth noting that ρ_{ij} does not directly capture the similarity between products i and j. For example, if products i and j are very similar but product i is a bad deal, then this may discourage customers from clicking on product j, resulting in a smaller ρ_{ij}. Generally speaking, if products i and j are very similar, then the ith and jth rows of attraction matrix $\boldsymbol{\rho}$ will be very similar. Moreover, since the similarities between products involve multiple attributes, such as brand, model, mileage, and production year, the click transition probability between products is a joint effect of these attributes. Particularly, the rank of the attraction matrix $\boldsymbol{\rho}$ further describes the number of potential attributes (latent factors) that influence the click transition probability, which will be discussed in detail in Section 4.3.1.

Limited products availability. The click transition probability (4.1) indicates that the click probability Θ_{ij}^S is proportional to the relative attraction value ρ_{ij}. This relation is consistent with observations in practice: If more products become available, customers' attention to each product will be smaller, and the probability of clicking on each product (including the no-purchase alternative) will decrease. If there are fewer available products, then customers are more likely to leave without purchase. Since the product availability for different customers may vary, the click transition probability and purchase probability also vary among products. This brings additional challenges in estimating the parameters in MDAC from dynamic availability. Our MDAC enables a fast algorithm to recover $\boldsymbol{\rho}$ under dynamic availabilities, which will be discussed in detail in Section 4.3.6. Besides, the dynamic availability implies that the optimal pricing strategy also varies among customers. Our MDAC model enables retailers to adapt the optimal pricing strategy to various availability, which will be further discussed in Section 4.4.

Joint pricing effect. In the MDAC model, for the customer who is currently browsing product i, she can purchase product i in two ways: She can either purchase this product immediately, or she can transition back to product i and purchase it after some future transitions to other states. Thus, the final purchase probability of product i depends on the instant purchase function μ of all products, not only the price of product i. This purchase probability is mathematically related to the general Markov chain-based choice model (GMCCM) in Dong, Simsek, and Topaloglu, 2019; Kleywegt and Shao, 2022; Goutam, Goyal, and Soret, 2019, when all products in GMCCM are within the assortment. In this case, the purchase probability of MDAC is the same as the choice probability of GMCCM models, which means that MDAC is able to describe the joint pricing effect on the purchase behavior.

However, these two models are conceptually different and describe distinct scenarios. The GMCCM focuses on the final purchase probability of products. Their transition probability is fixed and independent of the assortment. This independence limits the capability of GMCCM to characterize click transition behaviors under limited and dynamic availability. On the other hand, our MDAC focuses on the role of prices and is effectively designed to capture these click transition behaviors.

Our MDAC model integrates the purchase and click behaviors of customers. It is worth noting that in our click model, we do not consider the effect of ranking, recommendation, or limited display on the click behaviors. By (4.1), the click transition probability only depends on the set of available products and the attraction matrix $\boldsymbol{\rho}$. Our MDAC is also supported empirically by the real-world dataset. As shown later in Figure 4.2 in Section 4.5.3, given the same estimated parameters of $\boldsymbol{\rho}$, λ, and μ, when utilizing the state information of one customer revealed by her current clicked product, we can reduce the prediction error for her final purchase behavior. This shows the benefit of using click data to predict the purchase probability via real-world data.

4.3 Estimation of MDAC using Click Data

In this section, we propose algorithms to efficiently estimate the parameters in the MDAC using both the click data and the purchase data. These parameters include the attraction matrix $\boldsymbol{\rho}$, the instant purchase probability $\mu(i;\mathbf{p})$, and the arrival probability λ. We first discuss the formulation and algorithm for estimating the attraction matrix $\boldsymbol{\rho}$ under the MDAC model. Next, we propose algorithms to estimate the instant purchase probability $\mu(i;\mathbf{p})$. We further analyze the generalization error bound for the estimation of the low-rank attraction matrix. Finally, we provide an efficient algorithm to estimate the attraction matrix under the dynamic availability of products.

Given the available product set S, the click transition probability Θ^S can be estimated through the click transition data collected under this available product set S. Specifically, if a customer clicks product j immediately after clicking product i, without purchasing product i or leaving the platform, the system records a valid click transition pair as (i,j) or $(i,0)$. In Section 4.5, we provide more details about how to identify the valid click transition pairs from the click data. The set of valid click transition pairs under available product set S is denoted by $\mathbb{C}_S = \{(i_1^S, j_1^S), (i_2^S, j_2^S), ..., (i_{N_S}^S, j_{N_S}^S)\}$, where N_S is the total number of valid click transition pairs under available product set S. Thus, a simple estimation of Θ_{ij}^S is the proportion of click transition pairs from i to j within $\mathbb{C}_S$. However, this simple estimation would require a large number of click transition data to obtain a small estimation error, especially when the number of products is large. Thus, to address the scalability issue, we explore the similarities among products in the next section.

4.3.1 Similarities Among Products

When the number of products n is large, estimating the attraction matrix $\boldsymbol{\rho}$ accurately is challenging because there are $n \times n$ entries within the matrix. Even if the estimation error for one entry is small, the cumulative estimation error for the entire attraction matrix can be very large due to the high dimensionality of products, which may further result in the sub-optimality of the pricing decisions. To overcome this issue, we realize that some products may share some common properties or lie in the same categories. Utilizing the potential similarities between products, we can group the click transition behaviors into several patterns. The number of patterns can be much smaller than the total number of products. Thus, by utilizing similarities between products, we can reduce the size of the search space when estimating the attraction matrix $\boldsymbol{\rho}$. It can accelerate the estimation process and yield a smaller estimation error.

To characterize similarities between products, we assume the attraction matrix $\boldsymbol{\rho}$ to be low-rank. Specifically, the matrix's rank is at most r, potentially much smaller than n. This low-rank structure suggests the existence of r latent factor transition patterns influencing the click transition probability across the product set. The reason is as explained below. The matrix $\boldsymbol{\rho}$ with rank r can be decomposed as $\boldsymbol{\rho} = \sum_{t=1}^{r} u_t v_t^T$, where $u_t, v_t \in \mathbb{R}^{n+1}$ are some vectors. These vectors u_t and v_t can be interpreted as follows. Take a used car trading website as an example, factors influencing click transition probability include car attributes like mileage, production year, and other latent factors. If there are r such factors, the attraction matrix's rank is r. For each factor t, the ith entry in vector u_t represents the impact of product i on the factor t. Intuitively, a superior product with a high value in factor t will likely encourage customers to click on other products also scoring high in that factor, hence positively impacting factor t and making u_t^i positive. For example, if factor t represents mileage, and product i is a good car with high mileage, then product i has a positive impact on the mileage attribute. In this case, u_t^i is a positive number. Conversely, if product i is a low-quality car with a high mileage, then product i has a negative impact on the mileage attribute, resulting in a negative u_t^i. Next, the jth entry in vector v_t represents the impact on the click transition probability to product j from factor t. Intuitively, if product j has a high value in factor t, then factor t has a large impact on its click probability. For example, if product j has a higher mileage, then the attribute t has a larger impact on product j, resulting in a larger v_t^j. Thus, by summing over all possible attributes, the click transition probability in the matrix $\sum_{t=1}^{r} u_t v_t^T$ for each product can be described as a linear combination of effect from these r potential factors. This low-rank structure significantly mitigates the complexity of estimating the attraction matrix. We verified the low-rank structure of the attraction matrix $\boldsymbol{\rho}$ using a real-world dataset in Section 4.5.2.

When the rank of the attraction matrix $\boldsymbol{\rho}$ is one, there is only one transition pattern among products and our MDAC will reduce to the traditional MNL choice model. In this case, for each click transition, customers will simply resample one product (or no-purchase option) from the set of available products. When the rank is larger than one, our click model is able to characterize heterogeneous transition patterns that depend on the current browsing

product.

Before introducing estimation algorithms, we first provide some additional notations for the click data. Recall that $\mathbb{C}_S$ is the set of click transitions under available product set S. We use X_{ij} to denote the indicator matrix whose entry in the i^{th} row and the j^{th} column is one while the other entries are all zero. We use $A \cdot B$ to denote the sum of the entrywise multiplication of two matrices A and B.

4.3.2 Estimating the Attraction Matrix with Low-Rank Structures

To estimate the attraction matrix $\boldsymbol{\rho}$, we maximize the log-likelihood function of the click transitions. To address the low-rank structure of the click attraction matrix $\boldsymbol{\rho}$, we add a term $\gamma\|\boldsymbol{\rho}\|_*$ to the objective function, where $\|\cdot\|_*$ is the nuclear norm of the matrix and γ is a positive multiplier for the regularization term. The value of γ will be specified later. In particular, the formulation for the estimation problem is as follows:

$$\min_{\boldsymbol{\rho}} \quad \mathcal{L}(\boldsymbol{\rho}) := \ \ell_{mle}(\boldsymbol{\rho}) + \gamma\|\boldsymbol{\rho}\|_* \tag{4.2}$$

$$s.t. \quad \sum_{j\in[\bar{n}]} \rho_{ij} = 1, \quad \forall i \in [n] \tag{4.2a}$$

$$\rho_{ij} \geq 0, \quad \forall i \in [n],\ \forall j \in [\bar{n}] \tag{4.2b}$$

$$\rho_{00} = 1, \rho_{0j} = 0, \quad \forall j \neq 0, j \in [\bar{n}]. \tag{4.2c}$$

Constraints (4.2a) and (4.2b) require each row of $\boldsymbol{\rho}$ to be a probability simplex, where constraints (4.2a) require the sum of each row of $\boldsymbol{\rho}$ to be one. Constraint (4.2c) restricts the no-purchase alternative to be an absorbing state. In objective function (4.2), $\ell_{mle}(\boldsymbol{\rho})$ is the negative log-likelihood function of the click transition data. In the next paragraphs, exact expressions of $\ell_{mle}(\boldsymbol{\rho})$ are provided.

Suppose that the clickstream data are collected from various sets of available products. Under (4.1), for one available product set S, the negative log-likelihood function $\ell_{mle}(\boldsymbol{\rho}; S)$ can be written as:

$$\ell_{mle}(\boldsymbol{\rho}; S) = \frac{1}{N_S} \sum_{t\in\mathbb{C}_S} \left[\ln(\sum_{k\in S} X_{i_t k} \cdot \boldsymbol{\rho}) - \ln(X_{i_t j_t} \cdot \boldsymbol{\rho}) \right].$$

Let $\mathcal{S}_t$ denote the set of available product sets shown by time t. Then, $\ell_{mle}(\boldsymbol{\rho}) = \frac{1}{|\mathcal{S}_t|} \sum_{S\in\mathcal{S}_t} \ell_{mle}(\boldsymbol{\rho}; S)$. The objective $\mathcal{L}(\boldsymbol{\rho})$ is

$$\mathcal{L}(\boldsymbol{\rho}) = \frac{1}{|\mathcal{S}_t|} \sum_{S\in\mathcal{S}_t} \ell_{mle}(\boldsymbol{\rho}; S) + \gamma\|\boldsymbol{\rho}\|_*. \tag{4.3}$$

Given one available product set S, the negative log-likelihood function of click transition matrix Θ^S can be reduced to $\ell_{mle}(\Theta^S) = \frac{1}{N_S} \sum_{t\in\mathbb{C}_S} \left[\ln\left(\sum_{j\in S} X_{i_t j} \cdot \Theta^S \right) - \ln(X_{i_t j_t} \cdot \Theta^S) \right]$. In

this case, the objective function (4.3) regarding to the available product set S is

$$\mathcal{L}(\Theta^S) = \ell_{mle}(\Theta^S) + \gamma\|\Theta^S\|_* = \frac{1}{N_S}\sum_{t\in\mathbb{C}_S}\left[\ln\Big(\sum_{j\in S} X_{i_t j}\cdot\Theta^S\Big) - \ln(X_{i_t j_t}\cdot\Theta^S)\right] + \gamma\|\Theta^S\|_*. \tag{4.4}$$

Solving Problem (4.2) is non-trivial because although the nuclear norm is convex, $\ell_{mle}(\boldsymbol{\rho})$ is not convex; thus, the objective function $\mathcal{L}(\boldsymbol{\rho})$ is not convex. However, $\mathcal{L}(\boldsymbol{\rho})$ is restricted convex (see Appendix C.1.1), which implies that we can use the subgradient projection method to solve (4.2). This algorithm is provided in Appendix 4.3.2.1, which is similar to the gradient projection method in Fu and Ge, 2021. At each iteration, we used subgradient descent to update $\boldsymbol{\rho}$ and then project $\boldsymbol{\rho}$ to the feasible space. As will be shown later in the proof of Lemma 4.3.3, $\mathcal{L}(\boldsymbol{\rho})$ in Equation (4.3) is restricted convex in the feasible region, so Algorithm 4 converges to the global optimal points, e.g., see Gafni, Bertsekas, et al., 1982 and Calamai and Moré, 1987.

4.3.2.1 Subgradient Projection Algorithm for Estimating the Transition Matrix

In this section, we provide the algorithm for solving (4.2) when estimating the transition matrix. This subgradient projection algorithm is provided in Algorithm 4. The step size in Algorithm 4 can be chosen according to the generalized version of the Armijo rule in Bertsekas, 1976. The projection step is equivalent to projecting each row of $\boldsymbol{\rho}$ to the probability simplex, and the projection oracle can follow the procedures in W. Wang and Carreira-Perpinán, 2013.

Algorithm 4 Subgradient projection method for estimating the transition matrix

1: **Input:** Sequences of click data $\{X_{i_t j_t}\}_{t\in\mathbb{C}_S}$. An oracle Proj($\boldsymbol{\rho}$) to project $\boldsymbol{\rho}$ to the feasible region satisfying (4.2a), (4.2b) and (4.2c). Sequences of the step sizes a_m, the maximum iteration number M_{iter}, and the tolerance level v_{tol}
2: **Initialization:** Set $\boldsymbol{\rho}^{(0)}$ as $\rho_{ij} = \frac{1}{\bar{n}}$, $\forall i \in [n], j \in [\bar{n}]$. Set the iteration number m_{iter} as 0
3: **While** ($m_{iter} \le M_{iter}$):
4: Calculate the subgradient of objective function (4.2) at point $\boldsymbol{\rho}^{m_{iter}}$ as $\nabla\mathcal{L}_{m_{iter}}$
5: Let $\boldsymbol{\rho}^{m_{iter}+1} \leftarrow \text{Proj}(\boldsymbol{\rho}^{m_{iter}} - a_{m_{iter}}\nabla\mathcal{L}_{m_{iter}})$
6: **If** the improvement of the objective function (4.2) is less than v_{tol}:
7: **Break**
8: **Else:**
9: $m_{iter} \leftarrow m_{iter} + 1$
10: **Return** $\boldsymbol{\rho}^{m_{iter}}$

The subgradient of the objective (4.2), $\nabla\mathcal{L}(\boldsymbol{\rho})$ can be calculated in Lemma 4.3.1. We denote the subgradient of the nuclear norm $\|\boldsymbol{\rho}\|_*$ by $\nabla\|\boldsymbol{\rho}\|_*$.

Lemma 4.3.1. *For the MDAC model, $\mathcal{L}(\boldsymbol{\rho})$ is restricted convex in the feasible space. The subgradient of $\mathcal{L}(\boldsymbol{\rho})$ can be computed by:*

$$\nabla\mathcal{L}(\boldsymbol{\rho}) = \frac{1}{|\mathcal{S}_t|}\sum_{S\in\mathcal{S}_t}\frac{1}{N}\sum_{t=1}^{N}\Big[\frac{\sum_{j\in S}X_{i_tj}}{\sum_{j\in S}X_{i_tj}\cdot\boldsymbol{\rho}} - \frac{X_{i_tj_t}}{X_{i_tj_t}\cdot\boldsymbol{\rho}}\Big] + \gamma\nabla\|\boldsymbol{\rho}\|_*.$$

4.3.3 Learning Purchase Behaviors

We next estimate the instant purchase probability $\mu(i;\mathbf{p})$ using the purchase data, given the estimation of the attraction matrix $\boldsymbol{\rho}$. In our study, we assume the instant purchase probability $\mu(i;\mathbf{p})$ in the MDAC has the following form:

$$\mu(i;\mathbf{p}) = e^{-\alpha_i p_i} \text{ for all } i \in S,$$

where α_i is the price elasticity of the instant purchase probability for product i and where p_i is the price of product i. Note that $\mu(i;\mathbf{p})$ does not depend on the prices of other products, which is a commonly adopted assumption in Dong, Simsek, and Topaloglu, 2019 and Kleywegt and Shao, 2022. To estimate the function $\mu(i;\mathbf{p})$, it suffices to estimate the price elasticity vector $\boldsymbol{\alpha}\in\mathbb{R}^n$.

Suppose that N_i^B is the total number of click transition pairs starting with product i when product i is available. Let W_{ii} be the number of times that customers purchase product i right after clicking product i within these N_i^B transitions. Then, the log-likelihood function of price elasticity α_i can be written as

$$\ell_{\text{purchase}}(\alpha_i) = -\alpha_i p_i W_{ii} + (N_i^B - W_{ii})\ln(1-e^{-\alpha_i p_i}). \tag{4.5}$$

We can then estimate α_i by maximizing the negative log-likelihood function (i.e., $\hat{\alpha}_i = \arg\max_{\alpha_i}\ell_{\text{purchase}}(\alpha_i)$). Lemma 4.3.2 shows that $\ell_{\text{purchase}}(\alpha_i)$ is concave with respect to α_i.

Lemma 4.3.2. *The negative log-likelihood function $\ell_{purchase}(\alpha_i)$ is concave with respect to α_i.*

4.3.4 Estimation of Arrival Probability

In this section, we discuss the methods to estimate arrival probability λ_i. We use the purchase data to estimate the arrival probability.

The purchase probability of product i for one customer, given the available product set S and arrival rate λ can be written as $\pi(\lambda;i,S)$ in (4.6) by Goutam, Goyal, and Soret, 2019 as

$$\pi(\lambda;i,S) = \lambda^T(I_n - Diag(1-\mu(i;\boldsymbol{p}))\boldsymbol{\rho})^{-1}\Pi(S)e_i, \tag{4.6}$$

where

$$\Pi(S) = \begin{bmatrix} \mu(1;S,p_1) & 0 & ... & 0 & (1-\mu(1;S,p_1))\rho_{10} \\ 0 & \mu(2;S,p_2) & ... & 0 & (1-\mu(2;S,p_2))\rho_{20} \\ ... & & & ... & ... \\ 0 & 0 & ... & \mu(n;S,p_n) & (1-\mu(n;S,p_n))\rho_{n0} \end{bmatrix}.$$

Suppose the total customer viewing available product set S is K^S and within these customers, the number of customers who purchase product i is w_i^S, then the log-likelihood function is:

$$\mathcal{L}_{\text{arrival}}(\lambda) = \sum_{S\in\mathcal{S}}\sum_{i\in S} w_i^S \ln(\pi(\lambda; i, S)) + \sum_{S\in\mathcal{S}}\sum_{i\in S}(K^S - w_i^S)\ln(1-\pi(\lambda; i, S)). \tag{4.7}$$

Then, we estimate $\pi(\lambda; i, S)$ using

$$\hat{\pi}(\lambda; i, S) = \arg\max_{\pi(\lambda;i,S)} \mathcal{L}_{\text{arrival}}(\pi(\lambda; i, S)),$$

which is a concave function of $\pi(\lambda; i, S)$. After having $\hat{\pi}(\lambda; i, S)$, we can solve linear equations (4.6) to get the estimation $\hat{\lambda}$.

As will be shown in Section 4.4.1, the estimation of arrival probability λ does not influence the optimal pricing decisions, so it does not change the revenue or the regret analysis of online pricing.

4.3.5 Estimation Error Bound with Low-rank Structure

In this section, given the finite size of the training set, we quantify the estimation error bounds for the parameters in MDAC. These estimation error bounds will prepare us to justify our greedy pricing policy in Section 4.4. For the estimation error of the instant price elasticity α_i, the estimation error, $|\hat{\alpha}_i - \alpha_i|$ can be shown to converge to zero at rate $\mathcal{O}(1/\sqrt{N_i^B})$ (see, e.g., Theorem 1 in L. Li, Yu Lu, and Zhou, 2017), when N_i^B click transition pairs are collected.

Now, we focus on the estimation error bound for the obtained transition matrix $\hat{\Theta}^S$ by solving (4.2). Most literature on low-rank matrix recovery considers cases where the noise of each entry is independent. In Problem (4.2), however, since customers' click behavior is a function of all the entries in one row, when maximizing the likelihood function of customers' click data, the noise of the entries in one row are correlated. Therefore, the sample complexity from the cases where the noise is independent is not applicable to our setting.

We use $\|\cdot\|_F$ to denote the Frobenius norm of a matrix. We first focus on the availability-constrained click model under a given available product set. Recall that the objective function is given in (4.4), and Θ^S in (4.4) is a submatrix of $\boldsymbol{\rho}$. Assumption 4.3.1 assumes the boundedness of each entry in Θ^S. The lower bound of Θ^S in Assumption 4.3.1 ensures that there exists a positive minimum probability for transitioning between any two products. The upper bound $\Theta_{ij}^S \le \frac{\beta_2}{|S|}$ basically assumes that the transition probability is upper bounded by $\mathcal{O}(1/|S|)$, which is reasonable since customers' attention to each product may decrease when more products become available. The boundedness of transition probability is a common assumption in the low-rank estimation literature with state transitions (e.g., Theorem 3 in Kallus and Udell, 2020, and Assumption 1 in Z. Zhu et al., 2021.)

Assumption 4.3.1. *There exist constants β_0 and β_2, such that $0 < \beta_0 \le \beta_2 \le |S|$ and $\beta_0 \le \Theta_{ij}^S \le \frac{\beta_2}{|S|}$, for all $i \in S$ and $j \in \bar{S}$.*

For simplicity, we denote $\min\{\beta_0, \frac{1}{\beta_2}\}$ by β_1, and thus, we have $\beta_1 \leq \Theta_{ij}^S \leq \frac{1}{\beta_1|S|}$, for all $i, j \in \bar{S}$.

Given any available product set S, Lemma 4.3.3 provides the finite-sample error bound of the estimator $\hat{\Theta}^S$ for the MDAC model, where the true value of Θ^S is denoted by Θ^{*S}.

Lemma 4.3.3 (Non-asymptotic error bound with fixed availability)**.** *Suppose that Assumption 4.3.1 holds. Given one set of available products S, suppose that there exists a constant c_1 such that the number of clicks under this availability satisfies $N_S \leq \frac{2c_1|S|^2}{9\beta_2^2}$. Then, for any integer $r \leq |S|$ and parameter $\tau \geq 1$, setting $\gamma = \frac{1}{2}\sqrt{\frac{8\tau \ln(2|S|)}{N_S\beta_1}}$, with probability at least $1 - 4(2|S|)^{-\tau/c_1}$, any solution $\|\hat{\Theta}^S\|$to Problem (4.2) satisfies*

$$\|\hat{\Theta}^S - \Theta^{*S}\|_F \leq \frac{128}{\beta_1^2}\sqrt{\frac{2\tau\tilde{r}\ln(2|S|)}{N_S\beta_1}}, \tag{4.8}$$

where $\tilde{r} = \max\{rank(\Theta^{*S}), r\}$.

When the rank of Θ^{*S} is r, Lemma 4.3.3 demonstrates that the estimation error in terms of the Frobenius norm grows in the order of $\tilde{\mathcal{O}}(\sqrt{r\ln(|S|)})$, and converges to zero at rate $\tilde{\mathcal{O}}(\sqrt{\frac{1}{N_S}})$. Note that this result is smaller than the order $\tilde{\mathcal{O}}(\sqrt{r|S|})$ in Kallus and Udell, 2020 and is in the same order as Theorem 1 in Z. Zhu et al., 2021, although our assumptions are different. In Lemma 4.3.3, we assume $N_S \leq \mathcal{O}(|S|^2)$, which is a technical requirement of the proof, (same as Theorem 3 in Kallus and Udell, 2020, Remark 3 in Z. Zhu et al., 2021). When N_S is larger than $\mathcal{O}(|S|^2)$, the platform has sufficient click data to estimate the transition matrix, and does not need to consider the low-rank structure of it. In Lemma 4.3.3, τ is a parameter to control the probability and estimation error. In the setting of online pricing in Section 4.4.2, we will show how to set the value of τ.

Lemma 4.3.3 provides an error bound for the estimation for a given available product set S. When the availability of products is dynamically changing, we can still solve Problem (4.2) to obtain the estimation of the attraction matrix. However, since the click transition behaviors are different under various availabilities, the derivation of the estimation error bound is intricate. Thus, in Section 4.3.6, we provide another algorithm that allows us to derive the error bound under various availabilities, by utilizing the results of Lemma 4.3.3.

4.3.6 Estimation Under Dynamic Availability

In reality, the range of available products is often limited due to exogenous factors such as stockouts or discontinuation. In response to these uncontrollable external factors, online retailers strive to adjust to any potential fluctuations in product availability. This adaptation necessitates understanding the entire attraction matrix, making it crucial for retailers to estimate this matrix based on the limited and dynamically changing pool of available products.

In this section, we provide the algorithm as well as the error bound for estimating the entire attraction matrix $\boldsymbol{\rho}$ under various available product sets.

The estimation algorithm for the entire attraction matrix $\boldsymbol{\rho}$ under various availabilities is stated in Algorithm 5. The basic idea is that we first estimate the transition matrix under each single available product set to obtain the estimation of submatrix Θ_{ij}^S. Next, we re-scale each submatrix to combine the results from different availabilities and get the estimation of the full attraction matrix $\boldsymbol{\rho}$. In this case, the error bound of the estimation of the entire attraction matrix $\boldsymbol{\rho}$ can be obtained by combining the error bounds of the estimated submatrix Θ^S of each available product set. To recover the entire attraction matrix, the variety of product availability should be sufficient to reveal the transition behaviors between any pair of products. To describe this condition, we define the *cover* of all products in Definition 4.3.1.

Definition 4.3.1 (Cover of all products.)**.** *We say that a set of available product sets $\mathbb{S}_c$ is a cover of all products, if for any pair of products (i, j), $i, j \in [n]$, there exists some available product set $S \in \mathbb{S}_c$ such that both products are available, i.e., $i \in S$ and $j \in S$.*

Proposition 4.3.1 further shows that the cover of all products is a necessary condition to estimate the entire attraction matrix under various availabilities.

Proposition 4.3.1 (Necessary condition for the estimation under various availabilites.)**.** *Suppose that the click data are collected under a set of available product sets $\mathbb{S}_c$. Under the MDAC model, a necessary condition for the existence of the unique estimator of entire attraction matrix $\boldsymbol{\rho}$ by solving Problem (4.2) with $\gamma = 0$ is that the set of available product sets $\mathbb{S}_c$ is a cover of all products.*

Proposition 4.3.1 shows that when $\gamma = 0$, there are multiple estimators for Problem (4.2) if $\mathbb{S}_c$ is not a cover of all products. In this case, when $\gamma > 0$, the estimator with the smallest nuclear norm would minimize objective (4.2), which may not be consistent with the true attraction matrix $\boldsymbol{\rho}$. Thus, Proposition 4.3.1 implies that although each available product set does not have to include all products, the set of available product sets has to be a cover of all products, in order to estimate the entire attraction matrix. Therefore, in Algorithm 5, to recover the entire attraction matrix, we assume the input is a cover of all products. Because the numbers of click transition data under different availabilities are varied, the estimation error bounds for each submatrix are also different. In order to get an estimation of $\boldsymbol{\rho}$ with a small estimation error, we keep removing the estimation result with the largest estimation error bound until the remaining available product sets are no longer a cover of all products. This allows us to identify a cover whose largest error bound is minimized. Then, we can integrate the estimation of the remaining submatrix and get the estimation of $\boldsymbol{\rho}$. Let $\mathcal{S}$ denote the set of all possible product available sets S, and the cardinality of $\mathcal{S}$ is denoted by $|\mathcal{S}|$.

Theorem 4.3.1 shows the error bound for Algorithm 5 when estimating the attraction matrix in the MDAC model. We use err_S to denote the estimation error for Θ^S using Lemma 4.3.3, given the rank as r.

Algorithm 5 Estimate the attraction matrix under dynamic availability

Input: A set of available product sets $\mathcal{S}$ which is a cover of all products; the click transition data under each available product set S, for all $S \in \mathcal{S}$; The rank of the attraction matrix r.
For each available product set $S \in \mathcal{S}$:
  Estimate Θ^S by solving Problem (4.2).
$\mathbb{S}_c \leftarrow \mathcal{S}$.
Sort the available product sets in nonincreasing order according to the value of $\ln(|S|)/N_S$. After sorting, denote the sequence of available product sets by $S_{(1)}, S_{(2)}, ..., S_{(|\mathcal{S}|)}$.
For $i = 1, 2, ..., |\mathcal{S}|$:
  If $\mathbb{S}_c \backslash \{S_{(i)}\}$ is still a cover of all products:
    Remove $S_{(i)}$ from Set $\mathbb{S}_c$.
  Else:
    Break.
Solve the following equations to get the estimation of attraction matrix $\boldsymbol{\rho}$:

$$\begin{cases} \frac{\rho_{ij}}{\rho_{i0}} & = \frac{\Theta^S_{ij}}{\Theta^S_{i0}}, \quad \exists S \in \mathbb{S}_c, \forall i, j \in [n], \\ \sum_{j \in [n]} \rho_{ij} & = 1, \quad \forall i \in [n]. \end{cases} \tag{4.9}$$

Return matrix $\boldsymbol{\rho}$ and the cover $\mathbb{S}_c$.

Theorem 4.3.1. *Let $\hat{\boldsymbol{\rho}}$ and $\mathbb{S}_c$ be the outputs of Algorithm 5. Under the same setting of Lemma 4.3.3, with probability at least $1 - 4\sum_{S \in \mathbb{S}_c}(2|S|)^{-\tau/c_1}$, the estimated attraction matrix $\hat{\boldsymbol{\rho}}$ satisfies $\|\hat{\boldsymbol{\rho}} - \boldsymbol{\rho}^*\|_F \leq \sqrt{\sum_{S \in \mathbb{S}_c} err_S^2}$.*

Theorem 4.3.1 indicates that the estimation error of $\boldsymbol{\rho}$ can be upper bounded by the ℓ_2-norm of the vector for the estimation error bounds err_S within the final cover $\mathbb{S}_c$. The largest size of output $\mathbb{S}_c$ is at most n^2, so combining with Lemma 4.3.3, we obtain that the estimation error is no more than $\tilde{\mathcal{O}}(n\sqrt{\frac{r}{N_S^{\min}}})$, where $N_S^{\min}$ is the smallest number of clicks for one available product set in $\mathbb{S}_c$. In reality, to constitute a cover, the size of output $\mathbb{S}_c$ does not need to be as large as n^2. For example, if more than half of all products are available each time, the size of $\mathbb{S}_c$ can be as small as n, so the estimation error bound in Theorem 4.3.1 can be in the order of $\tilde{\mathcal{O}}(\sqrt{\frac{nr}{N_S^{\min}}})$. These small error bounds prepare us to demonstrate the effectiveness of the greedy pricing policy under dynamic availability of products.

4.4 Pricing from the Click Data

In this section, we study how to infer the optimal pricing strategies from the click data. First, in Section 4.4.1, we reveal some insights on the change of optimal prices when click behaviors

change. Second, in Section 4.4.2, we propose an online algorithm to learn parameters in the MDAC while simultaneously maximizing the revenue from pricing decisions. Finally, in Section 4.4.3, we analyze the regret of our exploration-free online pricing algorithm.

4.4.1 Optimal Static Prices

Given the available product set S and parameters $(\boldsymbol{\rho}, \boldsymbol{\alpha})$, the optimal prices depend on the instant purchase probability $\mu(i; \mathbf{p})$. Recall in Section 4.3.3, we assume that the instant purchase probability $\mu(i; \mathbf{p}) = e^{-\alpha_i p_i}$ for $i \in S$. For simplicity of expression and with a slight abuse of notation, we use $\mu(p_i) = \mu(i; \mathbf{p})$ to denote the instant purchase probability. Let $\mathfrak{R}(\mathbf{p})$ denote the expected revenue from one customer when the price vector of the available products is $\mathbf{p} \in \mathbb{R}^{|S|}$. Suppose the cost of product i is c_i, and the corresponding cost vector is $\mathbf{c}$. The expected revenue conditional on the current state i, denoted as r_i, can be decomposed into two parts: the instant revenue and future revenue. The instant revenue is the expected revenue when customers purchase product i, if product i is available. Since the purchase probability is $\mu(p_i)$, this part can be written as $\mu(p_i)(p_i - c_i)$. The future revenue refers to the expected revenue when customers transition to other states, which can be written as $\big(1 - \mu(p_i)\big) \sum_{j \in [n]} \rho_{ij} r_j$. Therefore, by summing these two components, the expected revenue can be expressed as

$$r_i = \mu(p_i)(p_i - c_i) + \big(1 - \mu(p_i)\big) \sum_{j \in [n]} \rho_{ij} r_j. \tag{4.10}$$

Then, the total profit is $\mathfrak{R}(\mathbf{p}) = \sum_{i \in [n]} \lambda_i r_i$, where λ_i is the probability that a customer entering the system visits product i.

To maximize the total profit $\mathfrak{R}(\mathbf{p})$, we can utilize the iterative algorithm provided in Dong, Simsek, and Topaloglu, 2019. They provide an algorithm to compute the optimal price when the available product set is the entire product set. We extend their idea to a more general setting in Algorithm 6. The key idea is that at each iteration, for the available products within S, we set their prices to maximize the expected revenue including both the instant and future expected revenue. We keep iterating until the difference between two consecutive iterations $\|\mathbf{r}^t - \mathbf{r}^{t+1}\|_2$ is less than threshold ϵ_r.

In Algorithm 6, when achieving the optimal prices, the expected revenue of each product will achieve the optimal stationary revenue defined in Definition 4.4.1.

Definition 4.4.1 ((Optimal) Stationary Revenue)**.** *We call $r_i(\mathbf{p}; \boldsymbol{\rho}, \boldsymbol{\alpha})$ the stationary revenue of product i if $\{r_i(\mathbf{p}; \boldsymbol{\rho}, \boldsymbol{\alpha})\}_{i \in [n]}$ satisfies Equation (4.10). Moreover, given $\boldsymbol{\rho}$ and $\boldsymbol{\alpha}$, we call r_i^* the optimal stationary revenue of product i, if r_i^* satisfies:*

$$r_i^* = \begin{cases} \max\limits_{p_i} \left\{ \mu(p_i)(p_i - c_i) + \big(1 - \mu(p_i)\big) \sum\limits_{j \in [n]} \rho_{ij} r_j^* \right\}, & \text{if product } i \text{ is available,} \\ \sum\limits_{j \in [n]} \rho_{ij} r_j^*, & \text{otherwise.} \end{cases}$$

Algorithm 6 Optimal pricing in MDAC

Initialization: Initialize vector $\mathbf{r}^0 \in \mathbb{R}^n$ randomly
While (True):
  For all Product $i \in [n]$:
    If product i **is available**:
      $r_i^{t+1} = \max_{p_i}\{\mu(p_i)(p_i - c_i) + \big(1 - \mu(p_i)\big)\sum_{j\in[n]} \rho_{ij} r_j^t\}$
    Else:
      $r_i^{t+1} = \sum_{j\in[n]} \rho_{ij} r_j^t$
  If $\|\mathbf{r}^t - \mathbf{r}^{t+1}\|_2 \le \epsilon_r$: **Break**
Set $p_i \leftarrow \arg\max \big\{\mu(p_i)(p_i - c_i) + \big(1 - \mu(p_i)\big)\sum_{j\in[n]} \rho_{ij} r_j^t\big\}$, for all $i \in S$
Return p_i, for all $i \in S$

Furthermore, if $r_i^* > r_j^*$, we say that product i is more preferable *than product j.*

In Appendix C.2, we demonstrate that the adapted algorithm in Dong, Simsek, and Topaloglu, 2019 converges to the optimal stationary revenue. Using the concepts of the optimal stationary revenue, we characterize the relationship between optimal prices p_i and attraction matrix $\boldsymbol{\rho}$ in Proposition 4.4.1.

Proposition 4.4.1 (Optimal prices under the change of click transition probability)**.** *Given transition matrices $\boldsymbol{\rho}$ and $\boldsymbol{\rho}'$, suppose p_i and p_i' are the optimal prices of product i under the parameters $(\boldsymbol{\rho}, \boldsymbol{\alpha})$ and $(\boldsymbol{\rho}', \boldsymbol{\alpha})$ respectively. Suppose r_i^* is the current optimal stationary revenue from the customer at product i, under the attraction matrix $\boldsymbol{\rho}$. Then,*

1. *If $\sum_{j\in[n]} \rho_{ij} r_j^* \le \sum_{j\in[n]} \rho_{ij}' r_j^*$, for all product i, then we have $p_i' \ge p_i$, $\forall i \in [n]$;*
2. *If $\sum_{j\in[n]} \rho_{ij} r_j^* \ge \sum_{j\in[n]} \rho_{ij}' r_j^*$, for all product i, then we have $p_i' \le p_i$, $\forall i \in [n]$.*

Proposition 4.4.1 implies that when the attraction matrix changes from $\boldsymbol{\rho}$ to $\boldsymbol{\rho}'$, the change of optimal prices depends on the relations between $\sum_{j\in[n]} \rho_{ij} r_j^*$ and $\sum_{j\in[n]} \rho_{ij}' r_j^*$. The condition $\sum_{j\in[n]} \rho_{ij} r_j^* \le \sum_{j\in[n]} \rho_{ij}' r_j^*$ implies that customers are more likely to transition to the products that offer higher optimal stationary revenue under $\boldsymbol{\rho}$. Proposition 4.4.1 indicates that if this condition holds for all product i, then after the change from $\boldsymbol{\rho}$ to $\boldsymbol{\rho}'$, the optimal prices for all the products are non-decreasing. We note two special cases for Proposition 4.4.1. For small $\epsilon > 0$:

1. First, suppose that $\rho_{ik}' = \rho_{ik} + \epsilon$ and $\rho_{i0}' = \rho_{i0} - \epsilon$, for one pair of product $i, k \in [n]$; and the other elements of $\boldsymbol{\rho}'$ and $\boldsymbol{\rho}$ are the same. Then, by Proposition 4.4.1, we have that the optimal price $p_i' \ge p_i$, for all $i \in [n]$. This implies that if the probability of customers' transitioning to the no-purchase alternative is lower for some products, and the rest of the attraction matrix remains the same, then the prices for all the products are non-decreasing.

2. Second, suppose that $\rho'_{ik} = \rho_{ik} + \epsilon$ for one pair of product $i, k \in [n]$ and $\rho'_{im} = \rho_{im} - \epsilon$ for one pair of product $i, m \in [n]$; the other elements of $\boldsymbol{\rho}'$ and $\boldsymbol{\rho}$ are the same. Suppose that product k is more preferable than product m under the attraction matrix $\boldsymbol{\rho}$. Then, we have the optimal price $p'_i \geq p_i$, for all $i \in [n]$. This observation implies that if the probability of transitioning to some products that have higher revenue increases (or equivalently, if the probability of transitioning to some products that have fewer revenue decreases), and the rest of the attraction matrix remains the same, then the prices for all the products are non-decreasing.

Proposition 4.4.2 demonstrates that under some conditions, the optimal prices are independent of the transition probability ρ_{ij}.

Proposition 4.4.2. *If the attraction matrix $\boldsymbol{\rho}$ and cost vector $\mathbf{c}$ satisfy the following two conditions:*

1. $\rho_{i0} = \rho_{j0}$, $\forall i, j \in [n]$;
2. $c_i = c_j$, $\forall i, j \in [n]$;

Then the optimal prices for all the products are the same.

Proposition 4.4.2 shows that when the transition probabilities to the no-purchase state are the same for all the products and the costs are the same for all products, the optimal prices for all the products are the same, which are independent of the transition probability between the products. It implies that the parameters ρ_{i0} and c_i play an important role in distinguishing the optimal prices between products.

4.4.2 Greedy Online Pricing Policy

In practice, given the constantly increasing amount of click data, the retailer would like to update their pricing strategy promptly to maximize revenue. To achieve this goal, in this section, we consider the online pricing problem where we simultaneously optimize the prices of products and learn the parameters in the MDAC using the click data and purchase data. Our goal is to minimize the cumulative revenue loss compared to the optimal prices, which is defined as regret in (4.11).

Suppose the total number of customers is T and the price vector for customer t is $\mathbf{p}_t$, then the regret of the online pricing problem can be written as:

$$\text{Regret}(\{\mathbf{p}_t\}_{t=1}^T) = T \max_{\mathbf{p}} \{\mathfrak{R}(\mathbf{p})\} - \sum_{t=1}^{T} \mathfrak{R}(\mathbf{p}_t). \tag{4.11}$$

In the online pricing setting, for each customer, our pricing decisions directly impact her click and purchase behavior and thereby further impact the data collection process. Thus, optimal pricing policies should consider not only the revenue from the current customer but also the

data collection process that will influence future decisions. For example, we might need to leverage prices to incentivize customers to click on some rare products in the training set, in order to design better prices in the future. This is referred to as the "exploration-exploitation" tradeoff in the online learning. How to balance this tradeoff is the crux of designing online pricing policies.

In our click model, to minimize the regret, one interesting observation is that we do not need to do the exploration actively. In other words, our online pricing algorithm, Algorithm 7, is an exploration-free algorithm that selects the best price greedily to maximize the revenue from each customer.

Algorithm 7 Greedy Online Pricing Algorithm with Click Data

1: **Initialization:** Initialize the values of prices randomly.
2: **For** $t = 1, ..., T$ **do**
3: Customer t arrives.
4: Collect the click transition data and purchase behavior of customer t. Update the set of click transition pairs $\mathbb{C}_S$. Update N_i^B and W_{ii} for all products i.
5: Solve Problem (4.2) to update the estimation of attraction matrix $\hat{\boldsymbol{\rho}}$.
6: Update instant price elasticity $\hat{\alpha}_i \leftarrow \arg\max_{\alpha_i} \ell_{\text{purchase}}(\alpha_i)$ in (4.5), $\forall i \in [n]$.
7: Given the current availability of products, maximize the expected total revenue to obtain the optimal price $\mathbf{p}_t$, with parameter $(\hat{\boldsymbol{\rho}}, \hat{\boldsymbol{\alpha}})$.
8: Set price as $\mathbf{p}_t$.
9: **End For**

The basic idea of Algorithm 7 is as follows. For each customer t, we first collect the click and purchase data from this customer. Then, we update our estimation for the parameters in MDAC. Based on the new estimation, we optimize the prices to maximize the expected total revenue. Compared to traditional online pricing algorithms that require actively exploring the demands of different products at various prices, our online pricing algorithm is computationally easier. This exploration-free online algorithm also enjoys a small regret bound, which will be shown in Section 4.4.3.

4.4.3 Regret Analysis for Greedy Pricing Policy

To analyze the regret of Algorithm 7, we first make Assumption 4.4.1 on the boundedness of the instant purchase probability. Assumption 4.4.1 assumes the boundedness of prices and instant price elasticity, which is reasonable in practice.

Assumption 4.4.1. *The prices for products are bounded by $[\underline{p}, \overline{p}]$ where $\underline{p} > 0$; The instant price elasticity α_i for all products is bounded by $[\underline{\alpha}, \overline{\alpha}]$ where $\underline{\alpha} > 0$.*

Recall that Assumption 4.3.1 assumes that $\frac{\beta_2}{n} \geq \rho_{ij} \geq \beta_1$, for all $i, j \in [n]$. When Assumptions 4.3.1 and 4.4.1 hold, we have that $\exp(-\overline{p}\overline{\alpha}\beta_2) \leq \mu \leq \exp(-\underline{p}\underline{\alpha}n/\beta_1) \leq$

$\exp(-\underline{p}\underline{\alpha}/\beta_1) < 1$. For simplicity, we denote the lower and upper bounds for μ by $\underline{\mu}$ and $\overline{\mu}$, respectively. We use $\|\cdot\|_1$ to denote the maximum absolute sum of the rows in the matrix, i.e., $\|\boldsymbol{\rho}\|_1 = \max_{i\in\bar{S}} \sum_{j\in\bar{S}} |\rho_{ij}|$. We further denote $\max_{i\in\bar{S}} |\alpha_i|$ by $\|\boldsymbol{\alpha}\|_\infty$. Recall $r_i(\mathbf{p}; \boldsymbol{\rho}, \boldsymbol{\alpha})$ denotes the stationary revenue of the customer clicking product i, given the price vector $\mathbf{p}$ and parameters $(\boldsymbol{\rho}, \boldsymbol{\alpha})$. Before explaining how we derive the regret of our online exploration-free algorithm, we first show that r_i in Algorithm 6 has a Lipschitz property, with respect to the prediction error of $\boldsymbol{\rho}$ and $\boldsymbol{\alpha}$.

Lemma 4.4.1. *Suppose that Assumptions 4.3.1 and 4.4.1 hold. Assume* $\mathbf{p}_1$ *and* $\mathbf{p}_2$ *are the optimal price vectors with respect to the parameters* $(\boldsymbol{\rho}_1, \boldsymbol{\alpha}_1)$ *and* $(\boldsymbol{\rho}_2, \boldsymbol{\alpha}_2)$*, respectively. Then we have that* $r_i(\mathbf{p}_1; \boldsymbol{\rho}_1, \boldsymbol{\alpha}_1) - r_i(\mathbf{p}_2; \boldsymbol{\rho}_1, \boldsymbol{\alpha}_1) \le L_1\|\boldsymbol{\rho}_1 - \boldsymbol{\rho}_2\|_1 + L_2\|\boldsymbol{\alpha}_1 - \boldsymbol{\alpha}_2\|_\infty$*, for any product* $i \in [n]$*, where* $L_1 = \frac{2\overline{\alpha}\overline{p}^2}{\underline{\mu}+\underline{\alpha}\overline{p}}$ *and* $L_2 = \frac{(1+\overline{\alpha}^3\overline{p})\overline{p}}{\underline{\alpha}(\underline{\mu}+\underline{\alpha}\overline{p})}$.

We explain here why exploration-free algorithms work in theory: First, the instant purchase probability μ is assumed to be upper bounded by $\overline{\mu}$. It ensures that regardless of prices, each customer has some positive probability of transiting to any state. This implies that the amount of click data corresponding to any transition pair grows at least linearly in terms of the number of customers with high probability. Thus, according to Lemma 4.3.3, the estimation error of $\boldsymbol{\rho}$ decreases in the order of $\mathcal{O}(\sqrt{1/T})$, where T is the number of customers, irrespective of the prices. Second, the estimation error of the price elasticity α_i also decreases in the order of $\mathcal{O}(\sqrt{1/T})$ for all the products by the lower bound of the transition probability. Hence, the estimation error bounds of both parameters $\boldsymbol{\rho}$ and $\boldsymbol{\alpha}$ decrease in the order of $\mathcal{O}(\sqrt{1/T})$, which are independent of the prices. As shown in Lemma 4.4.1, the one-step regret can be upper bounded by the sum of the estimation errors of $\boldsymbol{\rho}$ and $\boldsymbol{\alpha}$. Therefore, combining all these results, our exploration-free algorithm achieves a regret of order $\tilde{\mathcal{O}}(\sqrt{T})$.

Lemma 4.4.2 (Regret bound under full availability)**.** *Suppose that Assumptions 4.3.1 and 4.4.1 hold. Assume the rank of the attraction matrix is* r *and all products are available. There exists a constant* C *such that when* $T \ge C(1 + \ln(\frac{1}{\delta}))$*, with probability* $1 - 3\delta$*, the regret of Algorithm 7 is at most*

$$\textit{Regret}(\{\mathbf{p}_t\}_{t=1}^T) \le \Big(\frac{2L_2}{\underline{p}} + \frac{c_2 L_1}{\beta_1^{2.5}}\sqrt{nr}\Big)\sqrt{T\ln\Big(\frac{nT}{\delta}\Big)},$$

for some constant $c_2 > 0$.

In Lemma 4.4.2, the regret is in the order of $\tilde{\mathcal{O}}(\sqrt{nrT})$. The order of T matches the lower bound of regret of the online pricing problem in J. Broder and Rusmevichientong, 2012. In the multi-product problem, the naive regret grows in $\mathcal{O}(n)$, by treating each product independently. In our study, by considering and utilizing the low-rank structure of the attraction matrix, we reduce the regret to $\tilde{\mathcal{O}}(\sqrt{nr})$, which is a significant reduction.

Next, we consider the case where the availability of products is limited and dynamically changing, for example, due to stockouts or logistic delays. Note that in order to minimize

regret, we do not need to recover the entire attraction matrix, especially when the availability at each time is limited. Thus, we consider a new pricing policy where each available product set is treated independently. In this case, at each iteration, given the product availability, we estimate the submatrix of the attraction matrix by using the previous click data collected under this available product set. Then, we optimize the prices of the available products based on the estimation of this submatrix. We refer to this pricing policy as the availability-focused pricing policy. Theorem 4.4.1 shows that in the worst case, for any sequence of product availabilities, the regret bound of our availability-focused pricing algorithm is sublinear with respect to the number of customers.

Theorem 4.4.1 (Worst-case regret bound under availability-focused pricing policy)**.** *For any sequences of the available product sets, under the same conditions as Lemma 4.4.2, the regret of availability-focused pricing policy satisfies that with probability at least* $1-2\delta$*, for any sequence of available product sets, the cumulative regret is at most*

$$\left(\frac{c_2 L_1}{\beta_1^{2.5}}\sqrt{nr2^n} + 2\frac{L_2 c_1}{\underline{p}(1-\overline{\mu})}\right)\sqrt{T\ln(T/\delta)}.$$

In Theorem 4.4.1, the regret bound is a square root function of the number of customers T. Since we consider the worst case of the available product sets, the scale of the regret bound is $\mathcal{O}(2^{n/2})$.

Note that the availability-focused pricing policy does not recover the entire attraction matrix. This means that when a specific combination of available products has never appeared before, the availability-focused pricing policy would perform badly because the data from other available product sets is not used. To address this issue, retailers need to estimate the preferences for all the products under dynamic availability. To recover the entire attraction matrix under the dynamic availability, as indicated by Proposition 4.3.1, the set $\mathbb{S}$ must be a cover of all products. Thus, we next consider a special case of product availability. Particularly, we assume that for each customer, the available product set is uniformly chosen from a set $\mathbb{S}$. In this case, the scale of the regret bound can be much smaller than $\mathcal{O}(2^n)$ in Theorem 4.4.1. We further denote the maximum number of available products for one customer by $\tilde{N}$. Then, Theorem 4.4.2 provides the regret bound under the dynamic availabilities.

Theorem 4.4.2 (Regret bound under dynamic i.i.d. availabilities)**.** *Suppose that at each iteration, the set of available products is uniformly selected from a set* $\mathbb{S}$*, where* $\mathbb{S}$ *is a cover of all products. Under the same conditions as Lemma 4.4.2, there exists a constant* C *such that when* $T \geq C|\mathbb{S}|(1+\ln(\frac{|\mathbb{S}|}{\delta}))$*, with probability* $1-4\delta$*, the regret of Algorithm 7 is at most*

$$\textit{Regret}(\{\mathbf{p}_t\}_{t=1}^T) \leq \Big(\frac{2L_2}{\underline{p}} + \frac{c_2 L_1}{\beta_1^{2.5}}\sqrt{\tilde{N}r}\Big)\sqrt{|\mathbb{S}|T\ln\Big(\frac{\tilde{N}T}{\delta}\Big)},$$

for some constant $c_2 > 0$*.*

The upper bound in Theorem 4.4.2 shows that our exploration-free online pricing algorithm achieves the regret in the order of $\tilde{\mathcal{O}}(\sqrt{\tilde{N}|\mathbb{S}|rT})$ under the dynamic availability. We compare this regret bound with the regret bounds in Lemma 4.4.2 and Theorem 4.4.1 as follows. Intuitively, when one customer can see at most $\tilde{N}$ available products, then in practice, we expect that the cardinality of $|\mathbb{S}|$ is $\tilde{\mathcal{O}}(n^2/\tilde{N})$ to constitute a cover of the entire product set. When $|\mathbb{S}| \leq \tilde{\mathcal{O}}(n^2/\tilde{N})$, the order of the regret bound in Theorem 4.4.2 is $\tilde{\mathcal{O}}(n\sqrt{rT})$, which is a significant reduction from the worst case in Theorem 4.4.1 and is slightly larger than the regret bound with full availability. The scale of the regret bound gets larger than the bound in Lemma 4.4.2 because when the availability of products dynamically changes, we need a longer time to accumulate the click data under various availabilities to estimate the entire attraction matrix. However, this regret bound is still sublinear with respect to the number of customers T. As a special case of Theorem 4.4.2, when $\tilde{N} = |\mathbb{S}|$, i.e., all products are available, the regret bound reduces to $\tilde{\mathcal{O}}(\sqrt{nrT})$ with $|\mathbb{S}| = 1$. This result is consistent with Lemma 4.4.2 under the full availability.

4.5 Numerical Experiments

In this section, we test our estimation algorithms and greedy pricing policies using the real-world click data. In Sections 4.5.1 - 4.5.4, we explain our use of real-world click data to demonstrate the effectiveness of our estimation algorithm for Problem (4.2). In Section 4.5.5, we discuss how we test the performance of our greedy online algorithm, Algorithm 7, using real-world data. In Section 4.5.6, we use synthetic data to verify the theoretical insights on the optimal prices and the impact of product availability.

4.5.1 Data Preprocessing For the Real-World Data

In our numerical experiments, we use the public click data in JD.com, 2020 to estimate the attraction matrix. In the original dataset, we have two tables describing the order information and click data. Table 4.1 shows one sample of the records in these two tables. In the first dataset, we have the IDs of items, users, and timestamps of one-click events. In the second dataset, we have the purchase information of one order, including the IDs of the order, user, item, prices, and purchase time.

JD.com, 2020 provided click data and purchase data for more than 31,000 products. To select the proper products for our experiments, we calculated the total number of clicks for all the products and selected the top 50 products. We then filtered the click data and purchase data to restrict our dataset to these 50 products, resulting in six million click records and more than 180,000 purchase records. Note that the first transition pair of each click trajectory depends on both the arrival probability and the attraction matrix. Thus, when estimating the transition matrix, for each click trajectory, we drop the first click transition pair, and add the rest of the click transition pairs into the set $\mathbb{C}_S$.

sku_ID	user_ID	request_time	channel
581d5b54c1	9a71d488b4	2018-03-01 01:09:25	mobile

order_ID	user_ID	sku_ID	order_time	final_unit_price
d0cf5cc6db	0abe9ef2ce	581d5b54c1	2018-03-01 17:14:25.0	79.0

Table 4.1: Sample of dataset

We connected the click records and purchases by their time and customer ID. To illustrate, for one customer, we divided the sequence of her click records into multiple click trajectories. Each click trajectory for this customer showed her entry into the system, her browsing of the products, and either her departure from the system or a purchase. We identified one click trajectory for one customer according to the following rules: If we found no click data for this customer after the last purchase or no click data for this customer during the past 24 hours, the new click record of this customer was set as the starting point for a new click trajectory. Thus, if we found no click behavior or purchase by a customer within 24 hours after the last click record, we assumed that this customer left the system without a purchase. In the MDAC, we assumed that a customer has to transit to the state of product i before she purchases product i. As a result, customers are assumed to click product i before purchasing it. However, in the dataset, there are cases in which customers may click product j instead of i where $j \neq i$, before purchasing product i. Therefore, for each purchase record, we added one click record of the same product at the same time artificially into the click dataset. Then, within the click trajectory—for example, $\{i_0, i_1, i_2, i_3, ..., i_t\}$—we generated the set of click transition pairs $\{(i_1, i_2), (i_2, i_3)...., (i_{t-1}, i_t)\}$, where i_t is the state of no-purchase or purchase. On average, each customer makes 5.72 clicks before purchasing a product or leaving without any purchase. These click sequences contain repeated clicks on the same products, which indicates the back-and-forth click behaviors of customers.

4.5.2 Estimation Methods

In this section, we provide the details of our estimation method and benchmark method. There are three sets of parameters, The attraction matrix $\boldsymbol{\rho}$, price elasticity $\boldsymbol{\alpha}$, and the arrival probability λ. Since we select 50 products with the top click rate, the attraction matrix $\boldsymbol{\rho}$ is a $\mathbb{R}^{51\times 51}$ matrix.

Firstly, we implement the benchmark method (the orange and green curves in Figure 4.2) in the following way. This estimation method only uses the purchase data. Due to the mathematical relations of the purchase probability between our MDAC model and the traditional GMCCM model, this benchmark can also be viewed as the estimation method for the GMCCM with sales data with fixed assortments. In this method, the attraction

matrix $\boldsymbol{\rho}$, price elasticity $\boldsymbol{\alpha}$, and the arrival probability λ are estimated by minimizing the negative log-likelihood function. However, optimizing these three sets of parameters jointly is non-trivial since the problem is non-convex. There is no literature that provides efficient algorithms to estimate these parameters, and it is still an open question to estimate these parameters jointly in MDAC in general cases. Thus, we adopt a heuristic alternative optimization algorithm. Particularly, we fixed the other parameters and optimized $\boldsymbol{\rho}$, α, and λ in sequence to minimize the negative likelihood function. We repeat iterations until the reduction is small or the number of iterations exceeds the threshold (which is set as 5). Within each iteration, we use the gradient descent to optimize $\boldsymbol{\rho}$, $\boldsymbol{\alpha}$, and λ. The gradients of each entry in $\boldsymbol{\rho}$, $\boldsymbol{\alpha}$, and λ are calculated as follows. Recall that the purchase probability in MDAC is given in (4.6). The derivative of the purchase probability for product i with respect to the attraction matrix $\boldsymbol{\rho}$ is

$$\frac{\partial \pi(\boldsymbol{\rho}; i, S)}{\partial \boldsymbol{\rho}} = \lambda (I_n - Diag(1 - \mu(i; \boldsymbol{p}))\boldsymbol{\rho})^{-1} Diag(1 - \mu(i; \boldsymbol{p})) \cdot \\ I_n (I_n - Diag(1 - \mu(i; \boldsymbol{p}))\boldsymbol{\rho})^{-1} ([\Pi(S)]_{\cdot,i})^T,$$

where $[\Pi(S)]_{\cdot,i}$ represents the ith column of the matrix $\Pi(S)$. The last column represents the no-purchase alternative.

The derivative of the no-purchase probability for the no-purchase options with respect to the attraction matrix $\boldsymbol{\rho}$ is

$$\frac{\partial \pi(\lambda; i, S)}{\partial \boldsymbol{\rho}} = \left(\lambda (I_n - Diag(1 - \mu(i; \boldsymbol{p}))\boldsymbol{\rho})^{-1}\right)^T Diag(1 - \mu(i; \boldsymbol{p})).$$

Throughout the experiments, we assume all products share the same α_i. When estimating α and λ, we also use gradient descent to minimize the negative log-likelihood function. In experiments, we use the numerical approximation to obtain a gradient of α and λ.

Secondly, we implement our proposed estimation method (the blue curve in Figure 4.2) in the following way. It uses both the purchase and click data. To leverage the click data, we estimate the attraction matrix $\boldsymbol{\rho}$ by solving Problem (4.2). The click data in the real world contains a lot of noise and outliers. Thus, we need to clean the data carefully. The data pre-processing part is the same as that for 50 products in Section 4.5.1 except for the following few rules:

- The channels of click data contain 'APP', 'PC', 'mobile', and 'others'. Among all channels, the click data within the 'APP' counts about 85% of the total clicks. Thus, after selecting ten products with the top click rates, we further restrict the click data in the channel of 'APP'. These ten products are the available product set set.

- To select the orders that are related to the clicks in the 'APP', we remove the purchase records that have no click data in 'APP' before the purchase. These purchases may come from 'PC' or other channels.

- To avoid the correlations between click data, we select only one click stream from one customer. In other words, if a customer has multiple purchase behaviors (including no-purchase) in the dataset, we only select the first click stream.
- We aggregate multiple consecutive clicks on the same products by the same customer within one minute. The single click without subsequent clicks or purchases is treated as noise and thus removed.

In the data cleaning process described above, we selected only one of the channels ('APP') and 10 specific products. This makes identifying no-purchase behavior challenging. Customers might click on products in other channels or on products not among the selected 50. Since we still record these as no-purchase behaviors in our click data set, the estimation of the transition probability to the no-purchase state, $\rho_{i,0}$, might be inaccurate. These inaccurate estimation of $\rho_{i,0}$ may cause bias for the pricing decision and estimation results. Thus, we re-estimate the transition probability to the no-purchase state in the following way. We re-estimate the value of $\rho_{i,0}$ based on the solutions of Problem (4.2). Particularly, suppose $w_{i,j}$ is estimation results by Algorithm 4 of Problem (4.2), $j \in [\bar{n}]$ in the training set, then the value of attraction matrix $\rho_{i,j}$ is

$$\rho_{i,j} = \begin{cases} (1-\rho_{i,0})\frac{w_{i,j}}{\sum_{j\in[\bar{n}]} w_{i,j}}, & j \neq 0 \\ \rho_{i,0}, & j = 0. \end{cases}$$

$\rho_{i,0}$ are optimized iteratively to minimize the negative likelihood function of the purchase probability. For α and λ, we adopt the alternative optimization method to update these values. When updating α, we minimize the likelihood of the clicks in Equation (4.5).

After the data cleaning process, we select 4000 customers in the test set, which counts about 25% of the total customers. For the rest of the customers, we select $5, 25, 45, .., 125$ customers into the training set for each trial. We run 10 independent trials in total. The 85% confidence intervals of the prediction error are shown in Figure 4.3.

When estimating the attraction matrix, since we use the nuclear norm to approximate the low-rank constraint, the optimal solution of problem (4.2) does not have an exact low-rank structure. Thus, we use the following rule to transform its output to a low-rank attraction matrix. Suppose the desired rank is r and the output attraction matrix is $\hat{\boldsymbol{\rho}}_\gamma$. We use SVD to decompose $\hat{\boldsymbol{\rho}}_\gamma$ to $\hat{\boldsymbol{\rho}}_\gamma = \hat{U}\hat{\Sigma}\hat{V}^T$, where $\hat{\Sigma}$ is a diagonal matrix, whose entries are the singular values of $\hat{\boldsymbol{\rho}}_\gamma$. Then, we keep the first r largest singular values in $\hat{\Sigma}$ and set the others to zero. This filtered diagonal matrix is denoted by Σ_r. Then, we obtain a low-rank matrix $\boldsymbol{\rho}_r$ by $\boldsymbol{\rho}_r = \hat{U}\Sigma_r\hat{V}^T$. Then, we renormalize each row of $\boldsymbol{\rho}_r$ such that the sum of the row is one, which yields an attraction matrix $\hat{\boldsymbol{\rho}}^r_{low}$.

The value of the penalty multiplier γ controls the rank of the output attraction matrix. Given a value of rank r, the value of γ is chosen by a grid search, such that the approximation error $\|\hat{\boldsymbol{\rho}}_\gamma - \hat{\boldsymbol{\rho}}^r_{low}\|_2$ is minimized. We initialize randomly and then use gradient descent to optimize. The step size is chosen adaptively according to Kallus and Udell, 2020. Figure 4.1

shows the sum of the mean squared error of all rows in the test set, versus the rank of the matrix. It shows that when the rank is less than 10, the total prediction error benefits from the increase in rank and that when the rank is larger than 10, the benefit is not much. Thus, we choose 10 as the rank of the attraction matrix.

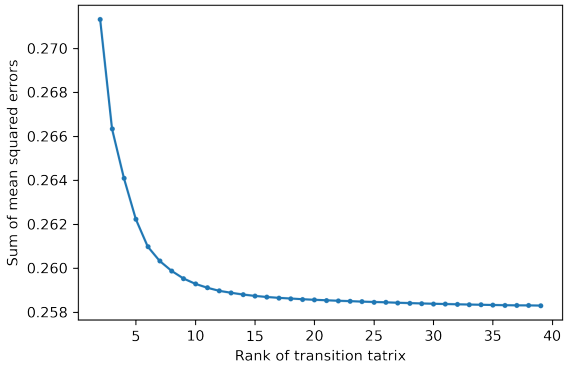


Figure 4.1: Rank selection.

In the next section, we demonstrate how to predict the purchase behaviors of each customer based on the estimation results of the parameters in the MDAC.

4.5.3 Prediction Methods

In this section, we provide the details of how to predict the purchase probability, given the estimated parameters in MDAC in Section 4.5.2. Additionally, we demonstrate that incorporating click data can reveal the currently browsed product and reduce prediction error, as evidenced by the results in Figure 4.2.

In Figures 4.2, the MSE for the prediction error is calculated as follows. Suppose there are T customers, n products. Customer t made w_t clicks in the click stream.

The average prediction error for the orange curve is calculated as:

$$\frac{1}{T}\frac{1}{n}\sum_{i\in[\bar{n}]}\sum_{t=1}^{T}\left(p_{i,t}-N_{i,t}\right)^2,$$

where $p_{i,t}$ is the predicted purchase probability for product i, $i \in [\bar{n}]$ for customer t.

The prediction error corresponding to the green curve and blue curve in Figure 4.2 and the blue curve in Figure 4.3 is:

$$\frac{1}{T}\frac{1}{n}\frac{1}{w_t}\sum_{i\in[n]}\sum_{t=1}^{T}\sum_{j=1}^{w_t}\left(p_{i,t}^{j}-N_{i,t}\right)^2,$$

where $p_{i,t}^{j}$ is the predicted purchase probability for product i for $i \in [\bar{n}]$ conditional on that the current state is the clicked product for click j of customer t.

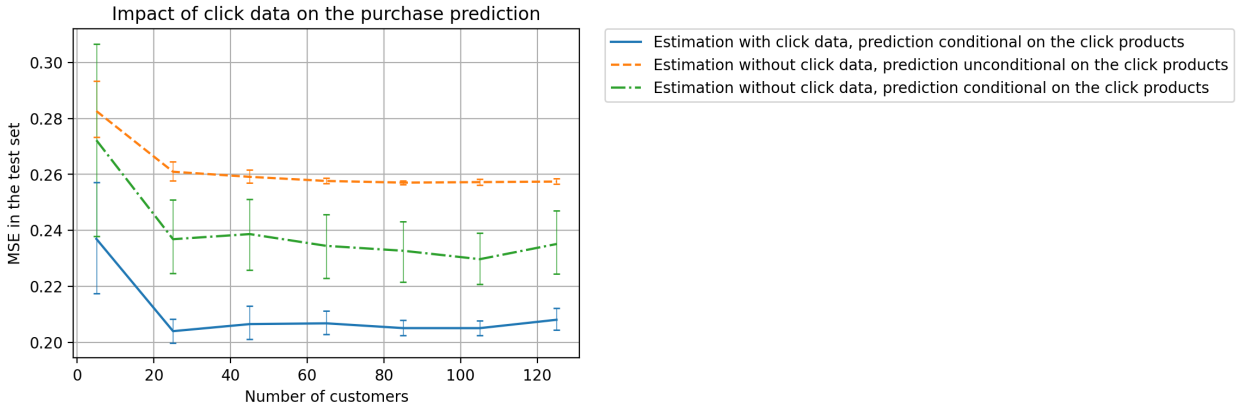


Figure 4.2: Predicting error of purchase probability for different estimating methods

The orange curve in Figure 4.2 shows the prediction error (mean squared error) of purchase behavior when the click data is not used in the estimation or prediction process. The blue and green curves use the same prediction methods but different estimation methods. These two curves show the prediction errors when utilizing the click data in the prediction process, by assuming that the current clicked product of one customer represents the state of this customer in MDAC. In the estimation process, the green curve only uses the sales data to estimate the parameters in MDAC, which is the same as the orange line, while the blue curve uses both click and sales data to estimate the parameters in MDAC.

Thus, Figure 4.2 demonstrates the benefit of click data in the following two ways:

- First, by comparing the green and orange curves, we note that these two methods have the same estimation process of the parameters, but the predictions of the purchase probability are different, and have different MSE. Thus, we conclude that the prediction that is conditional on the current clicked products has a lower MSE than the prediction that does not consider the clicked products. This supports the assumption that the current clicked product represents the current state of this customer in MDAC.

- Second, by comparing the green curve and blue curves, it shows that using both click and sales data to estimate the parameters in MDAC has a smaller MSE than the estimation method that only uses the sales data. This indicates the value of considering the click data when estimating the parameters in MDAC.

We would like to emphasize that, the estimation methods in Section 4.5.2 show a much better performance than the benchmarks in Figure 4.2, without utilizing the low-rank structure

of the transition matrix. When the number of products is large, like 50 products in Section 4.5.1, utilizing the low-rank structure of the attraction matrix could have an even better performance.

4.5.4 Revenue of Pricing Decisions

To illustrate the practical value of connecting MDAC and click behaviors, we further calculate the expected revenue of our click model and the benchmark method. In particular, we use the MDAC model to predict the purchase probability and compare the revenue of the pricing decisions based on the following two pricing methods:

- In the first method, we only use the sales data to estimate the parameters in MDAC and predict the purchase probability.
- In the second method, we use both click and sales data to estimate the parameters in MDAC.

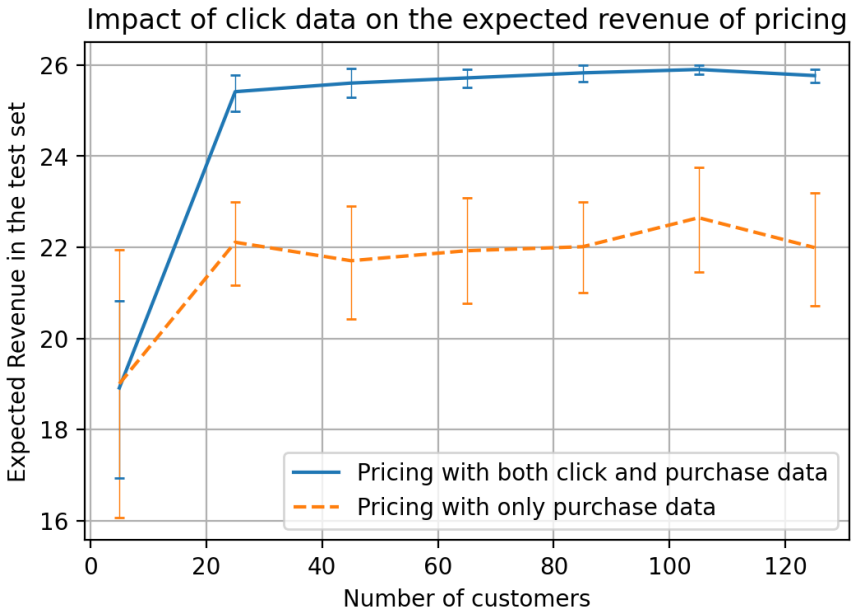


Figure 4.3: Expected revenue under MDAC model using click or click + sales data.

In Figure 4.3, the estimations of the parameters for the orange and blue curves are the same as Figure 4.2. The optimal prices under each set of parameters are obtained by Algorithm 6. After obtaining the optimal prices, the expected revenue in the test set is

calculated by the stationary revenue defined in Definition 4.4.1. The true parameters of the MDAC model in the test set, are estimated by both the click and purchase data. This estimation method in the test set is the same as the estimation method of the blue curve in Figure 4.2.

Our results demonstrate the effectiveness of leveraging click data to improve revenue. Figure 4.3 shows the expected revenue of the optimal pricing under two estimation methods. The orange curve represents the revenue from optimal pricing when the parameters are estimated by the purchase data only; the blue curve represents the revenue when parameters are estimated by both the purchase and click data. By comparing these two curves, we show that utilizing the click data can help to find the best prices and increase the revenue by approximately 18%.

4.5.5 Performance of Online Greedy Algorithms

We examine the performance of our proposed online Algorithm 7 using simulation. To generate the click transition pairs from the click trajectory, we used the following rules. We selected 10 products with the highest click rate as the entire product set. We selected the first $20,000$ click transition pairs as the training set and the following $10,000$ click transition pairs as the test set. Suppose the true value of the attraction matrix is the transformed attraction matrix with rank 10 under the training set in Section 4.5.1, which is denoted by $\boldsymbol{\rho}^*$. For simplicity, we assume all products share the same α_i in the numerical experiments. The true value of price elasticity α_i is also the MLE estimator in the training set, whose value is $\alpha_i^* = 0.0523$. The arrivals of customers are generated according to the real-world data in the training set. The prices of all the products we offer are unchanged for one customer. When this customer leaves the system, we will update the estimation of parameters in the model, and update the prices for all the products. The transition behaviors and purchase behaviors of customers are generated from the simulation under the true parameters $(\boldsymbol{\rho}^*, \boldsymbol{\alpha}^*)$. Since we cannot observe the cost of the products, the cost is assumed to be $c = 0.7 \times p \times (1 + \epsilon)$, where ϵ follows the standard Gaussian distribution.

As pointed out by Theorem 4.4.2, when the number of available products is limited, the set of various available product sets should constitute a cover of the entire product. In our numerical experiment, we change the number of available products from 10 to 8. When all of ten products are available, then the available product set itself is a cover. When only 9 out of 10 products are available, in order to constitute a cover, the minimum number of available product sets is 3, i.e., $|\mathbb{S}|$ is at least 3. When 8 products are available, we need at least 6 available product sets to form a cover. To compute the regret of each step, we evaluate $\mathfrak{R}(\mathbf{p}_t)$ by $\frac{1}{\bar{n}} \sum_{i \in [\bar{n}]} \hat{r}_i^t$, where $\hat{r}_i^t$ are the stationary revenues of product i, given the prices as p^t and the parameters as $(\boldsymbol{\rho}^*, \boldsymbol{\alpha}^*)$. For each setting, we run 20 independent trials and calculate the mean of one-step regret for 20 trials, which is $\mathfrak{R}(\mathbf{p}^*) - \mathfrak{R}(\mathbf{p}_t)$ and $\mathbf{p}^*$ denotes the optimal prices under true parameters $(\boldsymbol{\rho}^*, \boldsymbol{\alpha}^*)$.

The plot in Figure 4.4 shows the cumulative regret of online algorithms under different numbers of available products, with 95% confidence intervals. This shows that the regret of

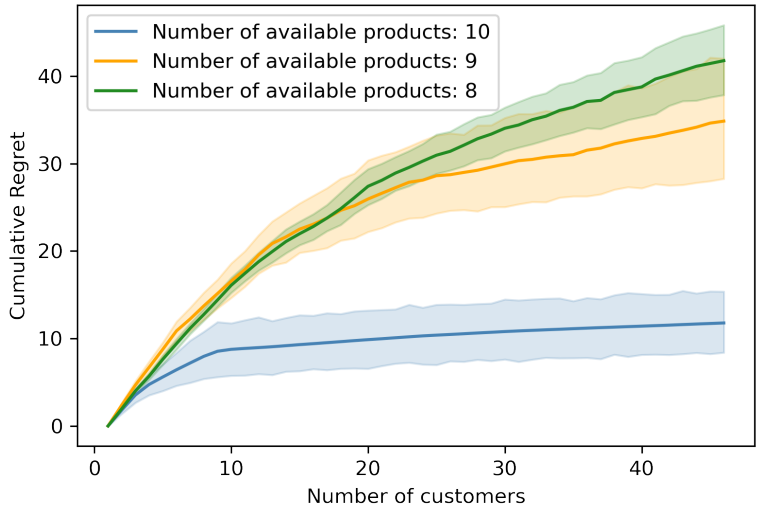


Figure 4.4: Cumulative regret for the exploration-free online algorithm under different sizes of the available product set

our exploration-free online learning algorithm achieves a sublinear regret. It further shows that when the number of available products gets smaller, the scale of regret of our algorithm gets larger because it needs more time to explore the entire product set. This insight is consistent with Theorem 4.4.2.

4.5.6 Synthetic Data

As the last part of the numerical experiment, we explain how we validate our proposed models using synthetic data. With the three sets of numerical experiments, discussed in Section 4.5.6, we demonstrate how the optimal pricing policies depend on the parameters of the MDAC.

Experimental setup. In the synthetic data set, we generate 50 products indexed by $1, 2, ..., 50$, whose costs are all zero. For product i, ρ_{i0} is set as $0.05 + 0.01 \times i$, so the transition probability to the no-purchase state increases as the index increases. The transition probabilities to other states are drawn uniformly for each product. Proposition 4.4.2 indicates that the optimal stationary revenue and optimal prices highly depend on ρ_{i0}. Therefore, this construction will generate a decreasing list of optimal stationary revenue and a decreasing list of optimal prices. In other words, product 1 has the highest optimal stationary revenue, and highest optimal prices. The value of price elasticity α_i in $\mu(\cdot)$ is set as 0.0523 for all

product i, which is the empirical maximizer of the likelihood function of purchase in Section 4.5.5. The optimal stationary revenue of product i is denoted by r_i^*.

Our first example illustrates the properties developed in Proposition 4.4.1, which indicates that the optimal prices of one product depend on both the optimal stationary revenue and the attraction matrix. The second example shows the non-monotonicity of the optimal stationary revenue with respect to the probability to transit to the no-purchase state. It also shows the impact of the cost.

4.5.6.1 Verification and insights of Proposition 4.4.1.

To verify Proposition 4.4.1, we change the attraction matrix as follows. We select the products indexed by 25 and 26, and reduce $\rho_{25,26}$ by 0.01. For each product $m \neq 26$, we increase $\rho_{25,m}$ by 0.01. Figure 4.5 shows the change of the optimal prices for products 25, 26 and m, with respect to the change of optimal stationary revenue $r_m^* - r_{26}^*$. The results, shown in different colors in Figure 4.5, also are shown in different scales. If this difference is positive, customers are more likely to click the products with higher optimal stationary revenue, after clicking product 25. Figure 4.5 shows that when customers are more likely to click the products with higher optimal stationary revenue, rather than the products with less optimal stationary revenue, the prices for all three products (products 25, 26, and m) would increase. On the other hand, when the difference is negative, the prices for all three products would decrease, which is consistent with Proposition 4.4.1. Proposition 4.4.1 further states that the trend of prices is the same for all the products; it is not limited to products 25, 26 and m. Note that in Figure 4.5, there is a spike for product m when $m = 25$. The reason is that when $m = 25$, the probability of transitioning to product 25 itself gets increased. Note that this difference in optimal prices of product 25 is shown in both the orange curve and the blue solid line, but with different scales.

4.5.6.2 Relations between the optimal stationary revenue and ρ_{i0}.

In the second example, we examine the relations between the optimal stationary revenue and ρ_{i0}. In this example, ρ_{i0} is the same as in the basic setting, but the transition probabilities to other states are generated randomly according to a Gaussian distribution and then rescaled such that the row sum is one. We select a set of products indexed by 25, 30, 35, 40, and 45 to increase its cost solely by 20 for each instance. The blue dots in Figure 4.6 demonstrate that there are some products with higher ρ_{i0} having higher optimal stationary revenue, which means that the optimal stationary revenue is not monotone on ρ_{i0}. It also shows that the cost increase of one product will reduce the optimal prices of other products, which is consistent with Lemma 4 in Dong, Simsek, and Topaloglu, 2019.

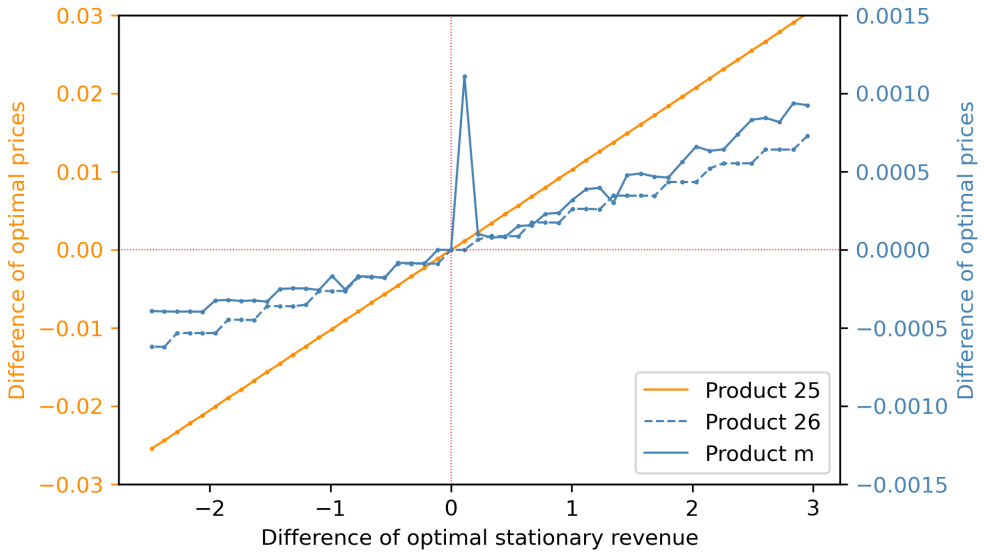


Figure 4.5: Difference of optimal prices vs. the difference of optimal stationary revenue.

4.5.6.3 Relations between the optimal prices and the size of available product set.

In this example, we examine the relations between the optimal prices and the size of available products. We consider 50 products in total. The attraction matrix is generated randomly. The instant price elasticity is the same as the previous setting. The cost for product i is $i+5$. We select the first X products as available products, where X is from 5 to 40 products. The optimal prices are obtained by Algorithm 6. We demonstrate the optimal prices for the first 4 products as examples in Figure 4.7. The x-axis is the number of available products.

Figure 4.7 shows that when more products become available, the prices for the first four products increase. To intuitively explain this trend, we first claim that when the optimal stationary revenue of one product decreases, the optimal prices of other products would decrease. This claim is shown in the proof of Proposition 4.4.1. Moreover, the fewer products are available, the more products deviate from their optimal prices. Consequently, their stationary revenue decreases. As a result, the optimal price of products decreases.

4.6 Concluding Remarks

In this chapter, we studied how to use the high-dimensional click transition data of customers to make pricing decisions for online retailers. We proposed the MDAC model that connects the click and purchase behaviors. We provided efficient algorithms to solve estimation

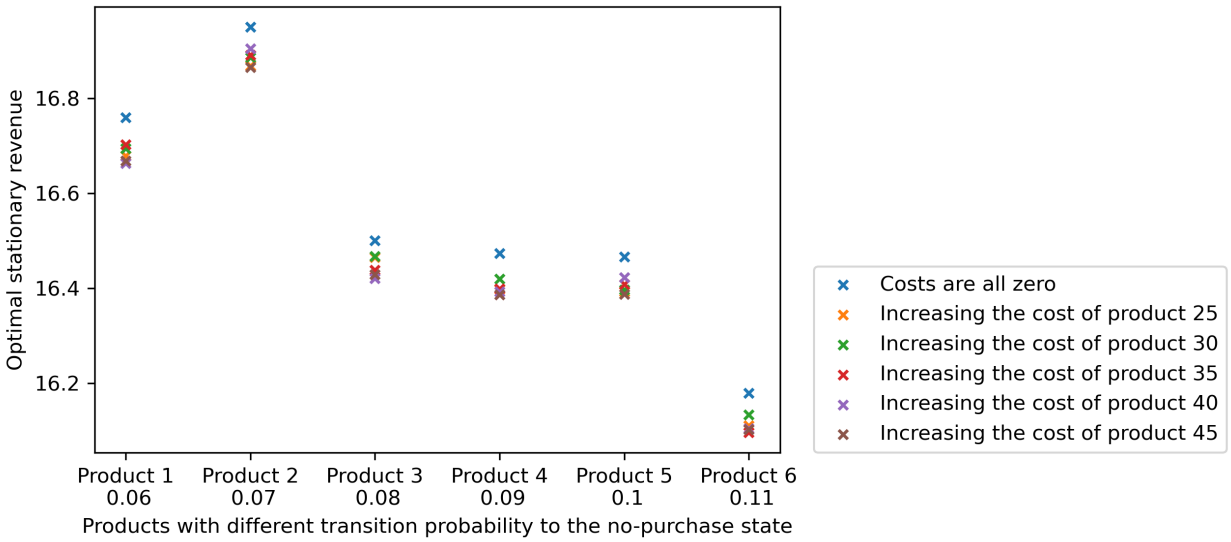


Figure 4.6: Optimal stationary revenue vs. ρ_{i0}.

problems under the dynamic availability of products. To address the scalability issue when estimating the parameters in the MDAC, we utilized the low-rank structure of the attraction matrix in the MDAC. Moreover, we studied the online pricing problem, and provided an exploration-free algorithm. We showed that by leveraging the click data, the estimation error bound and regret can be much smaller than the methods that use sales data only. We also conducted numerical experiments on the real-world dataset, which verified that our click models and greedy pricing policy can help retailers reduce the learn customers' preferences and increase revenue.

From this work, we see several intriguing directions for future research that are worth investigating. First, we assume that the instant purchase probability is a function of price but that the attraction matrix is independent of price. This assumption is consistent with most studies that incorporate prices into the Markov chain process (e.g., (Dong, Simsek, and Topaloglu, 2019; Kleywegt and Shao, 2022)). However, a more general scenario in which the attraction matrix depends on price is worth studying as it adds more model complexity to the attraction matrix. Secondly, our current model assumes the stationarity and homogeneity of customers. In future research, a nonstationary preference and personalized pricing strategy in the MDAC would be worth investigating.

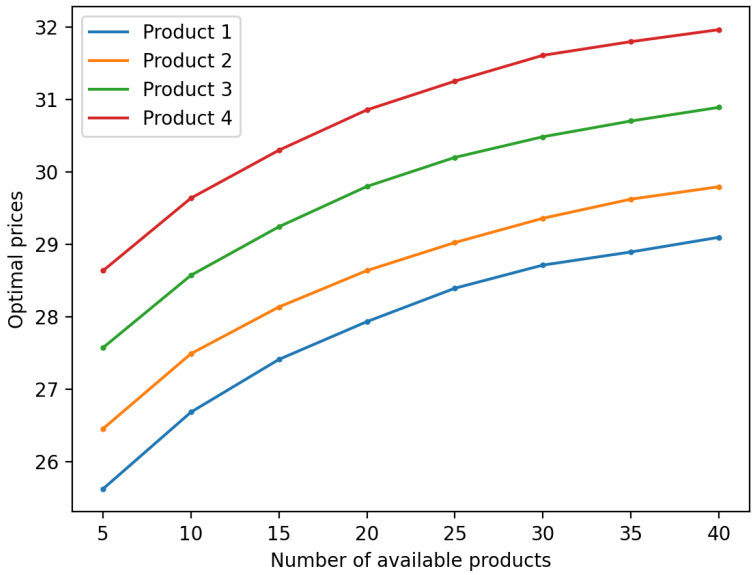


Figure 4.7: Optimal prices vs. Number of the available products

Bibliography

Agarwal, Anish et al. (2021). "Causal Matrix Completion". In: *arXiv preprint arXiv:2109.15154*.

Agrawal, Shipra et al. (2017). "Thompson sampling for the MNL-bandit". In: *arXiv preprint arXiv:1706.00977*.

— (2019). "Mnl-bandit: A dynamic learning approach to assortment selection". In: *Operations Research* 67.5, pp. 1453–1485.

Alptekinoğlu, Aydın and John H Semple (2016). "The exponomial choice model: A new alternative for assortment and price optimization". In: *Operations Research* 64.1, pp. 79–93.

Amaldoss, Wilfred and Chuan He (2018). "Reference-dependent utility, product variety, and price competition". In: *Management Science* 64.9, pp. 4302–4316.

Amos, Brandon and J Zico Kolter (2017). "Optnet: Differentiable optimization as a layer in neural networks". In: *International Conference on Machine Learning*. PMLR, pp. 136–145.

Aouad, Ali, Vivek Farias, and Retsef Levi (2021). "Assortment optimization under consider-then-choose choice models". In: *Management Science* 67.6, pp. 3368–3386.

Aouad, Ali, Jacob Feldman, et al. (2019). "The Click-Based MNL Model: A Novel Framework for Modeling Click Data in Assortment Optimization". In: *Available at SSRN 3340620*.

Aouad, Ali, Retsef Levi, and Danny Segev (2018). "Greedy-like algorithms for dynamic assortment planning under multinomial logit preferences". In: *Operations Research* 66.5, pp. 1321–1345.

Athey, Susan et al. (2021). "Matrix completion methods for causal panel data models". In: *Journal of the American Statistical Association* 116.536, pp. 1716–1730.

Balcan, Maria-Florina, Alina Beygelzimer, and John Langford (2009). "Agnostic active learning". In: *Journal of Computer and System Sciences* 75.1, pp. 78–89.

Balcan, Maria-Florina, Andrei Broder, and Tong Zhang (2007). "Margin based active learning". In: *International Conference on Computational Learning Theory*. Springer, pp. 35–50.

Ban, Gah-Yi and N Bora Keskin (2020). "Personalized dynamic pricing with machine learning: High dimensional features and heterogeneous elasticity". In: *Forthcoming, Management Science*.

Ben-Akiva, Moshe E and Steven R Lerman (1985). *Discrete choice analysis: theory and application to travel demand*. Vol. 9. MIT press.

Berbeglia, Gerardo, Agustín Garassino, and Gustavo Vulcano (2021). "A comparative empirical study of discrete choice models in retail operations". In: *Management Science*.

Berthet, Quentin et al. (2020). "Learning with differentiable pertubed optimizers". In: *Advances in neural information processing systems* 33, pp. 9508–9519.

Bertsekas, Dimitri P (1976). "On the Goldstein-Levitin-Polyak gradient projection method". In: *IEEE Transactions on automatic control* 21.2, pp. 174–184.

Bertsimas, Dimitris and Nathan Kallus (2020). "From predictive to prescriptive analytics". In: *Management Science* 66.3, pp. 1025–1044.

Bertsimas, Dimitris and Velibor V Mišić (2019). "Exact first-choice product line optimization". In: *Operations Research* 67.3, pp. 651–670.

Besbes, Omar, Yonatan Gur, and Assaf Zeevi (2016). "Optimization in online content recommendation services: Beyond click-through rates". In: *Manufacturing & Service Operations Management* 18.1, pp. 15–33.

Besbes, Omar and Assaf Zeevi (2009). "Dynamic pricing without knowing the demand function: Risk bounds and near-optimal algorithms". In: *Operations Research* 57.6, pp. 1407–1420.

Beygelzimer, Alina, Sanjoy Dasgupta, and John Langford (2009). "Importance weighted active learning". In: *Proceedings of the 26th annual international conference on machine learning*, pp. 49–56.

Blanchet, Jose, Guillermo Gallego, and Vineet Goyal (2016). "A markov chain approximation to choice modeling". In: *Operations Research* 64.4, pp. 886–905.

Boer, Arnoud V den and Bert Zwart (2015). "Dynamic pricing and learning with finite inventories". In: *Operations research* 63.4, pp. 965–978.

Broder, Josef and Paat Rusmevichientong (2012). "Dynamic pricing under a general parametric choice model". In: *Operations Research* 60.4, pp. 965–980.

Cai, Wenbin, Muhan Zhang, and Ya Zhang (2016). "Batch mode active learning for regression with expected model change". In: *IEEE transactions on neural networks and learning systems* 28.7, pp. 1668–1681.

Calamai, Paul H and Jorge J Moré (1987). "Projected gradient methods for linearly constrained problems". In: *Mathematical programming* 39.1, pp. 93–116.

Castro, Rui, Rebecca Willett, and Robert Nowak (2005). "Faster rates in regression via active learning". In: *NIPS*. Vol. 18, pp. 179–186.

Chao, Min-Te and WE Strawderman (1972). "Negative moments of positive random variables". In: *Journal of the American Statistical Association* 67.338, pp. 429–431.

Chen, Ningyuan et al. (2021). "Model-free assortment pricing with transaction data". In: *arXiv preprint arXiv:2101.02251*.

Chen, Xi and Yining Wang (2017). "A Note on a Tight Lower Bound for MNL-Bandit Assortment Selection Models". In: *arXiv preprint arXiv:1709.06109*.

Cheung, Wang Chi and David Simchi-Levi (2017). "Thompson Sampling for Online Personalized Assortment Optimization Problems with Multinomial Logit Choice Models". In: *Available at SSRN 3075658*.

Chuklin, Aleksandr, Ilya Markov, and Maarten de Rijke (2015). "Click models for web search". In: *Synthesis lectures on information concepts, retrieval, and services* 7.3, pp. 1–115.

Chung, Tsai-Hsuan et al. (2022). "Decision-Aware Learning for Optimizing Health Supply Chains". In: *arXiv preprint arXiv:2211.08507*.

Cohn, David, Les Atlas, and Richard Ladner (1994). "Improving generalization with active learning". In: *Machine learning* 15.2, pp. 201–221.

Craswell, Nick et al. (2008). "An experimental comparison of click position-bias models". In: *Proceedings of the 2008 international conference on web search and data mining*, pp. 87–94.

Dasgupta, Sanjoy, Daniel J Hsu, and Claire Monteleoni (2007). *A general agnostic active learning algorithm*. Citeseer.

Demirovic, Emir et al. (2020). "Dynamic programming for predict+ optimise". In: *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*. AAAI Press, pp. 1444–1451.

Demirović, Emir et al. (2019). "Predict+ optimise with ranking objectives: Exhaustively learning linear functions". In: *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019*. International Joint Conferences on Artificial Intelligence, pp. 1078–1085.

Désir, Antoine, Vineet Goyal, Srikanth Jagabathula, et al. (2021). "Mallows-smoothed distribution over rankings approach for modeling choice". In: *Operations Research* 69.4, pp. 1206–1227.

Désir, Antoine, Vineet Goyal, Danny Segev, et al. (2020). "Constrained assortment optimization under the Markov chain–based choice model". In: *Management Science* 66.2, pp. 698–721.

Dong, James, A Serdar Simsek, and Huseyin Topaloglu (2019). "Pricing problems under the markov chain choice model". In: *Production and Operations Management* 28.1, pp. 157–175.

Donti, Priya, Brandon Amos, and J Zico Kolter (2017). "Task-based end-to-end model learning in stochastic optimization". In: *Advances in neural information processing systems* 30.

El Balghiti, Othman et al. (2019). "Generalization bounds in the predict-then-optimize framework". In: *Advances in neural information processing systems* 32.

— (2022). "Generalization bounds in the predict-then-optimize framework". In: *Mathematics of Operations Research*.

Elmachtoub, Adam N and Paul Grigas (2022). "Smart "predict, then optimize"". In: *Management Science* 68.1, pp. 9–26.

Elmachtoub, Adam N, Henry Lam, et al. (2023). "Estimate-Then-Optimize Versus Integrated-Estimation-Optimization: A Stochastic Dominance Perspective". In: *arXiv preprint arXiv:2304.06833*.

Elmachtoub, Adam N, Jason Cheuk Nam Liang, and Ryan McNellis (2020). "Decision trees for decision-making under the predict-then-optimize framework". In: *International Conference on Machine Learning*. PMLR, pp. 2858–2867.

Farias, Vivek F, Andrew A Li, and Tianyi Peng (2021). "Uncertainty Quantification For Low-Rank Matrix Completion With Heterogeneous and Sub-Exponential Noise". In: *arXiv preprint arXiv:2110.12046*.

Feldman, Jacob B and Huseyin Topaloglu (2017). "Revenue management under the Markov chain choice model". In: *Operations Research* 65.5, pp. 1322–1342.

Feng, Yifan (2020). "Active Learning in Marketplaces and Online Platforms". PhD thesis. The University of Chicago.

Ferreira, Kris Johnson and Emily Mower (2022). "Demand Learning and Pricing for Varying Assortments". In: *Manufacturing & Service Operations Management*.

Fu, Lei and Dong-Dong Ge (2021). "A Gradient Descent Method for Estimating the Markov Chain Choice Model". In: *Journal of the Operations Research Society of China*, pp. 1–11.

Gafni, Eli M, Dimitri P Bertsekas, et al. (1982). *Convergence of a gradient projection method*. Laboratory for Information and Decision Systems.

Gallego, Guillermo and Wentao Lu (2021). "An Optimal Greedy Heuristic with Minimal Learning Regret for the Markov Chain Choice Model". In: *Available at SSRN 3810470*.

Gao, Pin et al. (2021). "Assortment optimization and pricing under the multinomial logit model with impatient customers: Sequential recommendation and selection". In: *Operations research* 69.5, pp. 1509–1532.

Gao, Ruijiang and Maytal Saar-Tsechansky (2020). "Cost-accuracy aware adaptive labeling for active learning". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 34, pp. 2569–2576.

Gao, Xiangyu et al. (2022). "Joint learning and optimization for multi-product pricing (and ranking) under a general cascade click model". In: *Management Science*.

Goutam, Kumar, Vineet Goyal, and Agathe Soret (2019). "A Generalized Markov Chain Model to Capture Dynamic Preferences and Choice Overload". In: *arXiv preprint arXiv:1911.06716*.

Grigas, Paul, Meng Qi, et al. (2021). "Integrated conditional estimation-optimization". In: *arXiv preprint arXiv:2110.12351*.

Guo, Fan et al. (2009). "Click chain model in web search". In: *Proceedings of the 18th international conference on World wide web*, pp. 11–20.

Hanneke, Steve (2007). "A bound on the label complexity of agnostic active learning". In: *Proceedings of the 24th international conference on Machine learning*, pp. 353–360.

— (2011). "Rates of convergence in active learning". In: *The Annals of Statistics*, pp. 333–361.

Ho, Chin Pang and Grani A Hanasusanto (2019). "On data-driven prescriptive analytics with side information: A regularized nadaraya-watson approach". In: *URL: http://www. optimization-online. org/DB FILE/2019/01/7043. pdf*.

Hu, Yichun, Nathan Kallus, and Xiaojie Mao (2022). "Fast rates for contextual linear optimization". In: *Management Science*.

Jagabathula, Srikanth and Paat Rusmevichientong (2017). "A nonparametric joint assortment and price choice model". In: *Management Science* 63.9, pp. 3128–3145.

Jannach, Dietmar et al. (2010). *Recommender systems: an introduction*. Cambridge University Press.

JD.com (2020). *2020 Data Driven Research Challenge - MSOM Society*. URL: `https://connect.informs.org/msom/events/datadriven2020` (visited on 03/18/2022).

Jun, Kwang-Sung et al. (2019). "Bilinear bandits with low-rank structure". In: *International Conference on Machine Learning*. PMLR, pp. 3163–3172.

Kääriäinen, Matti (2006). "Active learning in the non-realizable case". In: *International Conference on Algorithmic Learning Theory*. Springer, pp. 63–77.

Kallus, Nathan and Xiaojie Mao (2023). "Stochastic optimization forests". In: *Management Science* 69.4, pp. 1975–1994.

Kallus, Nathan and Madeleine Udell (2020). "Dynamic assortment personalization in high dimensions". In: *Operations Research* 68.4, pp. 1020–1037.

Kao, Yi-hao, Benjamin Roy, and Xiang Yan (2009). "Directed regression". In: *Advances in Neural Information Processing Systems* 22.

Katariya, Sumeet et al. (2017). "Bernoulli Rank-1 Bandits for Click Feedback". In: *arXiv preprint arXiv:1703.06513*.

Keskin, N Bora and Assaf Zeevi (2017). "Chasing demand: Learning and earning in a changing environment". In: *Mathematics of Operations Research* 42.2, pp. 277–307.

Kleywegt, Anton J and Hongzhang Shao (2022). "Revenue Management Under the Markov Chain Choice Model with Joint Price and Assortment Decisions". In: *arXiv preprint arXiv:2204.04774*.

Kotary, James et al. (2021). "End-to-end constrained optimization learning: A survey". In: *arXiv preprint arXiv:2103.16378*.

Krishnamurthy, Akshay et al. (2017). "Active learning for cost-sensitive classification". In: *International Conference on Machine Learning*. PMLR, pp. 1915–1924.

Kuznetsov, Vitaly and Mehryar Mohri (2015). "Learning theory and algorithms for forecasting non-stationary time series". In: *Advances in neural information processing systems* 28.

Lei, Yanzhe et al. (2022). "Joint product framing (display, ranking, pricing) and order fulfillment under the multinomial logit model for e-commerce retailers". In: *Manufacturing & Service Operations Management* 24.3, pp. 1529–1546.

Li, Lihong, Yu Lu, and Dengyong Zhou (2017). "Provably optimal algorithms for generalized linear contextual bandits". In: *International Conference on Machine Learning*. PMLR, pp. 2071–2080.

Li, Shukai et al. (2022). "Online Learning for Constrained Assortment Optimization under Markov Chain Choice Model". In: *Available at SSRN 4079753*.

Liu, Heyuan and Paul Grigas (2021). "Risk bounds and calibration for a smart predict-then-optimize method". In: *Advances in Neural Information Processing Systems* 34.

Liu, Mo et al. (2023). "Active Learning in the Predict-then-Optimize Framework: A Margin-Based Approach". In: *arXiv preprint arXiv:2305.06584*.

Liu, Qian and Garrett Van Ryzin (2008). "On the choice-based linear programming model for network revenue management". In: *Manufacturing & Service Operations Management* 10.2, pp. 288–310.

Liu, Qing and Neeraj Arora (2011). "Efficient choice designs for a consider-then-choose model". In: *Marketing Science* 30.2, pp. 321–338.

Loke, Gar Goei, Qinshen Tang, and Yangge Xiao (2022). "Decision-driven regularization: A blended model for predict-then-optimize". In: *Available at SSRN 3623006*.

Lu, Yangyi, Amirhossein Meisami, and Ambuj Tewari (2021). "Low-rank generalized linear bandit problems". In: *International Conference on Artificial Intelligence and Statistics*. PMLR, pp. 460–468.

Luce, R Duncan (1959). *Individual choice behavior*. John Wiley.

Lyu, Chengyi et al. (2021). "Assortment Optimization with Multi-Item Basket Purchase under Multivariate MNL Model". In: *Available at SSRN 3818886*.

Mandi, Jayanta and Tias Guns (2020). "Interior Point Solving for LP-based prediction+ optimisation". In: *Advances in Neural Information Processing Systems* 33, pp. 7272–7282.

Mandi, Jayanta, Peter J Stuckey, Tias Guns, et al. (2020). "Smart predict-and-optimize for hard combinatorial optimization problems". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 34-02, pp. 1603–1610.

McFadden, Daniel et al. (1973). *Conditional logit analysis of qualitative choice behavior*. Institute of Urban and Regional Development, University of California Oakland.

Miao, Sentao and Xiuli Chao (2021). "Dynamic joint assortment and pricing optimization with demand learning". In: *Manufacturing & Service Operations Management* 23.2, pp. 525–545.

Ho-Nguyen, Nam and Fatma Kılınç-Karzan (2022). "Risk guarantees for end-to-end prediction and optimization processes". In: *Management Science*.

Paul, Alice, Jacob Feldman, and James Mario Davis (2018). "Assortment optimization and pricing under a nonparametric tree choice model". In: *Manufacturing & Service Operations Management* 20.3, pp. 550–565.

Puterman, Martin L (2014). *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons.

Qiang, Sheng and Mohsen Bayati (2016). "Dynamic pricing with demand covariates". In: *Available at SSRN 2765257*.

Rakhlin, Alexander, Karthik Sridharan, and Ambuj Tewari (2015). "Sequential complexities and uniform martingale laws of large numbers". In: *Probability theory and related fields* 161, pp. 111–153.

Rusmevichientong, Paat, Zuo-Jun Max Shen, and David B Shmoys (2010). "Dynamic assortment optimization with a multinomial logit choice model and capacity constraint". In: *Operations research* 58.6, pp. 1666–1680.

Saar-Tsechansky, Maytal, Prem Melville, and Foster Provost (2009). "Active feature-value acquisition". In: *Management Science* 55.4, pp. 664–684.

Sauré, Denis and Assaf Zeevi (2013). "Optimal dynamic assortment planning with demand learning". In: *Manufacturing & Service Operations Management* 15.3, pp. 387–404.

Settles, Burr (2009). *Active learning literature survey*. University of Wisconsin-Madison Department of Computer Sciences.

Şimşek, A Serdar and Huseyin Topaloglu (2018). "An expectation-maximization algorithm to estimate the parameters of the markov chain choice model". In: *Operations Research* 66.3, pp. 748–760.

Sugiyama, Masashi and Shinichi Nakajima (2009). "Pool-based active learning in approximate linear regression". In: *Machine Learning* 75.3, pp. 249–274.

Talluri, Kalyan and Garrett Van Ryzin (2004). "Revenue management under a general discrete choice model of consumer behavior". In: *Management Science* 50.1, pp. 15–33.

Tang, Bo and Elias B Khalil (2022). "PyEPO: A PyTorch-based End-to-End Predict-then-Optimize Library for Linear and Integer Programming". In: *arXiv preprint arXiv:2206.14234*.

Tropp, Joel A (2012). "User-friendly tail bounds for sums of random matrices". In: *Foundations of computational mathematics* 12.4, pp. 389–434.

Tulabandhula, Theja, Deeksha Sinha, and Prasoon Patidar (2020). "Multi-purchase behavior: Modeling and optimization". In: *Available at SSRN 3626788*.

Wang, Weiran and Miguel A Carreira-Perpinán (2013). "Projection onto the probability simplex: An efficient algorithm with a simple proof, and an application". In: *arXiv preprint arXiv:1309.1541*.

Wang, Yanqiao and Zuo-Jun Max Shen (2017). *Joint optimization of capacitated assortment and pricing problem under the tree logit model*. Tech. rep. Technical report, University of California, Berkeley, CA.

Wilder, Bryan, Bistra Dilkina, and Milind Tambe (2019). "Melding the data-decisions pipeline: Decision-focused learning for combinatorial optimization". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 33-01, pp. 1658–1665.

Yang, Junwen and Yifan Feng (July 2023). "Nested Elimination: A Simple Algorithm for Best-Item Identification From Choice-Based Feedback". In: *Proceedings of the 40th International Conference on Machine Learning*. Ed. by Andreas Krause et al. Vol. 202. Proceedings of Machine Learning Research. PMLR, pp. 39205–39233. URL: https://proceedings.mlr.press/v202/yang23b.html.

Yurtsever, Alp and Suvrit Sra (2022). "CCCP is Frank-Wolfe in disguise". In: *Advances in Neural Information Processing Systems*. Ed. by Alice H. Oh et al. URL: https://openreview.net/forum?id=OGGQs4xFHrr.

Zheng, Zhiqiang and Balaji Padmanabhan (2006). "Selectively acquiring customer information: A new data acquisition problem and an active learning-based solution". In: *Management Science* 52.5, pp. 697–712.

Zhu, Taozeng, Jingui Xie, and Melvyn Sim (2022). "Joint estimation and robustness optimization". In: *Management Science* 68.3, pp. 1659–1677.

Zhu, Zeyuan Allen et al. (2010). "A novel click model and its applications to online advertising". In: *Proceedings of the third ACM international conference on Web search and data mining*, pp. 321–330.

Zhu, Ziwei et al. (2021). "Learning Markov models via low-rank optimization". In: *Operations Research*.

Appendix A

Proof for Chapter 1

A.0.1 Proofs for Section 2.3

Proof of Lemma 2.3.1. Without loss of generality, assume that $\nu_S(c_1) \geq \nu_S(c_2)$, i.e., we have that $0 \leq \|c_1 - c_2\| < \nu_S(c_1)$. We also claim that $\nu_S(c_2) > 0$. Indeed, since ν_S is a 1-Lipschitz distance function, it holds that $\nu_S(c_1) - \nu_S(c_2) \leq \|c_1 - c_2\| < \nu_S(c_1)$ and hence $\nu_S(c_2) > 0$. As above, let $\{v_j : j = 1, ..., K\}$ be the extreme points of S, i.e., $S = \text{conv}(v_1, \ldots, v_K)$. Since $\nu_S(c_1) > 0$ and $\nu_S(c_2) > 0$, both $w^*(c_1)$ and $w^*(c_2)$ must be extreme points solutions, i.e., $w^*(c_1) = v_{j_1}$ and $w^*(c_1) = v_{j_2}$ for some indices j_1 and j_2.

We now prove the lemma by contradiction. If $w^*(c_1) \neq w^*(c_2)$, then by (2.4), the following two inequalities hold:

$$\nu_S(c_1) \leq \frac{c_1^T(w^*(c_2) - w^*(c_1))}{\|w^*(c_2) - w^*(c_1)\|_*}, \quad \nu_S(c_2) \leq \frac{c_2^T(w^*(c_1) - w^*(c_2))}{\|w^*(c_1) - w^*(c_2)\|_*}.$$

We add up both sides of the above two inequalities, and get

$$\nu_S(c_1) + \nu_S(c_2) \leq \frac{(c_1 - c_2)^T(w^*(c_2) - w^*(c_1))}{\|w^*(c_2) - w^*(c_1)\|_*} \leq \frac{\|c_1 - c_2\|\|w^*(c_2) - w^*(c_1)\|_*}{\|w^*(c_2) - w^*(c_1)\|_*} = \|c_1 - c_2\|,$$

where the second inequality uses Hölder's inequality. Because $\|c_1 - c_2\| < \nu_S(c_1)$, we have that $\nu_S(c_1) + \nu_S(c_2) \leq \|c_1 - c_2\| < \nu_S(c_1)$. This implies that $\nu_S(c_2) < 0$, which contradicts that ν_S is a non-negative distance function. Thus, we conclude that $w^*(c_1) = w^*(c_2)$. □

Proof of Lemma 2.3.2. Let $t \geq 0$ be given. First, we show that, for any given $x \in \mathcal{X}$, $\inf_{h^* \in \mathcal{H}^*}\{\nu_S(h^*(x))\} \geq 2b_t$ implies that $\nu_S(h_t(x)) \geq b_t$. Indeed, since ν_S is a 1-Lipschitz distance function, we have that $|\nu_S(h^*(x)) - \nu_S(h_t(x))| \leq \|h_t(x) - h^*(x)\|$ for all $h^* \in \mathcal{H}^*$. Since $\mathcal{H}_\ell^* \subseteq \mathcal{H}^*$ and $\text{Dist}_{\mathcal{H}_\ell^*}(h_t) \leq b_t$, we have that $\text{Dist}_{\mathcal{H}^*}(h_t) \leq \text{Dist}_{\mathcal{H}_\ell^*}(h_t) \leq b_t$. Hence, for any $\epsilon > 0$, there exists $h^* \in \mathcal{H}^*$ satisfying

$$\nu_S(h^*(x)) - \nu_S(h_t(x)) \leq \|h_t(x) - h^*(x)\| \leq \|h_t - h^*\|_\infty \leq b_t + \epsilon$$

Since the result holds for all $\epsilon > 0$, we conclude that $\nu_S(h_t(x)) \geq \inf_{h^* \in \mathcal{H}^*}\{\nu_S(h^*(x))\} - b_t$. Furthermore, since $\inf_{h^* \in \mathcal{H}^*}\{\nu_S(h^*(x))\} \geq 2b_t$, it holds that $\nu_S(h_t(x)) \geq 2b_t - b_t = b_t$.

According to Algorithm 1, a label for x_t is always acquired at iteration $t \geq 1$ if $\nu_S(h_{t-1}(x_t)) < b_{t-1}$. Otherwise, if $\nu_S(h_{t-1}(x_t)) \geq b_{t-1}$, then a label is acquired with probability $\tilde{p}$. Therefore, using the argument above, the label probability at iteration t is

$$\begin{aligned}\mathbb{P}(\text{acquire a label for } x_t) &= \mathbb{P}(\nu_S(h_{t-1}(x_t)) < b_{t-1}) + \tilde{p}\mathbb{P}(\nu_S(h_{t-1}(x_t)) \geq b_{t-1}) \\ &\leq \mathbb{P}(\nu_S(h_{t-1}(x_t)) < b_{t-1}) + \tilde{p} \\ &\leq \mathbb{P}\left(\inf_{h^* \in \mathcal{H}^*}\{\nu_S(h^*(x_t))\} < 2b_{t-1}\right) + \tilde{p} \;\leq\; \Psi(2b_{t-1}) + \tilde{p}\end{aligned}$$

Then, the expected number of acquired labels after T total iterations is at most $\sum_{t=1}^{T}\mathbb{P}(\text{acquire a label for } x_t) \leq \tilde{p}T + \sum_{t=1}^{T}\Psi(2b_{t-1})$. □

A.0.2 Proofs for Section 2.4

Proof of Proposition 2.4.1. Since the reweighted loss $\ell^{\texttt{rew}}(h; z)$ is upper bounded by $\frac{\omega_\ell(\hat{\mathcal{C}},\mathcal{C})}{\tilde{p}^{\mathbb{I}\{\tilde{p}>0\}}}$, Proposition 2.4.1 is an immediate result by Hoeffding's inequality and by taking the union bounds for all the predictors with $\mathcal{H}$. □

Proof of Lemma 2.4.1. Let $t \in \{1, \ldots, T\}$ be fixed. Since $\tilde{p} = 0$, the re-weighted loss function can be written as $\ell^{\texttt{rew}}(h; z_t) = \ell(h(x_t), c_t)d_t^M = \ell(h(x_t), c_t)\mathbb{I}\{\nu_S(h_{t-1}(x_t)) < b_{t-1}\}$, where h refers to a generic $h \in \mathcal{H}$ throughout. Recall that $\ell(h; z_t)$ denotes $\ell(h(x_t), c_t)$ and notice that we have the following simple decomposition:

$$\ell(h; z_t) \;=\; \ell(h; z_t)(1 - d_t^M) + \ell(h; z_t)d_t^M,$$

Hence, by the definition of $\ell_t^{\text{f}}(h)$, we have

$$\mathbb{E}[\ell(h; z_t)|\mathcal{F}_{t-1}] \;=\; \ell_t^{\text{f}}(h) + \mathbb{E}[\ell(h; z_t)d_t^M|\mathcal{F}_{t-1}] \;=\; \ell_t^{\text{f}}(h) + \mathbb{E}[\ell^{\texttt{rew}}(h; z_t)|\mathcal{F}_{t-1}].$$

Since $(x_1, c_t), \ldots, (x_T, c_T)$ are i.i.d. random variables following distribution $\mathcal{D}$, (x_t, c_t) is independent of $\mathcal{F}_{t-1}$ and hence

$$R_\ell(h) = \mathbb{E}[\ell(h; z_t)] = \mathbb{E}[\ell(h; z_t)|\mathcal{F}_{t-1}] = \ell_t^{\text{f}}(h) + \mathbb{E}[\ell^{\texttt{rew}}(h; z_t)|\mathcal{F}_{t-1}]. \tag{A.1}$$

Consider (A.1) applied to both $h \in \mathcal{H}$ and $h^* \in \mathcal{H}_\ell^*$ and averaged over $t \in \{1, \ldots, T\}$ to yield:

$$R_\ell(h) - R_\ell(h^*) \;=\; \frac{1}{T}\sum_{t=1}^{T}\left(\ell_t^{\text{f}}(h) - \ell_t^{\text{f}}(h^*)\right) + \frac{1}{T}\sum_{t=1}^{T}\left(\mathbb{E}[\ell^{\texttt{rew}}(h; z_t)|\mathcal{F}_{t-1}] - \mathbb{E}[\ell^{\texttt{rew}}(h^*; z_t)|\mathcal{F}_{t-1}]\right) \tag{A.2}$$

Thus, by the definition of $Z_h^{\texttt{t}}$, (A.2) is equivalently written as:

$$R_\ell(h) - R_\ell(h^*) \;=\; \frac{1}{T}\sum_{t=1}^{T}\left(\ell_t^{\text{f}}(h) - \ell_t^{\text{f}}(h^*)\right) + \frac{1}{T}\sum_{t=1}^{T}Z_h^{\texttt{t}} + \frac{1}{T}\sum_{t=1}^{T}\left(\ell^{\texttt{rew}}(h; z_t) - \ell^{\texttt{rew}}(h^*; z_t)\right). \tag{A.3}$$

□

Proof of Lemma 2.4.2. The proof is by strong induction. For the base case of $t = 1$, part *(a)* follows since $H_0 = \mathcal{H}$ and part *(b)* follows since $b_0 \geq \sqrt{r_0/\eta} \geq \sqrt{\omega_\ell(\hat{\mathcal{C}}, \mathcal{C})/\eta}$, and thus $\sup_{h \in H_0} \left\{ \left| \ell_1^{\mathrm{f}}(h) - \ell_1^{\mathrm{f}}(h^*) \right| \right\} \leq \omega_\ell(\hat{\mathcal{C}}, \mathcal{C}) \leq \eta b_0^2$.

Now, consider $t \geq 2$ and assume that parts *(a)* and *(b)* hold for all $\tilde{t} \in \{1, \ldots, t-1\}$. Namely, for all $\tilde{t} \in \{1, \ldots, t-1\}$, the following two conditions hold: *(a)* $h^* \in H_{\tilde{t}-1}$, and *(b)* $\sup_{h \in H_{\tilde{t}-1}} \left\{ \left| \ell_{\tilde{t}}^{\mathrm{f}}(h) - \ell_{\tilde{t}}^{\mathrm{f}}(h^*) \right| \right\} \leq \eta b_{\tilde{t}-1}^2$. Then, our goal is to show that the two claims hold for t.

First, we prove *(a)*. Recall that h^* denotes the unique minimizer of the surrogate risk R_ℓ, and h_{t-1} denotes the predictor from iteration $t-1$ of Algorithm 1. By Lemma 2.4.1, we have that

$$R_\ell(h_{t-1}) - R_\ell(h^*) = \frac{1}{t-1} \sum_{i=1}^{t-1} \left(\ell_i^{\mathrm{f}}(h_{t-1}) - \ell_i^{\mathrm{f}}(h^*) \right) + \frac{1}{t-1} \sum_{i=1}^{t-1} Z_{h_{t-1}}^{\mathtt{i}} + \frac{1}{t-1} \sum_{i=1}^{t-1} \left(\ell^{\mathtt{rew}}(h_{t-1}; z_i) - \ell^{\mathtt{rew}}(h^*; z_i) \right).$$

Since $R_\ell(h_{t-1}) - R_\ell(h^*) \geq 0$, we have that

$$\begin{aligned} \frac{1}{t-1} \sum_{i=1}^{t-1} (\ell^{\mathtt{rew}}(h^*; z_i) - \ell^{\mathtt{rew}}(h_{t-1}; z_i)) &\leq \frac{1}{t-1} \sum_{i=1}^{t-1} \left(\ell_i^{\mathrm{f}}(h_{t-1}) - \ell_i^{\mathrm{f}}(h^*) \right) + \frac{1}{t-1} \sum_{i=1}^{t-1} Z_{h_{t-1}}^{\mathtt{i}} \\ &\leq \frac{1}{t-1} \sum_{i=1}^{t-1} \sup_{h \in H_{i-1}} \left\{ \left| \ell_i^{\mathrm{f}}(h) - \ell_i^{\mathrm{f}}(h^*) \right| \right\} + \frac{1}{t-1} \sum_{i=1}^{t-1} Z_{h_{t-1}}^{\mathtt{i}} \\ &\leq \frac{1}{t-1} \sum_{i=1}^{t-1} \eta b_{i-1}^2 \ + \ r_{t-1}, \end{aligned}$$

where the second inequality uses $h_{t-1} \in H_{t-2} \subseteq H_{i-1}$ for $i \in \{1, \ldots, t-1\}$, and the third inequality uses assumption *(b)* of induction and the assumption that $r_t \geq \sup_{h \in \mathcal{H}} \left\{ \left| \frac{1}{t} \sum_{i=1}^{t} Z_h^{\mathtt{i}} \right| \right\}$ for $t \geq 1$.

Recall from Algorithm 1 that the reweighted loss function at iteration $t-1$ is $\hat{\ell}^{t-1}(h) = \frac{1}{t} \sum_{(x,c) \in W_{t-1}} \ell(h(x), c) = \frac{1}{t-1} \sum_{i=1}^{t-1} \ell^{\mathtt{rew}}(h; z_i)$, and h_{t-1} is the corresponding minimizer over H_{t-2} hence $\frac{1}{t-1} \sum_{i=1}^{t-1} \ell^{\mathtt{rew}}(h_{t-1}; z_i) = \hat{\ell}^{t-1,*}$. By assumption *(a)* of induction, we have that $h^* \in H_{t-2}$. The above chain of inequalities shows that $\hat{\ell}^{t-1}(h^*) \leq \hat{\ell}^{t-1,*} + \frac{1}{t-1} \sum_{i=1}^{t-1} \eta b_{i-1}^2 + r_{t-1}$, hence $h^* \in H_{t-1}$ by definition in Line 20 of Algorithm 1.

Next, we prove *(b)* for t. Let $h \in H_{t-1}$ be fixed. By Assumption 2.3.1.(2) and since h^* is the unique minimizer in $\mathcal{H}_\ell^*$, we have that $\|h - h^*\|_\infty \leq \phi(R_\ell(h) - R_\ell^*)$. By Assumption

2.4.1, we then have that

$$\begin{aligned}
|\ell_t^{\mathrm{f}}(h) - \ell_t^{\mathrm{f}}(h^*)| &= \left|\mathbb{E}[\ell(h(x_t), c_t) - \ell(h^*(x_t), c_t)|d_t^M = 0]\mathbb{P}(d_t^M = 0|\mathcal{F}_{t-1})\right| \\
&\leq \left|\mathbb{E}[\ell(h(x_t), c_t) - \ell(h^*(x_t), c_t)|d_t^M = 0]\right| \\
&\leq \left|\mathbb{E}[\mathbb{E}[\ell(h(x_t), c_t) - \ell(h^*(x_t), c_t)|x_t]|d_t^M = 0]\right| \\
&\leq \eta\mathbb{E}[\|h(x_t) - h^*(x_t)\|^2|d_t^M = 0] \\
&\leq \eta\left(\phi(R_\ell(h) - R_\ell^*)\right)^2. \qquad \text{(A.4)}
\end{aligned}$$

By Lemma 2.4.1, we have that

$$\begin{aligned}
R_\ell(h) - R_\ell(h^*) &= \frac{1}{t-1}\sum_{i=1}^{t-1}\left(\ell_i^{\mathrm{f}}(h) - \ell_i^{\mathrm{f}}(h^*)\right) + \frac{1}{t-1}\sum_{i=1}^{t-1} Z_h^{\mathtt{i}} + \frac{1}{t-1}\sum_{i=1}^{t-1}\left(\ell^{\mathtt{rew}}(h; z_i) - \ell^{\mathtt{rew}}(h^*; z_i)\right) \\
&\leq \frac{1}{t-1}\sum_{i=1}^{t-1}\eta b_{i-1}^2 + r_{t-1} + \frac{1}{t-1}\sum_{i=1}^{t-1}\left(\ell^{\mathtt{rew}}(h; z_i) - \ell^{\mathtt{rew}}(h^*; z_i)\right) \\
&= \frac{1}{t-1}\sum_{i=1}^{t-1}\eta b_{i-1}^2 + r_{t-1} + \hat{\ell}^{t-1}(h) - \hat{\ell}^{t-1}(h^*), \qquad \text{(A.5)}
\end{aligned}$$

where the inequality follows by assumption *(b)* of induction since $h \in H_{t-1} \subseteq H_{i-1}$ for $i \in \{1, \ldots, t-1\}$ and the assumption that $r_t \geq \sup_{h\in\mathcal{H}}\left\{\left|\frac{1}{t}\sum_{i=1}^t Z_h^{\mathtt{i}}\right|\right\}$ for $t \in \{1, \ldots, T\}$, and the equality follows by the definition of the reweighted loss function in Algorithm 1. By assumption we have that $h \in H_{t-1}$ and by the proof of part *(a)*, we have that $h^* \in H_{t-1}$. Thus, since $\hat{\ell}^{t-1,*} = \min_{h\in H_{t-2}}\hat{\ell}^{t-1}(h)$ and $H_{t-1} \subseteq H_{t-2}$, we have that

$$\hat{\ell}^{t-1,*} \leq \hat{\ell}^{t-1}(h) \leq \hat{\ell}^{t-1,*} + r_{t-1} + \frac{1}{t-1}\sum_{i=1}^{t-1}\eta b_{i-1}^2,$$

$$\text{and} \quad \hat{\ell}^{t-1,*} \leq \hat{\ell}^{t-1}(h^*) \leq \hat{\ell}^{t-1,*} + r_{t-1} + \frac{1}{t-1}\sum_{i=1}^{t-1}\eta b_{i-1}^2,$$

hence $\hat{\ell}^{t-1}(h) - \hat{\ell}^{t-1}(h^*) \leq r_{t-1} + \frac{1}{t-1}\sum_{i=1}^{t-1}\eta b_{i-1}^2$ and by combining with (A.5) we have

$$R_\ell(h) - R_\ell(h^*) \;\leq\; \frac{2\eta}{t-1}\sum_{i=1}^{t-1} b_{i-1}^2 + 2r_{t-1}.$$

Combining the above inequality with (A.4) yields

$$|\ell_t^{\mathrm{f}}(h) - \ell_t^{\mathrm{f}}(h^*)| \;\leq\; \eta\left(\phi\left(\frac{2\eta}{t-1}\sum_{i=1}^{t-1} b_{i-1}^2 + 2r_{t-1}\right)\right)^2 \;=\; \eta b_{t-1}^2,$$

using the definition of b_{t-1}. Since $h \in H_{t-1}$ is arbitrary, the conclusion in part *(b)* follows. □

Proof of Theorem 2.4.1. We provide the proof of each part separately.

Part *(a)*. Recall that h^* is the unique minimizer of the surrogate risk R_ℓ under Assumption 2.4.1 and that h_T is the predictor from iteration T of Algorithm 1. By Lemma 2.4.1, we have the following decomposition:

$$\begin{aligned} R_\ell(h_T) - R_\ell(h^*) &= \frac{1}{T}\sum_{t=1}^{T}\left(\ell_t^{\mathrm{f}}(h_T) - \ell_t^{\mathrm{f}}(h^*)\right) + \frac{1}{T}\sum_{t=1}^{T} Z_{h_T}^{\mathtt{t}} + \frac{1}{T}\sum_{t=1}^{T}\left(\ell^{\mathtt{rew}}(h_T; z_t) - \ell^{\mathtt{rew}}(h^*; z_t)\right) \\ &= \frac{1}{T}\sum_{t=1}^{T}\left(\ell_t^{\mathrm{f}}(h_T) - \ell_t^{\mathrm{f}}(h^*)\right) + \frac{1}{T}\sum_{t=1}^{T} Z_{h_T}^{\mathtt{t}} + \hat{\ell}^T(h_T) - \hat{\ell}^T(h^*), \end{aligned} \tag{A.6}$$

where we recall that the empirical re-weighted loss in Algorithm 1 is $\hat{\ell}^T(h) := \frac{1}{T}\sum_{(x,c)\in W_T} \ell(h(x), c) = \frac{1}{T}\sum_{t=1}^{T} \ell^{\mathtt{rew}}(h; z_t)$ in this case (since $\tilde{p} = 0$).

We will show that $r_t \geq \sup_{h\in\mathcal{H}}\left\{\left|\frac{1}{t}\sum_{i=1}^{t} Z_h^{\mathtt{i}}\right|\right\}$ simultaneously for all $t \geq 1$ with probability at least $1-\delta$ in order to apply Lemma 2.4.2, again with probability at least $1-\delta$. Indeed, suppose that the conclusions of Lemma 2.4.2 do hold. Then, by part *(a)*, we have $h^* \in H_{T-1}$ and therefore $\hat{\ell}^T(h_T) \leq \hat{\ell}^T(h^*)$ by the update in Line 19 of Algorithm 1. By the nested structure of the H_t sets, we have $h_T \in H_{T-1} \subseteq H_{t-1}$ for all $t \in \{1, \ldots, T\}$ and therefore, by part *(b)*, we have that $|\ell_t^{\mathrm{f}}(h_T) - \ell_t^{\mathrm{f}}(h^*)| \leq \eta b_{t-1}^2$. Thus, combining these inequalities with (A.6) yields:

$$R_\ell(h_T) - R_\ell(h^*) \;\leq\; r_T + \frac{1}{T}\sum_{t=0}^{T-1}\eta b_t^2,$$

which is the result in part *(a)*.

It remains to show that $r_T \geq \sup_{h\in\mathcal{H}}\left\{\left|\frac{1}{T}\sum_{t=1}^{T} Z_h^{\mathtt{t}}\right|\right\}$ simultaneously for all $T \geq 1$ with probability at least $1-\delta$. For each $T \geq 1$, we apply Proposition 2.4.1 and plug in both $h, h^* \in \mathcal{H}$. Indeed, by considering the two sequences $\{\mathbb{E}[\ell^{\mathtt{rew}}(h; z_t)|\mathcal{F}_{t-1}] - \ell^{\mathtt{rew}}(h; z_t)\}$ and $\{\mathbb{E}[\ell^{\mathtt{rew}}(h^*; z_t)|\mathcal{F}_{t-1}] - \ell^{\mathtt{rew}}(h^*; z_t)\}$ and their differences, we have the following bound for any $\epsilon > 0$ with probability at least $1 - 2N_1 \exp\left(-\frac{2T\epsilon^2}{\omega_\ell^2(\hat{\mathcal{C}},\mathcal{C})}\right)$:

$$\sup_{h\in\mathcal{H}}\left\{\left|\frac{1}{T}\sum_{t=1}^{T} Z_h^{\mathtt{t}}\right|\right\} \leq 2\epsilon.$$

Considering $\epsilon = \omega_\ell(\hat{\mathcal{C}},\mathcal{C})\sqrt{\frac{\ln(2TN_1/\delta)}{T}} = r_T/2$ yields that the above bound holds with probability at least $1 - \frac{\delta^2}{2T^2N_1} > 1 - \frac{\delta}{2T^2}$. Finally, applying the union bound over all $T \geq 1$, we obtain that $r_T \geq \sup_{h\in\mathcal{H}}\left\{\left|\frac{1}{T}\sum_{t=1}^{T} Z_h^{\mathtt{t}}\right|\right\}$ simultaneously for all $T \geq 1$ with probability at least

$$1 - \frac{\delta}{2}\sum_{T=1}^{\infty}\frac{1}{T^2} = 1 - \frac{\delta\pi^2}{12} > 1 - \delta.$$

Part *(b)*. In this part of the proof, we do not assume that $\mathcal{H}_\ell^*$ is a singleton as the same proof will apply later for Theorem 2.4.3. For any $h^* \in \mathcal{H}^*$, the excess SPO risk can be written as

$$\begin{aligned} R_{\text{SPO}}(h_T) - R_{\text{SPO}}^* &= \mathbb{E}_{(x,c)\sim\mathcal{D}}[c^T(w^*(h_T(x)) - w^*(h^*(x)))] \\ &\leq \mathbb{E}_{(x,c)\sim\mathcal{D}}[\|c\|\|w^*(h_T(x)) - w^*(h^*(x))\|_*]. \end{aligned} \tag{A.7}$$

We apply the conclusion of part *(a)*, $R_\ell(h_T) - R_\ell^* \leq r_T + \frac{\eta}{T}\sum_{t=0}^{T-1} b_t^2$, with probability at least $1-\delta$. Specifically, we shall provide a bound on $\Delta_T^{\text{SPOa}} := \mathbb{E}[R_{\text{SPO}}(h_T) - R_{\text{SPO}}^* | R_\ell(h_T) - R_\ell^* \leq r_T + \frac{\eta}{T}\sum_{t=0}^{T-1} b_t^2]$. If $b_T = 0$, then $\phi(r_T) = 0$ and Assumption 2.3.1.(2) implies that $h_T(x) = h^*(x)$ almost everywhere and hence $\Delta_T^{\text{SPOa}} = 0$. Thus, part *(b)* follows immediately. Otherwise, assume that $b_T > 0$. Recall that, by Assumption 2.3.1.(1), we have that $\text{Dist}_{\mathcal{H}^*}(h_T) \leq \text{Dist}_{\mathcal{H}_\ell^*}(h_T)$. Then, by combining Assumption 2.3.1.(2) with part *(a)*, with probability at least $1-\delta$, for any $\epsilon > 0$ there exists $h^* \in \mathcal{H}^*$ such that for almost every $x \in \mathcal{X}$,

$$\|h_T(x) - h^*(x)\| \;\leq\; \phi\left(r_T + \frac{\eta}{T}\sum_{t=0}^{T-1} b_t^2\right) + \epsilon \;\leq\; \phi\left(2r_T + \frac{2\eta}{T}\sum_{t=0}^{T-1} b_t^2\right) + \epsilon \;=\; b_T + \epsilon. \tag{A.8}$$

Since S satisfies the strength property with parameter $\mu > 0$, we apply part *(a)* of Theorem 3 of El Balghiti et al., 2022 to yield (in our notation) for any $h^* \in \mathcal{H}^*$ and $x \in \mathcal{X}$:

$$\|w^*(h_T(x)) - w^*(h^*(x))\|_* \;\leq\; \left(\frac{1}{\mu \min\{\nu_S(h_T(x)), \nu_S(h^*(x))\}}\right)\|h_T(x) - h^*(x)\|. \tag{A.9}$$

Now, let $\gamma_T \geq 2b_T > 0$ be a given parameter. For a given $x \in \mathcal{X}$ we consider two cases: *(i)* $\nu_S(h^*(x)) > \gamma_T$, and *(ii)* $\nu_S(h^*(x)) \leq \gamma_T$. Under case *(i)*, we have by the 1-Lipschitzness of $\nu_S(\cdot)$ that $\nu_S(h_T(x)) \geq \nu_S(h^*(x)) - \|h_T(x) - h^*(x)\| > \gamma_T - b_T \geq \gamma_T - \gamma_T/2 = \gamma_T/2$. Thus, we have that $\min\{\nu_S(h_T(x)), \nu_S(h^*(x))\} \geq \gamma_T/2$. For case *(i)* we will combine together (A.7), (A.8), and (A.9), and use the fact that $\|c\| \leq \rho(\mathcal{C})$. For case *(ii)*, we apply the worst case bound $R_{\text{SPO}}(h_T) - R_{\text{SPO}}^* \leq \omega_S(\mathcal{C})$ and note that the probability of case *(ii)* occurring is at most $\mathbb{P}\left(\inf_{h^*\in\mathcal{H}^*}\{\nu_S(h^*(x))\} \leq \gamma_T\right) = \Psi(\gamma_T)$. Overall, we have

$$\Delta_T^{\text{SPOa}} \;\leq\; \rho(\mathcal{C})\|w^*(h_T(x)) - w^*(h^*(x))\|_* + \Psi(\gamma_T)\omega_S(\mathcal{C}) \;\leq\; \frac{2\rho(\mathcal{C})(b_T+\epsilon)}{\mu\gamma_T} + \Psi(\gamma_T)\omega_S(\mathcal{C}).$$

We take $\epsilon \to 0$, and since $\gamma_T \geq b_T$ is arbitrary we take the infimum over γ_T to yield part *(b)*.

Part *(c)*. Again, in this part of the proof, we do not assume that $\mathcal{H}_\ell^*$ is a singleton as the same proof will apply later for Theorems 2.4.2 and 2.4.3. Recall that for any $h^* \in \mathcal{H}^*$, the excess SPO risk can be written as

$$R_{\text{SPO}}(h_T) - R_{\text{SPO}}^* \;=\; \mathbb{E}_{(x,c)\sim\mathcal{D}}[c^T(w^*(h_T(x)) - w^*(h^*(x)))].$$

Again, we apply part *(a)* with probability at least $1-\delta$. Recall that, by Assumption 2.3.1.(1), we have that $\text{Dist}_{\mathcal{H}^*}(h_T) \leq \text{Dist}_{\mathcal{H}^*_\ell}(h_T)$. Then, by combining Assumption 2.3.1.(2) with part *(a)*, with probability at least $1-\delta$, for any $\epsilon > 0$ there exists $h^* \in \mathcal{H}^*$ such that for almost every $x \in \mathcal{X}$,

$$\|h_T(x) - h^*(x)\| \;\leq\; \phi\left(r_T + \frac{\eta}{T}\sum_{t=0}^{T-1} b_t^2\right) + \epsilon \;\leq\; \phi\left(2r_T + \frac{2\eta}{T}\sum_{t=0}^{T-1} b_t^2\right) + \epsilon \;=\; b_T + \epsilon. \quad \text{(A.10)}$$

For a given $x \in \mathcal{X}$ we consider two cases: *(i)* $\nu_S(h^*(x)) \geq 2b_T$, and *(ii)* $\nu_S(h^*(x)) < 2b_T$. Under case *(i)*, we have that $\max\{\nu_S(h_T(x)), \nu_S(h^*(x))\} \geq 2b_T > b_T + \epsilon$ for $\epsilon < b_T$; thus combining Lemma 2.3.1 and (A.10) yields that $w^*(h_T(x)) = w^*(h^*(x))$, and hence $R_{\text{SPO}}(h_T) - R^*_{\text{SPO}} = 0$, for almost every $x \in \mathcal{X}$ under case *(i)*. For case *(ii)*, we also apply the worst case bound $R_{\text{SPO}}(h_T) - R^*_{\text{SPO}} \leq \omega_S(\mathcal{C})$ and note that the probability of case *(ii)* occurring is at most $\mathbb{P}\left(\inf_{h^* \in \mathcal{H}^*}\{\nu_S(h^*(x))\} < 2b_T\right) \leq \Psi(2b_T)$. Therefore, overall we have with probability at least $1-\delta$,

$$R_{\text{SPO}}(h_T) - R^*_{\text{SPO}} \leq \Psi(2b_T)\omega_S(\mathcal{C}).$$

Part *(d)*. First note that, by Assumption 2.3.1.(2), we have that $\text{Dist}_{\mathcal{H}^*_\ell}(h_0) \leq \phi(R_\ell(h_0) - R^*_\ell) \leq \phi(\omega_\ell(\hat{\mathcal{C}}, \mathcal{C})) \leq \phi(r_0) \leq b_0$. Again, we apply part *(a)* with probability at least $1-\delta$. Indeed, when $R_\ell(h_T) - R^*_\ell \leq r_T + \frac{\eta}{T}\sum_{t=0}^{T-1} b_t^2$ holds, by Assumption 2.3.1.(2), we have that $\text{Dist}_{\mathcal{H}^*_\ell}(h_T) \leq \phi(R_\ell(h_T) - R^*_\ell) \leq \phi(r_T + \frac{\eta}{T}\sum_{t=0}^{T-1} b_t^2) \leq b_T$. Thus, part *(a)* implies $\text{Dist}_{\mathcal{H}^*_\ell}(h_t) \leq b_t$ holds simultaneously for all $t \geq 0$ with probability at least $1-\delta$. By Lemma 2.3.2, since $\tilde{p} = 0$, conditional on part *(a)*, the label complexity is at most $\sum_{t=1}^{T} \Psi(2b_{t-1})$. With probability at most δ, we consider the worst case label complexity T and hence arrive at the overall label complexity bound of $\sum_{t=1}^{T} \Psi(2b_{t-1}) + \delta T$.

□

Proof of Proposition 2.4.2. Let $\bar{h} \in \mathcal{H}$ satisfy the conditions in Assumption 2.4.2. We will show that $R_{\text{SPO+}}(\bar{h}) = 0$ and therefore, since $0 \leq R_{\text{SPO}}(h) \leq R_{\text{SPO+}}(h)$ for all $h \in \mathcal{H}$, we have that $R^*_{\text{SPO+}} = R^*_{\text{SPO}} = 0$ and $\bar{h}$ is a minimizer for both.

Recall that for prediction $\hat{c} \in \mathbb{R}^d$ and realized cost vector $c \in \mathbb{R}^d$, the SPO+ satisfies as

$$\begin{aligned}
\ell_{\text{SPO+}}(\hat{c}, c) &:= \max_{w \in S}\left\{(c - 2\hat{c})^T w\right\} + 2\hat{c}^T w^*(c) - c^T w^*(c) \\
&= -\min_{w \in S}\left\{(2\hat{c} - c)^T w\right\} + 2\hat{c}^T w^*(c) - c^T w^*(c) \\
&= (c - 2\hat{c})^T w^*(2\hat{c} - c) + 2\hat{c}^T w^*(c) - c^T w^*(c) \\
&= 2\hat{c}^T(w^*(c) - w^*(2\hat{c} - c)) + c^T(w^*(2\hat{c} - c) - w^*(c)).
\end{aligned}$$

Under Assumption 2.4.2 in the polyhedral case, we have by Lemma 2.3.1 that $w^*(\bar{h}(x)) = w^*(c)$ with probability one over $(x, c) \sim \mathcal{D}$. Similarly, we have that $\|(2\bar{h}(x) - c) - \bar{h}(x)\| = \|\bar{h}(x) - c\| \leq \varrho\nu_S(\bar{h}(x)) < \nu_S(\bar{h}(x))$, and hence $w^*(2\bar{h}(x) - c) = w^*(\bar{h}(x)) = w^*(c)$, with probability

one over $(x, c) \sim \mathcal{D}$. Therefore, we have that $R_{\text{SPO+}}(\bar{h}) = \mathbb{E}_{(x,c)\sim\mathcal{D}}[\ell_{\text{SPO+}}(\bar{h}(x), c)] = 0$ by the above expression for $\ell_{\text{SPO+}}(\hat{c}, c)$.

□

Proof of Theorem 2.4.2. We provide the proof of part *(a)*, as the proofs of parts *(b)* and *(c)* are completely analogous to Theorem 2.4.1.

Recall that $(x_1, c_1), (x_2, c_2), \ldots$ is the sequence of features and corresponding cost vectors of Algorithm 1. It is assumed that this sequence is an i.i.d. sequence from the distribution $\mathcal{D}$ and note that c_t is only observed when we do not reject x_t, i.e., when $d_t^M = \mathbb{I}(\nu_S(h_{t-1}(x_t)) < b_{t-1}) = 1$. In a slight abuse of notation, in this proof only, let us define $Z_h^{\mathfrak{t}} := \mathbb{E}[\ell_{\text{SPO+}}(h(x_t), c_t)] - \ell_{\text{SPO+}}(h(x_t), c_t) = R_{\text{SPO+}}(h) - \ell_{\text{SPO+}}(h(x_t), c_t)$. Following the template of Lemma 2.4.2, the main idea of the proof is to show that, when $r_t \geq \sup_{h\in\mathcal{H}} \left\{\left|\frac{1}{t}\sum_{i=1}^{t} Z_h^{\mathfrak{i}}\right|\right\}$ for all $t \geq 1$, we have that:

$$(A) \quad \max_{t\in\{1,\ldots,T\}} \{\ell_{\text{SPO+}}(h_T(x_t), c_t)\} = 0 \quad \text{with probability 1 for all } T \geq 1.$$

In other words, h_T achieves zero SPO+ loss across the entire sequence $(x_1, c_1), \ldots, (x_T, c_T)$. In fact, we show a strong result, which is that *(A)* holds for *all* minimizers of the empirical reweighted loss at iteration T. The proof of this is by strong induction, and we defer it to the end.

Notice that $\max_{t\in\{1,\ldots,T\}} \{\ell_{\text{SPO+}}(h_T(x_t), c_t)\} = 0$ implies, of course, that h_T achieves zero (and hence minimizes) empirical risk $\frac{1}{T}\sum_{t=1}^{T} \ell_{\text{SPO+}}(h(x_t), c_t)$.

Thus, using $r_t \geq \sup_{h\in\mathcal{H}} \left\{\left|\frac{1}{t}\sum_{i=1}^{t} Z_h^{\mathfrak{i}}\right|\right\}$ for all $t \geq 1$, we have that

$$R_{\text{SPO+}}(h_T) = R_{\text{SPO+}}(h_T) - \frac{1}{T}\sum_{t=1}^{T} \ell_{\text{SPO+}}(h_T(x_t), c_t) = \frac{1}{T}\sum_{t=1}^{T} Z_{h_T}^{\mathfrak{t}} \leq r_T,$$

which is the result of part *(a)*. To control the probability that $r_t \geq \sup_{h\in\mathcal{H}} \left\{\left|\frac{1}{t}\sum_{i=1}^{t} Z_h^{\mathfrak{i}}\right|\right\}$, we observe that there are at most N_1 candidate predictors. As pointed out by Kuznetsov and Mohri, 2015, in the i.i.d. case, for $T \geq 1$ and for any $\epsilon > 0$, we have the same convergence result as Proposition 2.4.1.

In Proposition 2.4.1, considering $\epsilon = \omega_\ell(\hat{\mathcal{C}}, \mathcal{C})\sqrt{\frac{\ln(2TN_1/\delta)}{T}} = r_T$ and following the same reasoning as in the proof of Theorem 2.4.1 yields that $r_T \geq \sup_{h\in\mathcal{H}} \left\{\left|\frac{1}{T}\sum_{t=1}^{T} Z_h^{\mathfrak{t}}\right|\right\}$ simultaneously for all $T \geq 1$ with probability at least $1 - \delta$.

Proof of Claim *(A)*. It remains to show that *(A)* holds for all $T \geq 1$, which we prove by strong induction. In fact, we prove a stronger variant of *(A)* as follows. Recall that $\hat{\ell}^T(h) = \frac{1}{T}\sum_{(x,c)\in W_T} \ell_{\text{SPO+}}(h(x), c)$ is the empirical reweighed loss at iteration T. Define $H_T^0 := \{h \in \mathcal{H} : \hat{\ell}^T(h) = 0\} = \{h \in \mathcal{H} : \ell_{\text{SPO+}}(h(x), c) = 0 \text{ for all } (x, c) \in W_T\}$. The set H_T^0 is exactly the set of minimizers of $\hat{\ell}^T(h)$, with probability 1, since Proposition 2.4.2 (in

particular $R^*_{\text{SPO+}} = 0$) implies that $\hat{\ell}^T(h^*) = 0$ with probability 1. Hence, $h_T \in H^0_T$ with probability 1.

Let us also define $\bar{H}^0_T := \{h \in \mathcal{H} : \ell_{\text{SPO+}}(h(x_t), c_t) = 0 \text{ for all } t = 1, \ldots, T\}$. Clearly, $\bar{H}^0_T \subseteq H^0_T$ for all $T \geq 1$. Note also that both collections of sets are nested, i.e., $H^0_T \subseteq H^0_{T-1} \subseteq \cdots H^0_1 \subseteq \mathcal{H}$ and $\bar{H}^0_T \subseteq \bar{H}^0_{T-1} \subseteq \cdots \bar{H}^0_1 \subseteq \mathcal{H}$. Now, we will use strong induction to prove, when $r_t \geq \sup_{h\in\mathcal{H}} \left\{\left|\frac{1}{t}\sum_{i=1}^t Z^{\mathtt{i}}_h\right|\right\}$ for all $t \geq 1$, we have that:

$$(\bar{A}) \quad H^0_T = \bar{H}^0_T \text{ with probability 1 for all } T \geq 1.$$

Note that *(Ā)* implies *(A)* since $h_T \in H^0_T$ with probability 1.

To prove the base case $T = 1$, we observe that $b_0 \geq \rho(\hat{\mathcal{C}})$, and thus, $\nu_S(h(x_1)) \leq \|h(x_1)\| \leq \rho(\hat{\mathcal{C}}) \leq b_0$ for any $h \in \mathcal{H}$ and any $x_1 \in \mathcal{X}$. Thus, we have that $d^M_1 = 1$ with probability 1 and the sample (x_1, c_1) is added to working set W_1. By definition of H^0_1, for $h \in H^0_1$ we have that $\ell_{\text{SPO+}}(h(x_1), c_1) = 0$. Hence, $h \in \bar{H}^0_1$ and so we have proven that $H^0_1 \subseteq \bar{H}^0_1$.

Now, consider $T \geq 2$ and assume that *(Ā)* holds for all $\tilde{T} \in \{1, \ldots, T-1\}$. We need to show that $H^0_T \subseteq \bar{H}^0_T$, so let $h \in H^0_T$ be given. By the induction hypothesis, we have that $h \in H^0_{T-1} = \bar{H}^0_{T-1}$, and therefore we have $\ell_{\text{SPO+}}(h(x_t), c_t) = 0$ for all $t \in \{1, \ldots, T-1\}$. Thus, to show that $h \in \bar{H}^0_T$, it suffices to show that $\ell_{\text{SPO+}}(h(x_T), c_T) = 0$ with probability 1.

There are two cases to consider. First, if $d^M_T = 1$, then the sample (x_T, c_T) is added to working set W_T and thus, by definition of H^0_T, for $h \in H^0_T$ we have that $\ell_{\text{SPO+}}(h(x_T), c_T) = 0$. Hence, $h \in \bar{H}^0_T$ and so we have proven that $H^0_T \subseteq \bar{H}^0_T$.

Second, let us consider the case where $d^M_T = 0$ so we do not acquire the label c_T. In this case, we have that $W_T = W_{T-1}$, $H^0_T = H^0_{T-1}$, and, by the rejection criterion, $\nu_S(h_{T-1}(x_T)) \geq b_{T-1}$. For the given $h \in H^0_T$, to show that $\ell_{\text{SPO+}}(h(x_T), c_T) = 0$ with probability 1 recall from the proof of Proposition 2.4.2 that is suffices to show that $w^*(2h(x_T) - c_T) = w^*(c_T)$ with probability 1 over c_T drawn from the conditional distribution given x_T.

To prove this, first note that

$$R_{\text{SPO+}}(h) = R_{\text{SPO+}}(h) - \frac{1}{T-1}\sum_{t=1}^{T-1} \ell_{\text{SPO+}}(h(x_t), c_t) = \frac{1}{T-1}\sum_{t=1}^{T-1} Z^{\mathtt{t}}_h \leq r_{T-1},$$

where the first equality uses that $h \in H^0_T = H^0_{T-1} = \bar{H}^0_{T-1}$ and the inequality uses the assumption that $r_t \geq \sup_{h\in\mathcal{H}} \left\{\left|\frac{1}{t}\sum_{i=1}^t Z^{\mathtt{i}}_h\right|\right\}$ for all $t \geq 1$. By similar reasoning, we have that $h_{T-1} \in H^0_{T-1} = \bar{H}^0_{T-1}$ satisfies $R_{\text{SPO+}}(h_{T-1}) \leq r_{T-1}$. Let $\epsilon > 0$ be fixed. Now, by Assumption 2.3.1 and Proposition 2.4.2, there exists $h^*_0 \in \mathcal{H}^*_{\text{SPO+}}$ such that

$$\|h(x_T) - h^*_0(x_T)\| \leq \phi(R_{\text{SPO+}}(h)) + \epsilon \leq \phi(r_{T-1}) + \epsilon \leq \frac{\tau(1-\varrho)}{\tau(1-\varrho)+2} b_{T-1} + \epsilon,$$

and there exists $h^*_1 \in \mathcal{H}^*_{\text{SPO+}}$ such that

$$\|h_{T-1}(x_T) - h^*_1(x_T)\| \leq \phi(R_{\text{SPO+}}(h_{T-1})) + \epsilon \leq \phi(r_{T-1}) + \epsilon \leq \frac{\tau(1-\varrho)}{\tau(1-\varrho)+2} b_{T-1} + \epsilon,$$

where we have used $b_{T-1} = (1 + \frac{2}{\tau(1-\rho)})\phi(r_{T-1})$ in both inequalities above. Since both $h_0^*, h_1^* \in \mathcal{H}^*_{\text{SPO+}}$, according to the proof of Proposition 2.4.2, we have that $w^*(2h_0^*(x_T) - c_T) = w^*(c_T) = w^*(2h_1^*(x_T) - c_T)$ with probability 1. By the rejection criterion, $\nu_S(h_{T-1}(x_T)) \geq b_{T-1}$, and the 1-Lipschitzness of ν_S, we have

$$\begin{aligned}\nu_S(h_1^*(x_T)) &= \nu_S(h_1^*(x_T) - h_{T-1}(x_T) + h_{T-1}(x_T)) \geq \nu_S(h_{T-1}(x_T)) - \|h_1^*(x_T) - h_{T-1}(x_T)\| - \epsilon \\ &\geq b_{T-1} - \frac{\tau(1-\varrho)}{\tau(1-\varrho)+2} b_{T-1} - \epsilon = \frac{2}{\tau(1-\varrho)+2} b_{T-1} - \epsilon.\end{aligned}$$

By the second part of Assumption 2.4.2, we have that

$$\nu_S(h_0^*(x_T)) \geq \tau \left(\sup_{h' \in \mathcal{H}^*_{\text{SPO+}}} \{\nu_S(h'(x_T))\} \right) \geq \tau \nu_S(h_1^*(x_T)) \geq \frac{2\tau}{\tau(1-\varrho)+2} b_{T-1} - \epsilon\tau.$$

By viewing $2h(x_T) - c_T$ and $2h_0^*(x_T) - c_T$ as c_1 and c_2 in Lemma 2.3.1, we have

$$\|(2h(x_T) - c_T) - (2h_0^*(x_T) - c_T)\| = 2\|h(x_T) - h_0^*(x_T)\| \leq \frac{2\tau(1-\varrho)}{\tau(1-\varrho)+2} b_{T-1} + 2\epsilon.$$

By the 1-Lipschitzness of ν_S and the first part of Assumption 2.4.2, we have

$$\begin{aligned}\nu_S(2h_0^*(x_T) - c_T) &\geq \nu_S(h_0^*(x_T)) - \|h_0^*(x_T) - c_T\| \geq (1-\varrho)\nu_S(h_0^*(x_T)) \\ &\geq (1-\varrho)\left(\frac{2\tau}{\tau(1-\varrho)+2} b_{T-1} - \epsilon\tau \right) = \frac{2\tau(1-\varrho)}{\tau(1-\varrho)+2} b_{T-1} - (1-\varrho)\epsilon\tau.\end{aligned}$$

By taking $\epsilon \to 0$ and considering an appropriate convergent subsequence in the compact set $\mathcal{H}^*_{\text{SPO+}}$, the two inequalities above are satisfied for some $\bar{h}_0^* \in \mathcal{H}^*_{\text{SPO+}}$ with $\epsilon = 0$. In particular, this implies that the conditions in Lemma 2.3.1 are satisfied and we have that $w^*(2h(x_T) - c_T) = w^*(2\bar{h}_0^*(x_T) - c_T) = w^*(c_T)$ with probability 1. Hence, we have shown that $\ell_{\text{SPO+}}(h(x_T), c_T) = 0$ with probability 1, and so we have proven that $H_T^0 \subseteq \bar{H}_T^0$.

□

Proof of Theorem 2.4.3. We provide the proof of part *(a)*, as the proofs of parts *(b)*, *(c)*, and *(d)* are completely analogous to Theorem 2.4.1.

Let us now prove part *(a)*. When $T = 0$, part *(a)* holds by the definition of $r_0 \geq \omega_\ell(\hat{\mathcal{C}}, \mathcal{C})$. Otherwise, let $T \geq 1$ be given. For any $t \in \{1, \ldots, T\}$, recall that the re-weighted loss function at iteration t is in this case given by $\ell^{\texttt{rew}}(h; z_t) := d_t^M \ell(h(x_t), c_t) + (1 - d_t^M)(q_t/\tilde{p})\ell(h(x_t), c_t)$. Since $\tilde{p} > 0$, and q_t is a random variable that independent of x_t, c_t, and d_t^M, we condition on the two possible values of $q_t \in \{0, 1\}$ and obtain the following decomposition:

$$\begin{aligned}\mathbb{E}[\ell^{\texttt{rew}}(h; z_t)|\mathcal{F}_{t-1}] &= \mathbb{E}[\ell(h(x_t), c_t) d_t^M | \mathcal{F}_{t-1}] + \mathbb{E}[\ell(h(x_t), c_t)(1 - d_t^M)|\mathcal{F}_{t-1}] \\ &= \mathbb{E}[\ell(h(x_t), c_t)|\mathcal{F}_{t-1}] = \mathbb{E}[\ell(h(x_t), c_t)] = R_\ell(h),\end{aligned}$$

where we have also used that (x_t, c_t) is independent of $\mathcal{F}_{t-1}$. In other words, the conditional expectation of re-weighted surrogate loss at iteration t equals the surrogate risk. Consider the above applied to both $h \in \mathcal{H}$ and $h^* \in \mathcal{H}_\ell^*$ and averaged over $t \in \{1, \ldots, T\}$ to yield:

$$R_\ell(h) - R_\ell(h^*) \;=\; \frac{1}{T}\sum_{t=1}^{T} \left(\mathbb{E}[\ell^{\texttt{rew}}(h; z_t)|\mathcal{F}_{t-1}] - \mathbb{E}[\ell^{\texttt{rew}}(h^*; z_t)|\mathcal{F}_{t-1}]\right) \tag{A.11}$$

As before, we denote the discrepancy between the expectation and the true excess re-weighted loss of predictor h at time t by $Z_h^{\texttt{t}}$, i.e., $Z_h^{\texttt{t}} := \mathbb{E}[\ell^{\texttt{rew}}(h; z_t) - \ell^{\texttt{rew}}(h^*; z_t)|\mathcal{F}_{t-1}] - (\ell^{\texttt{rew}}(h; z_t) - \ell^{\texttt{rew}}(h^*; z_t))$. Recall that the empirical re-weighted loss in Algorithm 1 is $\hat{\ell}^T(h) = \frac{1}{T}\left(\sum_{(x,c)\in W_T} \ell(h(x), c) + \frac{1}{\tilde{p}}\sum_{(x,c)\in \tilde{W}_T} \ell(h(x), c)\right) = \frac{1}{T}\sum_{t=1}^{T} \ell^{\texttt{rew}}(h; z_t)$. Thus, (A.11) is equivalently written as:

$$R_\ell(h) - R_\ell(h^*) \;=\; \frac{1}{T}\sum_{t=1}^{T} Z_h^{\texttt{t}} + \frac{1}{T}\sum_{t=1}^{T} \left(\ell^{\texttt{rew}}(h; z_t) - \ell^{\texttt{rew}}(h^*; z_t)\right), \tag{A.12}$$

for any $h \in \mathcal{H}$. To bound the term $\frac{1}{T}\sum_{t=1}^{T} Z_h^{\texttt{t}}$, we apply Proposition 2.4.1 twice to both h and h^*, with $\epsilon \leftarrow \frac{\omega_\ell(\hat{\mathcal{C}}, \mathcal{C})}{\tilde{p}}\sqrt{\frac{4\ln(2TN_1/\delta)}{T}}$. Then, by considering their differences using the union bound we have that

$$\sup_{h\in\mathcal{H}} \left|\frac{1}{T}\sum_{t=1}^{T} Z_h^{\texttt{t}}\right| \;\leq\; 2\epsilon = r_T,$$

with probability at least $1 - \frac{\delta}{2T^2}$. Since h_T is the minimizer of the empirical re-weighted loss $\hat{\ell}^T(h)$ over $\mathcal{H}$, we have that $\ell^{\texttt{rew}}(h_T; z_t) - \ell^{\texttt{rew}}(h^*; z_t) \leq 0$ in (A.12) and we obtain that $R_\ell(h) - R_\ell(h^*) \leq r_T$ with probability at least $1 - \frac{\delta}{2T^2}$. Finally, applying the union bound over all $T \geq 1$, we obtain that $R_\ell(h) - R_\ell(h^*) \leq r_T$ simultaneously for all $T \geq 1$ with probability at least $1 - \delta$, which is the result of part *(a)*.

□

A.0.3 Proofs for Section 2.5

Proof of Proposition 2.5.1. Since $\phi(\epsilon) = C_\phi\sqrt{\epsilon}$, and $C_\phi \in (0, \frac{1}{36L^2})$, we set $\bar{C} = \sqrt{\frac{r_1}{5L}}$. We use induction to prove that $b_T \leq \bar{C}/T^{-1/4}$, for all T. For simplicity, we ignore the log term when analyzing the order, and assume that $r_t \leq \frac{r_1}{\sqrt{t}}$.

We assume $b_t \leq \bar{C}/t^{-1/4}$, for $1 \leq t \leq T-1$.

Then, since $b_t = 2\phi(2r_t + \frac{2L}{t}\sum_{i=0}^{t-1} b_i^2)$, we have that when $t = T$,

$$\begin{aligned} b_t &= 2C_\phi \sqrt{r_t + \frac{2L}{t}\sum_{i=0}^{t-1} b_i^2} \\ &\leq 2C_\phi \sqrt{\frac{r_1}{\sqrt{t}} + \frac{2L}{t}\sum_{i=0}^{t-1} b_i^2} \\ &\leq 2C_\phi \sqrt{\frac{r_1}{\sqrt{t}} + \frac{2L}{t}\sum_{i=0}^{t-1} \frac{\bar{C}^2}{\sqrt{i}}} \\ &\leq 2C_\phi \sqrt{\frac{r_1}{\sqrt{t}} + \frac{4L}{t}\bar{C}\sqrt{t}}. \end{aligned}$$

The first inequality is by $r_t \leq \frac{r_1}{\sqrt{t}}$. The last inequality is from the fact that $\frac{1}{t}\sum_{i=0}^{t-1} i^{-1/2} \leq 2\sqrt{t}$.

Then, we plug in the value of $\bar{C}$, we have that $b_t \leq \bar{C}/t^{-1/4}$, when $t = T$. Thus, $R_{\text{SPO+}}(h_T) - R^*_{\text{SPO+}} \leq \tilde{\mathcal{O}}(T^{-1/2})$. Consequently, for the polyhedral case, $R_{\text{SPO}}(h_T) - R^*_{\text{SPO}} \leq 2\Psi(2b_T)\omega_S(\mathcal{C}) \leq \tilde{\mathcal{O}}(T^{-\kappa/4})$. For the strongly-convex feasible region, we set $\gamma_T = b$, and then we can obtain the same order $\tilde{\mathcal{O}}(T^{-\kappa/4})$ for the excess SPO risk.

Next, we consider the bound for the label complexity. We set δ as a very small number, for example, $\delta \leq \tilde{\mathcal{O}}(1/T^3)$, so we can ignore the last term in the label complexity in part *(d)*. Then, we have that $\mathbb{E}[n_t] \leq \tilde{\mathcal{O}}(2\sum_{t=1}^{T}\Psi(2b_t))$. Because $b_t \leq \tilde{\mathcal{O}}(T^{-\kappa/4})$, we have that $\sum_{t=1}^{T}\Psi(2b_t) \leq \tilde{\mathcal{O}}(T^{1-\kappa/4})$. Then, we can obtain the label complexity in Proposition 2.5.1 depending on the value of κ.

□

Proof of Proposition 2.5.2. We first consider the label complexity. By the part *(d)* in Theorem 2.4.3, the total label complexity $\mathbb{E}[n_t]$ is at most

$$\begin{aligned} \tilde{p}T + 2\sum_{t=1}^{T}\Psi(2b_t) &= \tilde{p}T + \sum_{t=1}^{T} 2\Psi\left(2\phi\left(\frac{1}{\tilde{p}}\sqrt{2\ln(t/\delta)/t}\right)\right) \\ &\leq \tilde{p}T + \sum_{t=1}^{T} C' \cdot \left(\frac{1}{\tilde{p}}\right)^{\frac{\kappa}{2}} (\ln(t/\delta)/t)^{\kappa/4} \\ &\leq \tilde{\mathcal{O}}\left(\tilde{p}T + \left(\frac{1}{\tilde{p}}\right)^{\frac{\kappa}{2}} (T\ln T)^{1-\kappa/4}\right). \end{aligned}$$

□

The first inequality is because of assumptions 2.6.1 and 2.5.1. The second inequality is because of the integration. To minimize the order of T, we set $\tilde{p} = T^{-\frac{k}{2(k+2)}}$. Then, the label complexity $\mathbb{E}[n_t]$ is at most $\tilde{\mathcal{O}}\left(T^{1-\frac{k}{2(k+2)}}\right)$ for $\kappa > 0$.

Next, since $r_T \leq \tilde{\mathcal{O}}(\frac{1}{\sqrt{T}\tilde{p}}) = \tilde{\mathcal{O}}(T^{-\frac{1}{\kappa+2}})$, we obtain the risk bounds for the surrogate loss. Since ϕ is a square root function. The SPO risk is at most $2\Psi(2\phi(r_T)) \leq \tilde{\mathcal{O}}(T^{-\frac{\kappa}{2(\kappa+2)}})$. □

Proof of Proposition 2.5.3. The reason why the excess surrogate risk in Proposition 2.5.2 is larger than $\tilde{\mathcal{O}}(T^{-1/2})$ is because $r_T \leq \tilde{\mathcal{O}}(\frac{1}{\tilde{p}\sqrt{T}})$. Indeed, when $\tilde{p} \leftarrow T^{-\frac{\kappa}{2(\kappa+2)}}$, then $r_T \leq \tilde{\mathcal{O}}\left(T^{\left(\frac{\kappa}{2(\kappa+2)}-\frac{1}{2}\right)}\right)$, which is larger than $\tilde{\mathcal{O}}(T^{-1/2})$. Moreover, the dependence on $\tilde{p}$ comes from the bound on the re-weighted loss, since the re-weighted loss is upper-bounded by $\frac{\omega_\ell(\hat{\mathcal{C}},\mathcal{C})}{\tilde{p}}$. When $T \to \infty$, $\tilde{p} \to 0$, and thus, the re-weighted loss tends to infinity.

Given the output predictor h_T at iteration T, recall that $Z^{\mathtt{t}}_{h_T} := \mathbb{E}[\ell^{\mathtt{rew}}(h_T; z_t) - \ell^{\mathtt{rew}}(h^*; z_t)|\mathcal{F}_{t-1}] - (\ell^{\mathtt{rew}}(h_T; z_t) - \ell^{\mathtt{rew}}(h^*; z_t))$. Since $\mathbb{E}[Z^{\mathtt{t}}_{h_T}] = 0$, we have that $\sum_{t=1}^T Z^{\mathtt{t}}_{h_T}$ is a martingale.

Thus, if we can further remove the dependence on $\tilde{p}$ and show that $Z^{\mathtt{t}}_{h_T}$ is finite for all $T \geq 0$, then we can apply the Azuma's Inequality and achieve the convergence rate $\tilde{\mathcal{O}}(T^{-1/2})$ for $\frac{1}{T}\sum_{t=1}^T Z^{\mathtt{t}}_{h_T}$.

By the Lipschitz property, we have that

$$|\ell(h_T(x), c) - \ell(h^*(x), c)| \leq L_\kappa \|h_T(x) - h^*(x)\|, \tag{A.13}$$

for all $x \in \mathcal{X}$. Recall that $\ell^{\mathtt{rew}}(h; z_t) := d_t^M \ell(h(x_t), c_t) + (1 - d_t^M)\frac{q_t}{\tilde{p}}\ell(h(x_t), c_t)$. Since when $d_t^M = 1$, $\ell^{\mathtt{rew}}(h; z_t)$ is obviously upper bounded by $\omega_S(\hat{\mathcal{C}}, \mathcal{C})$, and thus $Z^{\mathtt{t}}_{h_T}$ is obviously bounded. Therefore, to show $Z^{\mathtt{t}}_{h_T}$ is bounded, it suffices to consider the case when $d_t^M = 0$. When $d_t^M = 0$, we have that $\ell^{\mathtt{rew}}(h; z_t) = \frac{q_t}{\tilde{p}}\ell(h(x_t), c_t)$. Since $\tilde{p}_t \geq \alpha_1 \|h_T(x) - h^*(x)\|$, we have that

$$\frac{q_t}{\tilde{p}}|\ell(h_T(x), c) - \ell(h^*(x), c)| \leq \frac{1}{\tilde{p}}|\ell(h_T(x), c) - \ell(h^*(x), c)| \leq L_\kappa/\alpha_1 = \tilde{\mathcal{O}}(1).$$

The above implies that $Z^{\mathtt{t}}_{h_T}$ is also bounded when $d_t^M = 0$. We denote the upper bound of $Z^{\mathtt{t}}_{h_T}$ by $\sqrt{C_1} > 0$. Thus, when applying Azuma's inequality to the sequence $\sum_{t=1}^T Z^{\mathtt{t}}_{h_T}$, and taking the average, we can remove the dependence on $\tilde{p}$ and have that

$$\left|\frac{1}{T}\sum_{t=1}^T Z^{\mathtt{t}}_{h_T}\right| \leq \epsilon,$$

with probability at least $1 - 2N_1 e^{-\frac{\epsilon^2 T}{2C_1}}$.

Recall that

$$R_\ell(h_T) - R_\ell(h^*) \;=\; \frac{1}{T}\sum_{t=1}^T Z^{\mathtt{t}}_{h_T} + \frac{1}{T}\sum_{t=1}^T \left(\ell^{\mathtt{rew}}(h_T; z_t) - \ell^{\mathtt{rew}}(h^*; z_t)\right).$$

Since $\ell^{\mathtt{rew}}(h_T; z_t) - \ell^{\mathtt{rew}}(h^*; z_t) \leq 0$, we conclude that $R_\ell(h_T) - R_\ell(h^*)$ is at most ϵ. Setting $2N_1 e^{-\frac{\epsilon^2 T}{2C_1}} = \delta$, we obtain that $R_\ell(h_T) - R_\ell(h^*) \leq 2\sqrt{\frac{2C_1 \ln(2N_1/\delta)}{T}}$ with probability at least $1 - \delta$. Thus, we conclude that $R_\ell(h_T) - R_\ell(h^*)$ converges to zero at rate $\tilde{\mathcal{O}}(\sqrt{\ln(T)/T})$.

Finally, to derive the upper bound for the expected number of acquired labels $\mathbb{E}[n_T]$, by Theorem 2.4.3, we have that $\|h_t(x) - h^*(x)\| \leq \phi(r_t)$. Since $r_t \leq \tilde{\mathcal{O}}(T^{-\frac{1}{\kappa+2}})$, we further have that

$$\begin{aligned}
\mathbb{E}[n_T] &\leq \sum_{t=1}^{T} [\|h_t(x) - h^*(x)\| + \Psi(2b_{t-1})] + \delta T \\
&\leq \sum_{t=1}^{T} [\phi(r_t) + \Psi(2b_{t-1})] + \delta T \\
&\leq \sum_{t=1}^{T} \left[\tilde{\mathcal{O}}(T^{-\frac{1}{2(\kappa+2)}}) + \tilde{\mathcal{O}}(T^{-\frac{\kappa}{2(k+2)}})\right] + \delta T \\
&\leq \tilde{\mathcal{O}}(T^{1-\frac{1}{2(\kappa+2)}}) + \tilde{\mathcal{O}}(T^{1-\frac{\kappa}{2(k+2)}}) + \delta T.
\end{aligned}$$

Thus, we have that $\mathbb{E}[n_t] \leq \tilde{\mathcal{O}}(T^{1-\frac{\min\{\kappa,1\}}{2(\kappa+2)}})$ when $\delta \leq T^{-2}$.

□

A.0.4 Proofs for Section 2.6

Proof of Lemma 2.6.1. For given $x \in \mathcal{X}$, let $\bar{c} = \mathbb{E}[c|x]$ and $\Delta = h(x) - h^*(x)$. According to Theorem 1 in Elmachtoub and Grigas (2022), it holds that

$$\mathbb{E}[\ell_{\text{SPO+}}(h(x), c) - \ell_{\text{SPO+}}(h^*(x), c)|x] = \mathbb{E}[(c + 2\Delta)^T (w^*(c) - w^*(c + 2\Delta))|x].$$

Without loss of generality, we assume $d_S > 0$. Otherwise, the constant Ξ_S will be zero and the bound will be trivial.

Define the function $\iota(\kappa) := \frac{\kappa^2}{M}$ for $\kappa \in [0, \frac{M}{2}]$ and $\iota(\kappa) := \kappa - \frac{M}{4}$ for $\kappa \in [\frac{M}{2}, \infty)$, where $M > 1$ is a scaler which is larger than σ. Let $\kappa = \|\Delta\|_2$ and $A \in \mathbb{R}^{d\times d}$ be an orthogonal matrix such that $A^T\Delta = \kappa \cdot e_d$ for $e_d = (0, \ldots, 0, 1)^T$. We implement a change of basis and let the new basis be $A = (a_1, \ldots, a_d)$. With a slight abuse of notation, we keep the notation the same after the change of basis, for example, now the vector Δ equals $\kappa \cdot e_d$. Rewrite c as $c = (c', \xi)$, where $c' \in \mathbb{R}^{d-1}$ and $\xi \in \mathbb{R}$. Define $\bar{c'} := \mathbb{E}[c']$ and $\bar{\xi'} = \mathbb{E}[\xi]$. Then by applying the results in Lemma 7 of H. Liu and Grigas (2021), for any $\tilde{\kappa} \in (0, \kappa]$, it holds that

$$\mathbb{E}\left[(c + 2\Delta)^T (w^*(c) - w^*(c + 2\Delta))|x\right] \geq \frac{\alpha\tilde{\kappa}\kappa e^{-\frac{3\tilde{\kappa}^2 + 3\bar{\xi}^2 + \|\bar{c'}\|_2^2}{2\sigma^2}}}{4\sqrt{2\pi\sigma^2}} \cdot \Xi_S d_S.$$

Let $\tilde{\kappa} = \min\{\kappa, \sigma\}$, it holds that

$$\mathbb{E}[\ell_{\text{SPO+}}(h(x), c) - \ell_{\text{SPO+}}(h^*(x), c)|x] \geq \frac{\alpha\Xi_S}{4\sqrt{2\pi}e^{\frac{3(1+\beta^2)}{2}}} \cdot \min\left\{\frac{\kappa^2}{M}, \kappa\right\}.$$

Define the function $\iota(\kappa) := \frac{\kappa^2}{M}$ for $\kappa \in [0, \frac{M}{2}]$ and $\iota(\kappa) := \kappa - \frac{M}{4}$ for $\kappa \in [\frac{M}{2}, \infty)$, we have $\iota(\kappa)$ is the convex biconjugate of $\min\left\{\frac{\kappa^2}{M}, \kappa\right\}$. By taking the expectation on x, it holds that

$$\mathbb{E}[\ell_{\text{SPO+}}(h(x), c) - \ell_{\text{SPO+}}(h^*(x), c)] \geq \frac{\alpha \Xi_S}{4\sqrt{2\pi} e^{\frac{3(1+\beta^2)}{2}}} \cdot \mathbb{E}_x[\iota(\|h(x) - h^*(x)\|)]$$

Since $M \geq \max\{\sigma, 1\}$, taking $M = 2\rho(\hat{\mathcal{C}})$, we obtain that

$$R_{\text{SPO+}}(h) - R_{\text{SPO+}}(h^*) \geq \frac{\alpha \Xi_S}{8\sqrt{2\pi}\rho(\hat{\mathcal{C}}) e^{\frac{3(1+\beta^2)}{2}}} \cdot \mathbb{E}_x[\|h(x) - h^*(x)\|^2]$$

Then, combining the result with Assumption 2.6.1, we obtain Lemma 2.6.1. □

Proof of Lemma 2.6.2. For given $x \in \mathcal{X}$, let $\bar{c} = \mathbb{E}[c|x]$ and $\Delta = h(x) - h^*(x)$. By applying the results in Theorem C.2 of H. Liu and Grigas (2021), it holds that

$$\mathbb{E}\left[\ell_{\text{SPO+}}(h(x), c) - \ell_{\text{SPO+}}(h^*(x), c)|x\right] \geq \frac{\mu_S^2 r^{1/2}}{2^{1/2} L_S^{5/2}} \cdot \mathbb{E}_{c|x}\left[\|c + 2\Delta\| - \frac{c^T(c+2\Delta)}{\|c\|}\right].$$

It is easy to verify that $\|c + 2\Delta\| - \frac{c^T(c+2\Delta)}{\|c\|} \geq 0$, for any $c \in \mathbb{R}^d$.

For any Δ, we define a subspace $\Delta^{\perp,\beta_1}$ by $\Delta^{\perp,\beta_1} = \{c \in \mathcal{C} : 0 \leq c^T\Delta \leq \beta_1\|\Delta\|\|c\|\}$. By Definition 2.6.1, we have that $\mathbb{P}_{c|x}(c \in \Delta^{\perp,\beta_1}) \geq \beta_2 > 0$, for any $x \in \mathcal{X}$.

Thus, we have that

$$\begin{aligned}
\mathbb{E}\left[\|c + 2\Delta\| - \frac{c^T(c+2\Delta)}{\|c\|}|x\right] &\geq \mathbb{P}_{c|x}(c \in \Delta^{\perp,\beta_1})\mathbb{E}\left[\|c + 2\Delta\| - \frac{c^T(c+2\Delta)}{\|c\|}\Big| c \in \Delta^{\perp,\beta_1}\right] \\
&\geq \beta_2 \mathbb{E}\left[\|c + 2\Delta\| - \frac{c^T(c+2\Delta)}{\|c\|}\Big| c \in \Delta^{\perp,\beta_1}\right] \\
&= \beta_2 \mathbb{E}\left[\|c + 2\Delta\| - \|c\| - \frac{2c^T\Delta}{\|c\|}\Big| c \in \Delta^{\perp,\beta_1}\right].
\end{aligned}$$

Next, we show that $\mathbb{E}\left[\|c + 2\Delta\| - \|c\| - \frac{2c^T\Delta}{\|c\|}\Big| c \in \Delta^{\perp,\beta_1}\right] \geq \frac{1}{16\rho(\mathcal{C})}\|\Delta\|^2$.

Since $\|c\| \leq \rho(\mathcal{C})$, we have that

$$\begin{aligned}
\|c + 2\Delta\| - \|c\| - \frac{2c^T\Delta}{\|c\|} &= \sqrt{\|c\|^2 + 4\|\Delta\|^2 + 4c^T\Delta} - \|c\| - \frac{2c^T\Delta}{\|c\|} \\
&\geq \sqrt{\|c\|^2 + 4\|\Delta\|^2 + 4\beta_1\|c\|\|\Delta\|} - \|c\| - 2\beta_1\|\Delta\| \qquad \text{(A.14)} \\
&\geq \sqrt{\rho(\mathcal{C})^2 + 4\|\Delta\|^2 + 4\beta_1\rho(\mathcal{C})\|\Delta\|} - \rho(\mathcal{C}) - 2\beta\|\Delta\| \qquad \text{(A.15)}
\end{aligned}$$

Inequality (A.14) is because $\sqrt{\|c\|^2 + 4\|\Delta\|^2 + 4\beta_1\|c\|\|\Delta\|} - \|c\| - 2\beta_1\|\Delta\|$ is a decreasing function of β_1. Inequality (A.15) is because $\sqrt{\|c\|^2 + 4\|\Delta\|^2 + 4\beta_1\|c\|\|\Delta\|} - \|c\| - 2\beta_1\|\Delta\|$ is a decreasing function of $\|c\|$ when $\beta_1 < 1$.

Since $\|\Delta\| \leq 2\rho(\mathcal{C}, \hat{\mathcal{C}})$, by minimizing function $f(x) = \frac{\sqrt{\rho(\mathcal{C})^2 + 2\beta_1\rho(\mathcal{C})x + x^2} - \rho(\mathcal{C}) - \beta_1 x}{x^2}$ over $(0, 2\rho(\mathcal{C}, \hat{\mathcal{C}}))$, we have that

$$\frac{\sqrt{\rho(\mathcal{C})^2 + 4\|\Delta\|^2 + 4\beta_1\rho(\mathcal{C})\Delta} - \rho(\mathcal{C}) - 2\beta_1\|\Delta\|}{\|\Delta\|^2} > \min\left\{\frac{2(1-\beta_1^2)}{\rho(\mathcal{C}, \hat{\mathcal{C}})}, \frac{\sqrt{17+8\beta} - 1 - 4\beta_1}{4\rho(\mathcal{C}, \hat{\mathcal{C}})}\right\}$$

Thus, we have that $\mathbb{E}\left[\|c + 2\Delta\| - \|c\| \middle| c \in \Delta^{\perp, \beta_1}\right] \geq \min\left\{\frac{2(1-\beta_1^2)}{\rho(\mathcal{C}, \hat{\mathcal{C}})}, \frac{\sqrt{17+8\beta_1} - 1 - 4\beta_1}{4\rho(\mathcal{C}, \hat{\mathcal{C}})}\right\} \|\Delta\|^2$.
Thus, we conclude that for all $\mathbb{P} \in \mathcal{P}_{\beta_1, \beta_2}$, it holds that

$$\|h(x) - h^*(x)\|^2 \leq \frac{\mu_S^2 r^{1/2}}{2^{1/2}\beta_2 L_S^{5/2}} \cdot \min\left\{\frac{2(1-\beta_1^2)}{\rho(\mathcal{C}, \hat{\mathcal{C}})}, \frac{\sqrt{17+8\beta_1} - 1 - 4\beta_1}{4\rho(\mathcal{C}, \hat{\mathcal{C}})}\right\}^{-1} \cdot \mathbb{E}_{c|x}\left[\ell_{\text{SPO+}}(h(x), c) - \ell_{\text{SPO+}}(h^*(x), c)\right].$$

Taking the expectation of both sides with x, we obtain that

$$\mathbb{E}[\|h(x) - h^*(x)\|^2] \leq \frac{\mu_S^2 r^{1/2}}{2^{1/2}\beta_2 L_S^{5/2}} \min\left\{\frac{2(1-\beta_1^2)}{\rho(\mathcal{C}, \hat{\mathcal{C}})}, \frac{\sqrt{17+8\beta_1} - 1 - 4\beta_1}{4\rho(\mathcal{C}, \hat{\mathcal{C}})}\right\}^{-1} \cdot (R_{\text{SPO+}}(h) - R_{\text{SPO+}}(h^*)).$$

Then, combining the result with Assumption 2.6.1, we obtain Lemma 2.6.2. □

Proof of Lemma 2.6.3. Since the hypothesis class is well-specified, we denote $h^*(x)$ by $\bar{c}$, given $x \in \mathcal{X}$. Then, we define $\Delta = c_1 - \bar{c}$. According to Theorem 1 in Elmachtoub and Grigas (2022), we have that the excess SPO+ risk at x for the prediction c_1 is

$$\mathbb{E}[\ell_{\text{SPO+}}(\bar{c} + \Delta, c) - \ell_{\text{SPO+}}(\bar{c}, c)|x] = \mathbb{E}[(c + 2\Delta)^T(w^*(c) - w^*(c + 2\Delta))|x]$$

According to Lemmas 1 and 2 in H. Liu and Grigas (2021), we have that for any $c_1, c_2 \in \mathcal{C}$, it holds that

$$c_1^T(w^*(c_2) - w^*(c_1)) \leq \frac{L_S^2\rho(\mathcal{C})\sqrt{r - f_{\min}}}{\sqrt{2}\mu_S^{1.5}} \left\| \frac{c_1}{\|c_1\|} - \frac{c_2}{\|c_2\|} \right\|^2.$$

Replacing c_1 and c_2 with $c + 2\Delta$ and c, we obtain that

$$\mathbb{E}[\ell_{\text{SPO+}}(\bar{c} + \Delta, c) - \ell_{\text{SPO+}}(\bar{c}, c)|x] \leq \frac{L_S^2\rho(\mathcal{C})\sqrt{r - f_{\min}}}{\sqrt{2}\mu_S^{1.5}} \mathbb{E}\left[\left\| \frac{c}{\|c\|} - \frac{c + 2\Delta}{\|c + 2\Delta\|} \right\|^2\right].$$

Thus, to prove Lemma 2.6.3, it suffices to show that $\left\| \frac{c}{\|c\|} - \frac{c+2\Delta}{\|c+2\Delta\|} \right\| \leq \frac{4}{\beta}\|\Delta\|$ for any realized c and any $\Delta \in \mathbb{R}^d$.

We consider two cases: (1) $\|2\Delta\| \geq \|c\|$, and (2) $\|2\Delta\| \leq \|c\|$.

In the first case, since $\|c\| \geq \beta$, we have that $\|2\Delta\| \geq \|c\| \geq \beta$. Since $\left\| \frac{c}{\|c\|} - \frac{c+2\Delta}{\|c+2\Delta\|} \right\| \leq 2$, we have that $\left\| \frac{c}{\|c\|} - \frac{c+2\Delta}{\|c+2\Delta\|} \right\| \leq \frac{4\|\Delta\|}{\beta}$.

In the other case, when $\|2\Delta\| \leq \|c\|$, we have

$$\left\| \frac{c}{\|c\|} - \frac{c+2\Delta}{\|c+2\Delta\|} \right\| = \sqrt{2 - 2\frac{c^T(c+2\Delta)}{\|c\|\|c+2\Delta\|}}.$$

We use $\theta \in [0, \frac{\pi}{2})$ to denote the angle between c and $c+2\Delta$, then, we have $\left\| \frac{c}{\|c\|} - \frac{c+2\Delta}{\|c+2\Delta\|} \right\| = \sqrt{2 - 2\cos(\theta)} \leq 2\sin(\theta)$. Since $2\|\Delta\| \geq 2\|c\|\sin(\theta)$, we have that $\left\| \frac{c}{\|c\|} - \frac{c+2\Delta}{\|c+2\Delta\|} \right\| \leq \frac{2\|\Delta\|}{\|c\|} \leq \frac{2}{\beta}\|\Delta\|$. Thus, we obtain Lemma 2.6.3. □

Appendix B

Proof for Chapter 2

B.1 Proofs in Section 3.3

Proof of Theorem 3.3.1. We denote the predictor $f\left(\mathcal{S}_{t-1}\bigcup\{(\xi_t,\mathbf{y})\}\right)$ by h_f. By the definition of the value of one data point, since the predictions for different types are independent, so the predictions on $\xi \neq \xi_t$ remain the same for h_{t-1} and h_f. Thus, we have that $\ell(h_{t-1}(\xi),\mathbb{E}[\mathbf{y}|\xi]) = \ell(h_{t-1}(\xi),\mathbb{E}[\mathbf{y}|\xi])$, for $\xi \neq \xi_t$. Then, the value of one data point can be written as

$$\begin{aligned} V(\xi_t;\mathcal{S}_{t-1}) &= \beta\cdot\mathsf{Regret}(h_{t-1}) - \beta\cdot\mathbb{E}\left[\mathsf{Regret}\left(f\left(\mathcal{S}_{t-1}\bigcup\{(\xi_t,\mathbf{y})\}\right)\right)\Big|\xi_t\right] \\ &= \beta\mathbb{E}[\ell(h_{t-1}(\xi),\mathbb{E}[\mathbf{y}|\xi])] - \beta\mathbb{E}[\ell(h_f(\xi),\mathbb{E}[\mathbf{y}|\xi])] \\ &= \mu(\xi_t)\beta\left[\ell(h_{t-1}(\xi_t),\mathbb{E}[\mathbf{y}|\xi_t]) - \ell(h_f(\xi_t),\mathbb{E}[\mathbf{y}|\xi_t])\right]. \end{aligned}$$

Since $\ell(h_f(\xi_t),\mathbb{E}[\mathbf{y}|\xi_t]) \geq 0$, we have that

$$V(\xi_t;\mathcal{S}_{t-1}) \leq \mu(\xi_t)\beta\ell(h_{t-1}(\xi_t),\mathbb{E}[\mathbf{y}|\xi_t]) = \mu(\xi_t)\beta\mathbb{E}[\mathbf{y}^T|\xi_t](w^*(\mathbb{E}[\mathbf{y}|\xi_t]) - w^*(h_{t-1}(\xi_t))). \quad \text{(B.1)}$$

For simplicity, we denote $h_{t-1}(\xi_t)$ by $\hat{y}$ and denote the prediction error for $\hat{y}$ by ρ_{t-1}, i.e., $\|\mathbb{E}[\mathbf{y}^T|\xi_t] - \hat{y}\| \leq \rho_{t-1}$, we have the following upper bound for $\mathbb{E}[\mathbf{y}^T|\xi_t](w^*(\mathbb{E}[\mathbf{y}|\xi_t] - w^*(\hat{y}))$:

$$\mathbb{E}[\mathbf{y}^T|\xi_t](w^*(\mathbb{E}[\mathbf{y}|\xi_t] - w^*(\hat{y})) \leq \max_{y\in\mathcal{Y}:\|y-\hat{y}\|\leq\rho_{t-1}} : y^T(w^*(y) - w^*(\hat{y})). \quad \text{(B.2)}$$

For simplicity, we denote the above upper bound by $V_M^O(\xi_t,\hat{\mathbf{y}},\rho_{t-1})$. Then, we have that the value of one data point is at most $\mu(\xi_t)\beta V_M^O(\xi_t,\hat{\mathbf{y}},\rho_{t-1})$.

Next, let us consider the upper bound for $V_M^O(\xi_t,\hat{\mathbf{y}},\rho_{t-1})$. Since $\hat{y}^T(w^*(\hat{y}) - w^*(y)) \geq 0$, for any $y,\hat{y}\in\mathcal{Y}$, we have that

$$y^T(w^*(y) - w^*(\hat{y})) \leq y^T(w^*(y) - w^*(\hat{y})) + \hat{y}^T(w^*(\hat{y}) - w^*(y)) \leq \|y-\hat{y}\|\|w^*(\hat{y}) - w^*(y)\|.$$

Since $w^*(\hat{\mathbf{y}})$ and $w^*(y)$ are binary vectors with z ones and $d-z$ zeros, we have that $\|w^*(\hat{y}) - w^*(y)\| \leq \sqrt{2\min\{z,d-z\}}$. Together with the fact that $\|y-\hat{y}\| \leq \rho_{t-1}(\xi)$, we obtain that $V_M^O(\xi_t,\hat{\mathbf{y}},\rho_{t-1}) \leq \sqrt{2\min\{z,d-z\}}\rho_{t-1}$.

On the other hand, if $\nu_S(\mathbf{y}) > \rho$, by Lemma 1 in M. Liu et al., 2023, we have that $w^*(\hat{y}) = w^*(\mathbb{E}[\mathbf{y}|\xi_t])$, and thereby $\ell(h_{t-1}(\xi_t), \mathbb{E}[\mathbf{y}|\xi_t])$ and $V_M^O(\xi_t, \hat{\mathbf{y}}, \rho_{t-1})$ are both zero. Thus, we have that

$$V_M^O(\xi_t, \hat{\mathbf{y}}, \rho_{t-1}) \leq \sqrt{2\min\{z, d-z\}}\rho_{t-1}\mathbb{I}\{\nu_S(\hat{\mathbf{y}}) \leq \rho_{t-1}\}.$$

Combining it with (B.1), we obtain the upper bounds of the value of one data point in Theorem 3.3.1.

□

Before providing the proof of Theorem 3.6.1, we first provide Lemma B.1.1, which provides two upper bounds for the sum of the expectation of $U_M(\xi, h_t(\xi), \rho_t(\xi))$.

Lemma B.1.1. *Under the same setting as Theorem 3.6.1, suppose at each iteration t, for any type ξ, $U_M(\xi, h_{t-1}(\xi), \rho_{t-1}(\xi)) \geq c_{min}$ and $n_t(\xi) > 0$. Then, given $\delta \in (0,1)$, we have the following two upper bounds for $\sum_{\xi\in[m]} \mathbb{E}\left[U_M(\xi, h_{T-1}(\xi), \rho_{T-1}(\xi))|\xi\right]$ with probability at least $1 - me^{-\frac{\underline{\mu}p_{min}t}{8}} - \frac{\delta}{t^2}$,*

$$\sum_{\xi\in[m]} \mathbb{E}\left[U_M(\xi, h_{T-1}(\xi), \rho_{T-1}(\xi))|\xi\right] \leq$$
$$\beta \min\left\{\sum_{\xi\in[m]} 4\mu(\xi)\sqrt{\min\{z, d-z\}}\eta_{\mathcal{Y}}\sqrt{\frac{d\ln(t/\delta)}{p_{min}\mu(\xi)t}}, \sqrt{d}\eta_{\mathcal{Y}}\Psi\Big(4\eta_{\mathcal{Y}}\sqrt{\frac{2d\ln(t/\delta)}{p_{min}\mu(\xi)t}}\Big)\right\}.$$

Proof of Lemma B.1.1. Because $U_M(\xi, h_{t-1}(\xi), \rho_{t-1}(\xi))$ is a non-increasing sequence for each ξ, $U_M(\xi, h_{t-1}(\xi), \rho_{t-1}(\xi)) \geq c_{\mathsf{min}}$ implies that we offer some incentive between $[c_{\mathsf{min}}, c_{\mathsf{max}}]$ for all customers before iteration t. Thus, the probability of acquiring the label of any type ξ at one iteration before t is at least $p_{\mathsf{min}}\mu(\xi) \geq p_{\mathsf{min}}\underline{\mu}$. Next, we prove two upper bounds in Lemma B.1.1 respectively.

First upper bound for $\sum_{\xi\in[m]} \mathbb{E}[U_M(\xi, h_t(\xi), \rho_t(\xi))|\xi]$**.** Since the radius of $\mathcal{Y}$ is at most $2\eta_{\mathcal{Y}}$, by Hoeffding's inequality, when the number of observations up to time t, under type ξ is $n_t(\xi)$, for any $\delta > 0$, with probability at least $1 - \delta/t^2$, $\|\mathbb{E}[\mathbf{y}|\xi] - h(\xi)\| \leq 2\eta_{\mathcal{Y}}\sqrt{\frac{d\ln(t/\delta)}{n_t(\xi)}} = \rho_t(\xi)$. Next, we provide an upper bound for $\mathbb{E}[\rho_t(\xi)]$.

Since the probability of taking the survey is at least $p_{\mathsf{min}}\mu(\xi)$ for any ξ, $\mathbb{E}[\frac{1}{1+n_t(\xi)}] \leq \mathbb{E}[\frac{1}{1+Z}]$, where Z is a binomial random variable with success probability at least $p_{\mathsf{min}}\mu(\xi)$ and t trials. By M.-T. Chao and Strawderman, 1972, we have that $\mathbb{E}[\frac{1}{1+Z}] \leq \frac{1}{p_{\mathsf{min}}\mu(\xi)t}$. Besides, $\frac{1}{n_t(\xi)} \leq \frac{2}{1+n_t(\xi)}$, when $n_t(\xi) > 0$. Therefore, when $n_t(\xi) > 0$, $\mathbb{E}[\frac{1}{n_t(\xi)}] \leq \mathbb{E}[\frac{2}{n_t(\xi)+1}] \leq \frac{2}{p_{\mathsf{min}}\mu(\xi)t}$.

Therefore, for one given ξ, when $n_t(\xi) > 0$, by Jensen's inequality, we have that

$$\mathbb{E}[\rho_t(\xi)|n_t(\xi) > 0] \leq \eta_{\mathcal{Y}}\sqrt{\mathbb{E}\Big[\frac{2d\ln(t/\delta)}{n_t(\xi)}|n_t(\xi) > 0\Big]} \leq 2\eta_{\mathcal{Y}}\sqrt{\frac{2d\ln(t/\delta)}{p_{\mathsf{min}}\mu(\xi)t}}.$$

Therefore, if $t < \underline{t}_\xi$, with probability at least $1 - me^{t\ln(1-p_{\min}\underline{\mu})} - \delta/t^2$, $n_t(\xi) > 0$, and thereby,

$$\begin{aligned}\mathbb{E}[U_M(\xi, h_t(\xi), \rho_t(\xi))|\xi] &\le \beta\mathbb{E}[\sqrt{2\min\{z, d-z\}}\mu(\xi)\rho_t(\xi)\mathbb{I}\{\nu_S(\hat{\mathbf{y}}) \le \rho_t(\xi_t)\}|\xi] \\ &\le \beta\mathbb{E}[\sqrt{2\min\{z, d-z\}}\mu(\xi)\rho_t(\xi)|\xi] \\ &\le \beta\mu(\xi)\sqrt{2\min\{z, d-z\}}\mathbb{E}[\rho_t(\xi)|\xi] \\ &\le \beta\mu(\xi)\sqrt{2\min\{z, d-z\}}\eta_{\mathcal{Y}} 2\sqrt{\frac{2d\ln(t/\delta)}{p_{\min}\mu(\xi)t}}.\end{aligned}$$

Therefore, when $t < \underline{t}_\xi$, the first upper bound for $\sum_{\xi\in[m]}\mathbb{E}[U_M(\xi, h_t(\xi), \rho_t(\xi))|\xi]$ is:

$$\sum_{\xi\in[m]}\mathbb{E}[U_M(\xi, h_t(\xi), \rho_t(\xi))|\xi] \le \beta\sum_{\xi\in[m]} 4\mu(\xi)\sqrt{\min\{z, d-z\}}\eta_{\mathcal{Y}} \cdot\sqrt{\frac{d\ln(t/\delta)}{p_{\min}\mu(\xi)t}}. \tag{B.3}$$

This upper bound does not incorporate the information of $\mathbb{I}\{\nu_S(\hat{\mathbf{y}}) \le \rho_t(\xi_t)\}$.

Second upper bound for $\sum_{\xi\in[m]}\mathbb{E}[U_M(\xi, h_t(\xi), \rho_t(\xi))|\xi]$**.** Next, we consider another upper bound for $\sum_{\xi\in[m]}\mathbb{E}[U_M(\xi, h_t(\xi), \rho_t(\xi))|\xi]$ when $c_{\min} = 0$ that focuses on $\mathbb{I}\{\nu_S(\hat{\mathbf{y}}) \le \rho_t(\xi_t)\}$.

$$\begin{aligned}\mathbb{E}[U_M(\xi, h_t(\xi), \rho_t(\xi))|\xi] &\le \beta\mathbb{E}[\sqrt{2\min\{z, d-z\}}\mu(\xi)\rho_t(\xi)\mathbb{I}\{\nu_S(\hat{\mathbf{y}}) \le \rho_t(\xi)\}|\xi] \\ &\le \beta\mathbb{P}(\nu_S(\hat{\mathbf{y}}) \le \rho_t(\xi))\mathbb{E}[\sqrt{2\min\{z, d-z\}}\mu(\xi)\rho_t(\xi)|\nu_S(\hat{\mathbf{y}}) \le \rho_t(\xi)] \\ &\le \beta\mu(\xi)\sqrt{d}\eta_{\mathcal{Y}}\mathbb{P}(\nu_S(\hat{\mathbf{y}}) \le \rho_t(\xi)).\end{aligned} \tag{B.4}$$

The last inequality is because we relax $\sqrt{2\min\{z, d-z\}}\rho_t(\xi)$ to $\sqrt{d}\eta_{\mathcal{Y}}$. Next, we show that $\mathbb{P}\big(\nu_S(\hat{\mathbf{y}}) \le \rho_t(\xi)\big) \le \mathbb{P}\big(\nu_S(\bar{\mathbf{y}}(\xi)) \le 2\rho_t(\xi)\big)$. When $\nu_S(\bar{\mathbf{y}}(\xi)) > 2\rho_t(\xi)$, since the distance function ν_S is a 1-Lipschitz function, we have that $\nu_S(\bar{\mathbf{y}}(\xi)) = \nu_S(\bar{\mathbf{y}}(\xi) - \hat{\mathbf{y}}(\xi) + \hat{\mathbf{y}}(\xi)) \le \|\hat{\mathbf{y}}(\xi) - \bar{\mathbf{y}}(\xi)\| + \nu_S(\hat{\mathbf{y}}(\xi))$. Since $\|\hat{\mathbf{y}}(\xi) - \bar{\mathbf{y}}(\xi)\| \le \rho_t(\xi)$ with probability at least $1-\delta$, we have that $\nu_S(\bar{\mathbf{y}}(\xi)) \le \rho_t(\xi) + \nu_S(\hat{\mathbf{y}}(\xi))$. Thus, $2\rho_t(\xi) < \nu_S(\bar{\mathbf{y}}(\xi)) \le \rho_t(\xi) + \nu_S(\hat{\mathbf{y}}(\xi))$, which implies $\rho_t(\xi) \le \nu_S(\hat{\mathbf{y}}(\xi))$. Therefore, $\mathbb{I}\{\nu_S(\bar{\mathbf{y}}(\xi)) > 2\rho_t(\xi)\}$ is true implies $\mathbb{I}\{\rho_t(\xi) \le \nu_S(\hat{\mathbf{y}}(\xi))\}$ is true. Thus, we have that $\mathbb{P}\big(\nu_S(\hat{\mathbf{y}}) \le \rho_t(\xi)\big) \le \mathbb{P}\big(\nu_S(\bar{\mathbf{y}}(\xi)) \le 2\rho_t(\xi)\big)$.

Then, following (B.4), we have with probability at least $1 - \delta/t^2$,

$$\mathbb{E}[U_M(\xi, h_t(\xi), \rho_t(\xi))|\xi] \le \beta\mu(\xi)\sqrt{d}\eta_{\mathcal{Y}}\mathbb{P}\big(\nu_S(\bar{\mathbf{y}}(\xi)) \le 2\rho_t(\xi)\big).$$

Since $\mathbb{I}\big\{\nu_S(\bar{\mathbf{y}}(\xi)) \ge 2\rho_t(\xi)\big\} \ge \mathbb{I}\big\{\nu_S(\bar{\mathbf{y}}(\xi)) \ge \alpha_t, \text{ and } \alpha_t \ge 2\rho_t(\xi)\big\}$, for some constant α_t that is independent of ξ, we have that $\mathbb{I}\big\{\nu_S(\bar{\mathbf{y}}(\xi)) \le 2\rho_t(\xi)\big\} \le \mathbb{I}\big\{\nu_S(\bar{\mathbf{y}}(\xi)) \le \alpha_t, \text{ or } \alpha_t \le 2\rho_t(\xi)\big\} \le \mathbb{I}\big\{\nu_S(\bar{\mathbf{y}}(\xi)) \le \alpha_t\big\} + \mathbb{I}\big\{\alpha_t \le 2\rho_t(\xi)\big\}$. As a result, for any given ξ, when $\alpha_t > 2\rho_t(\xi)$, we have

$$\begin{aligned}\mathbb{P}\big(\nu_S(\bar{\mathbf{y}}(\xi)) \le 2\rho_t(\xi)\big) &\le \mathbb{E}\left[\mathbb{I}\big\{\nu_S(\bar{\mathbf{y}}(\xi)) \le \alpha_t\big\} + \mathbb{I}\big\{\alpha_t \le 2\rho_t(\xi)\big\}\right] \\ &\le \mathbb{E}\left[\mathbb{I}\big\{\nu_S(\bar{\mathbf{y}}(\xi)) \le \alpha_t\big\}\right] = \mathbb{I}\big\{\nu_S(\bar{\mathbf{y}}(\xi)) \le \alpha_t\big\},\end{aligned}$$

where the last equality is because the expectation is regarding the randomness of n_t and there is no randomness in the expectation.

We set $\alpha_t = 4\eta_{\mathcal{Y}}\sqrt{\frac{2d\ln(t/\delta)}{p_{\mathsf{min}}\mu(\xi)t}}$. We have that

$$\begin{aligned}\mathbb{P}\big(\alpha_t \le 2\rho_t(\xi)\big) &= \mathbb{P}\left(4\eta_{\mathcal{Y}}\sqrt{\frac{2d\ln(t/\delta)}{p_{\mathsf{min}}\mu(\xi)t}} \le 2\rho_t(\xi)\right)\\ &= \mathbb{P}\left(4\eta_{\mathcal{Y}}\sqrt{\frac{2d\ln(t/\delta)}{p_{\mathsf{min}}\mu(\xi)t}} \le 4\eta_{\mathcal{Y}}\sqrt{\frac{d\ln(t/\delta)}{n_t(\xi)}}\right)\\ &= \mathbb{P}\big(n_t(\xi) \le 0.5p_{\mathsf{min}}\mu(\xi)t\big)\end{aligned}$$

As shown before, when $t \le \underline{t}_\xi$, we have

$$\mathbb{P}\big(n_t(\xi) \le 0.5p_{\mathsf{min}}\mu(\xi)t\big) \le \mathbb{P}\big(Z \le 0.5p_{\mathsf{min}}\mu(\xi)t\big),$$

where Z is a Binomial random variable with success probability at least $p_{\mathsf{min}}\mu(\xi)$ and number of trials t. Thus, by Chernoff's bound on Z, we have that

$$\mathbb{P}\big(n_t(\xi) \le 0.5p_{\mathsf{min}}\mu(\xi)t\big) \le e^{-\frac{\mu(\xi)p_{\mathsf{min}}t}{8}} \le e^{-\frac{\underline{\mu}p_{\mathsf{min}}t}{8}}.$$

Thus, by the union bound, for all ξ, with probability at least $1 - me^{-\frac{\underline{\mu}p_{\mathsf{min}}t}{8}} - \delta/t^2$, $\alpha_t > 2\rho_t(\xi)$, and thereby,

$$\mathbb{P}\big(\nu_S(\bar{\mathbf{y}}(\xi)) \le 2\rho_t(\xi)\big) \le \mathbb{I}\big\{\nu_S(\bar{\mathbf{y}}(\xi)) \le \alpha_t\big\}.$$

Then, we have the second upper bound: With probability at least $1 - me^{-\frac{\underline{\mu}p_{\mathsf{min}}t}{8}} - \delta/t^2$,

$$\begin{aligned}\sum_{\xi\in[m]} \mathbb{E}[U_M(\xi, h_t(\xi), \rho_t(\xi))|\xi] &\le \beta\sum_{\xi\in[m]} \sqrt{d}\mu(\xi)\eta_{\mathcal{Y}}\mathbb{P}\big(\nu_S(\bar{\mathbf{y}}(\xi)) \le 2\rho_t(\xi)\big)\\ &\le \beta\sum_{\xi\in[m]} \sqrt{d}\eta_{\mathcal{Y}}\mu(\xi)\mathbb{I}\big\{\nu_S(\bar{\mathbf{y}}(\xi)) \le \alpha_t\big\}\\ &= \beta\sum_{\xi\in[m]} \sqrt{d}\eta_{\mathcal{Y}}\mu(\xi)\mathbb{I}\big\{\nu_S(\bar{\mathbf{y}}(\xi)) \le \alpha_t\big\}\\ &= \beta\sqrt{d}\eta_{\mathcal{Y}}\mathbb{P}\big(\nu_S(\bar{\mathbf{y}}(\xi)) \le \alpha_t\big)\\ &\le \beta\sqrt{d}\eta_{\mathcal{Y}}\Psi(\alpha_t).\end{aligned}$$

The last inequality is by the definition of function Ψ. Replacing α_t with $4\eta_{\mathcal{Y}}\sqrt{\frac{2d\ln(t/\delta)}{p_{\mathsf{min}}\mu(\xi)t}}$, The second upper bounds can be written as:

$$\sum_{\xi\in[m]} \mathbb{E}[U_M(\xi, h_t(\xi), \rho_t(\xi))|\xi] \le \beta\sqrt{d}\eta_{\mathcal{Y}}\Psi\Big(4\eta_{\mathcal{Y}}\sqrt{\frac{2d\ln(t/\delta)}{p_{\mathsf{min}}\mu(\xi)t}}\Big). \tag{B.5}$$

□

Proof of Theorem 3.3.2. By the definition of value of one data point, we have $V(\xi_t; \mathcal{S}_{t-1}) = \beta \cdot \mathsf{Regret}(h_{t-1}) - \beta \cdot \mathbb{E}_{y_t}\left[\mathsf{Regret}\left(f\left(\mathcal{S}_{t-1} \bigcup \{(\xi_t, \mathbf{y}_t)\}\right)\right) \middle| \xi_t\right]$. Since we assume the prediction models for different features are independent, the value of one data point can be reduced to

$$V(\xi_t; \mathcal{S}_{t-1}) = \beta\mu(\xi_t)\ell(h_{t-1}(\xi_t), \mathbb{E}[y|\xi_t]) - \beta\mu(\xi_t) \cdot \mathbb{E}_{y_t}\left[\ell(h_t(\xi_t), \mathbb{E}[y|\xi_t])\right].$$

In the last term, the randomness stems from possible new predictor h_t, which depends on the possible label outcome y_t. Thus, to prove Theorem 3.3.2, it suffices to show that there exists a distribution $\mathcal{D}$ and constant $K > 0$, such that

$$\ell(h_{t-1}(\xi_t), \mathbb{E}[y|\xi_t]) - \mathbb{E}_{y_t}\left[\ell(h_t(\xi_t), \mathbb{E}[y|\xi_t])\right] \geq K\rho_t(\xi_t)\mathbb{I}\{\nu_S(h_{t-1}(\xi_t)) \leq \rho_{t-1}(\xi_t)\}. \tag{B.6}$$

For the left-hand side,

$$\begin{aligned}
&\ell(h_{t-1}(\xi_t), \mathbb{E}[y|\xi_t]) - \mathbb{E}_{y_t}\left[\ell(h_t(\xi_t), \mathbb{E}[y|\xi_t])\right] \\
&= \mathbb{E}[y|\xi_t]^T(w^*(\mathbb{E}[y|\xi_t]) - w^*(h_{t-1}(\xi_t))) - \mathbb{E}[y|\xi_t]^T(w^*(\mathbb{E}[y|\xi_t]) - \mathbb{E}[w^*(h_t(\xi_t))]) \\
&= \mathbb{E}[y|\xi_t]^T(\mathbb{E}[w^*(h_t(\xi_t))] - w^*(h_{t-1}(\xi_t))).
\end{aligned}$$

We assume that the noise distribution satisfies the separability condition: For any outcome of ξ and y, the following condition holds:

$$\nu_S(\|\mathbb{E}[y|\xi] - y\|) < \nu_S(\mathbb{E}[y|\xi]).$$

This separability condition is studied in M. Liu et al., 2023 and implies the optimal Bayesian expected loss of the true predictor is zero. Because the prediction is determined by the average of the labeled data points, we have that under the separability condition, $\mathbb{E}[w^*(h_t(\xi_t))] = w^*(h_t(\mathbb{E}[y|\xi_t]))$.

To prove (B.6), we consider two cases depending on $\mathbb{I}\{\nu_S(h_{t-1}(\xi_t)) \leq \rho_{t-1}(\xi_t)\}$. When $\nu_S(h_{t-1}(\xi_t)) > \rho_{t-1}(\xi_t)$, by the upper bounds in Theorems 3.4.1 and 3.3.1, we have that the left-hand side of (B.6) is

$$\mathbb{E}[y|\xi_t]^T(\mathbb{E}[w^*(h_t(\xi_t))] - w^*(h_{t-1}(\xi_t))) = \mathbb{E}[y|\xi_t]^T(\mathbb{E}[w^*(E[y|\xi_t])] - w^*(h_{t-1}(\xi_t))) = 0.$$

Thus, both sides of (B.6) are zero.

Next, we consider the case where $\nu_S(h_{t-1}(\xi_t)) \leq \rho_{t-1}(\xi_t)$. In this case, we construct the distribution of $\mathbb{E}[y|\xi]$ further satisfies the following condition: There exists $K_1 > 0$, such that for any $\xi \in \mathcal{X}$, $w_0 \neq w^*(\mathbb{E}[y|\xi])$, we have

$$\mathbb{E}[y|\xi]^T(w^*(\mathbb{E}[y|\xi]) - w_0) \geq K_1.$$

Intuitively, K_1 represents the minimum revenue gap between the best assortment (recommendation) and the second-best assortment (recommendation) for any feature ξ.

Thus, the left-hand side of (B.6) is no less than K_1. Since $\rho_{t-1}(\xi_t) \leq \eta_{\mathcal{Y}}$, we have that

$$\mathbb{E}[y|\xi_t]^T(\mathbb{E}[w^*(E[y|\xi_t])] - w^*(h_{t-1}(\xi_t))) \cdot \frac{\eta_{\mathcal{Y}}}{K_1} \geq \rho_{t-1}(\xi_t).$$

Thus, (B.6) holds, and we obtain Theorem 3.3.2. □

B.2 Proofs in Section 3.4

Before proving Lemma 3.4.1, we first prove that the optimal revenue $g_a(u)$ is lipschitz with u in Lemma B.2.1.

Lemma B.2.1. *If $|u_i - u_i'| \le \varepsilon, \forall i \in [d]$, then the value $|g_a^*(u) - g_a^*(u')| \le z\eta_p\varepsilon$.*

Proof of Lemma B.2.1. We use $w^*(u)$ to denote $\arg\max_{w^T 1=z} : g(u, w)$. First, by Lemma A.3 in Agrawal et al., 2019, for any vector $\epsilon \ge 0$, we have that $g_a(u+\epsilon, w^*(u)) \ge g_a(u, w^*(u))$.

Therefore, if $0 \le u_i' - u_i \le \varepsilon, \forall i \in [d]$, then we have that

$$g_a^*(u') \le g_a(u+\varepsilon, w^*(u))$$

Considering the right hand side,

$$\begin{aligned} g_a(u+\varepsilon, w^*(u)) &= \frac{\sum_{i\in[d]}(u_i+\varepsilon)p_i w_i^*(u)}{1+(u+\varepsilon)^T w} \\ &= \frac{\sum_{i\in[d]} u_i p_i w_i^*(u) + \varepsilon\sum_{i\in[d]} p_i w_i^*(u)}{1+u^T w + z\varepsilon} \\ &\le \frac{\sum_{i\in[d]} u_i p_i w_i^*(u) + \varepsilon z\eta_p}{1+u^T w + z\varepsilon} \end{aligned}$$

Since $\eta_p \ge p_i$, we have

$$\frac{\sum_{i\in[d]} u_i p_i w_i^*(u)}{1+u^T w} \le \frac{\eta_p \sum_{i\in[d]} u_i w_i^*(u)}{1+u^T w} \le \eta_p.$$

Then, we define a temporary function $f(x) = \frac{a_1+a_2 x}{a_3+x}$, where $\frac{a_1}{a_3} \le a_2$, $a_1, a_2, a_3, x \ge 0$. Then, $f(x) \le \frac{a_1}{a_3} + \frac{a_2 x}{a_3+x}$. Then, we have that $f(x) \le \frac{a_1}{a_3} + \frac{a_2 x}{a_3}$. Returning to our setting, we replace a_1, a_2, a_3, x with $\sum_{i\in[d]} u_i p_i w_i^*(u), z\eta_p, 1+u^T w, \varepsilon$, we have that

$$\frac{\sum_{i\in[d]} u_i p_i w_i^*(u) + \varepsilon z\eta_p}{1+u^T w + z\varepsilon} \le \frac{\sum_{i\in[d]}(u_i)p_i w_i^*(u)}{1+u^T w} + \frac{z\eta_p}{1+u^T w}\varepsilon = g_a^*(u) + \frac{z\eta_p}{1+u^T w}\varepsilon \le g_a^*(u) + z\eta_p\varepsilon.$$

Thus, we have if $0 \le u_i' - u_i \le \varepsilon, \forall i \in [d]$,

$$g_a^*(u') \le g_a(u+\varepsilon, w^*(u)) \le g_a^*(u) + z\eta_p\varepsilon. \tag{B.7}$$

Next, we consider the lower bound for $g_a^*(u')$. Again, by Lemma A.3 in Agrawal et al., 2019, for any vector $\epsilon \ge 0$, $g_a^*(u-\epsilon) = g_a(u-\epsilon, w^*(u-\epsilon)) \le g_a(u, w^*(u-\epsilon)) \le g_a(u, w^*(u)) = g_a^*(u)$.

Thus, if $-\varepsilon \le u_i' - u_i \le 0, \forall i \in [d]$, then we have that

$$g_a^*(u') \ge g_a^*(u-\varepsilon). \tag{B.8}$$

In Equation (B.7), replacing u' and u with u and $u-\varepsilon$ respectively, we have that $g_a^*(u) \le g_a^*(u-\varepsilon) + z\eta_p\varepsilon$.

Thus, combining Equation (B.8), we have if $-\varepsilon \le u_i' - u_i \le 0, \forall i \in [d]$, then we have that

$$g_a^*(u') \ge g_a^*(u-\varepsilon) \ge g_a^*(u) - z\eta_p\varepsilon.$$

Therefore, we have if $|u_i - u_i'| \le \varepsilon, \forall i \in [d]$,

$$g_a^*(u) + z\eta_p\varepsilon \ge g_a^*(u') \ge g_a^*(u) - z\eta_p\varepsilon. \tag{B.9}$$

□

Proof of Lemma 3.4.1. Define $\Delta_g = g_a^*(u') - g_a^*(u)$, and $\Delta_u = u - u'$. When $\|\Delta_u\| \le \varepsilon$, by Lemma $B.2.1$, $\Delta_g \in [-z\eta_p\varepsilon, z\eta_p\varepsilon]$. We use $\Delta_{u,i}$ to denote the ith entry of Δ_u.

$$\begin{aligned}
|u_i(p_i - g_a^*(u)) - u_i'(p_i - g_a^*(u'))| &= |u_i(p_i - g_a^*(u)) - (u_i + \Delta_{u,i})(p_i - \Delta_g - g_a^*(u))| \\
&= |-\Delta_g u_i - \Delta_{u,i}(p_i - \Delta_g - g_a^*(u))| \\
&= |\Delta_g u_i + \Delta_{u,i}(p_i - \Delta_g - g_a^*(u))| \\
&\le |\Delta_g u_i + |\Delta_{u,i}|\eta_p| && \text{(B.10)} \\
&\le |z\eta_p\varepsilon u_i + |\Delta_{u,i}|\eta_p| && \text{(B.11)} \\
&\le |z\eta_p\varepsilon e^{\eta_{\mathcal{Y}}/\sigma} + \varepsilon\eta_p| && \text{(B.12)} \\
&= (z\eta_p e^{\eta_{\mathcal{Y}}/\sigma} + \eta_p)\varepsilon.
\end{aligned}$$

Inequality (B.10) is because the discounted value $p_i - \Delta_g - g_a^*(u)$ is less than η_p. Inequality (B.11) is from the bound of Δ_g and inequality (B.12) is from the bound of Δ_u. □

Proof of Theorem 3.4.1. The proof of Theorem 3.4.1 is from Lemma 3.4.1 and Theorem 3.3.1. By Lemma 3.4.1, when the radius of the confidence region of the utility $\mathbf{y}$ is ρ_t, the radius of the confidence region of the coefficient in the objective (3.6) is $\varkappa\rho_t$. Since Problem (3.6) is in the form of a product selection problem, we can directly use the result of Theorem 3.3.1. Since the value of one data point is also upper bounded by the maximum revenue loss, which is η_p, we obtain the upper bound in Theorem 3.4.1. □

B.3 Proofs in Sections 3.5 and 3.6

Proof of Theorem 3.5.1. First, we prove Claim *3.5.1.(1)*. Since $U(\xi_t; \mathcal{S}_{t-1})$ and $p(c)$ are universal for all $\xi \in [m]$, the offered incentive $c^*(U(\xi_t; \mathcal{S}_{t-1}), p)$ at each iteration t is universal for all types of customers. Hence, the probabilities of taking the survey are the same for all types. Thus, training set $\mathcal{S}_t$ can be viewed as a data set with i.i.d. samples.

We consider two cases: (1) $\beta\mathsf{Regret}(h_T) \ge c_{\mathsf{min}}$ and (2) $\beta\mathsf{Regret}(h_T) < c_{\mathsf{min}}$. When $\beta\mathsf{Regret}(h_T) \ge c_{\mathsf{min}}$, since $U(\xi_{T+1}; \mathcal{S}_T) \ge \beta\mathsf{Regret}(h_T)$, by the non-increasing condition of $U(\xi_{T+1}; \mathcal{S}_T)$, we have

$$U(\xi_T; \mathcal{S}_{T-1}) \ge U(\xi_{T+1}; \mathcal{S}_T) \ge \beta\mathsf{Regret}(h_T) \ge c_{\mathsf{min}}.$$

Thus, we offer incentives at least c_{min}, which means customers take the survey with probability at least p_{min}. Since each customer is independent, we can apply Chernoff's bound, and obtain that with probability at least $1 - e^{-p_{\mathsf{min}}T/8}$, the size of the training set is at least $\frac{p_{\mathsf{min}}T}{2}$. with at least $\frac{p_{\mathsf{min}}T}{2}$ samples. Thus, the risk of h_T is at most $\beta R_{\mathsf{s}}(p_{\mathsf{min}}T/2)$.

On the other hand, if $\beta\mathsf{Regret}(h_T) < c_{\mathsf{min}}$, we simply use c_{min} as the upper bound for the risk of h_T. Combining the upper bounds of these two cases, we have that with probability at least $1 - e^{-p_{\mathsf{min}}T/8}$,

$$\beta\mathsf{Regret}(h_T) \leq \beta R_{\mathsf{s}}(p_{\mathsf{min}}T/2) + c_{\mathsf{min}}.$$

Next, we prove Claim *3.5.1.(2)*. Since $U(\xi_t; \mathcal{S}_{t-1})$ is a non-increasing function, when $t \geq \underline{t}$, $U(\xi_t; \mathcal{S}_{t-1}) < c_{\mathsf{min}}$, which means we offer 0 incentives. Since the maximum incentive we offer is c_{max}, the cumulative label cost is at most $\min\{\underline{t}, T\}c_{\mathsf{max}}$. □

Proof of Proposition 3.5.1. We denote $p(c)\,(c - U)$ by $\mathtt{f}(c)$. Then, we have $\mathtt{f}'(c) = p'(c)[c - U] + p(c)$. We use $\tilde{c}^*$ to denote the minimizer of $f(c)$. By setting $\mathtt{f}'(c) = 0$, we have that $U = \tilde{c}^* + \frac{p(\tilde{c}^*)}{p'(\tilde{c}^*)}$.

Since $\mathtt{f}'(c) = p''(c)[c - U] + 2p'(c) \geq 0$ for all $c \leq U$, we have that $\tilde{c}^*$ is the minimizer of $\mathtt{f}(c)$. The derivative of $\tilde{c}^* + \frac{p(\tilde{c}^*)}{p'(\tilde{c}^*)}$ is

$$1 + \frac{(p'(\tilde{c}^*))^2 - p(\tilde{c}^*)p''(\tilde{c}^*)}{(p'(\tilde{c}^*))^2} = \frac{2(p'(\tilde{c}^*))^2 - p(\tilde{c}^*)p''(\tilde{c}^*)}{(p'(\tilde{c}^*))^2} \geq 0$$

The last inequality is by $p''(\tilde{c}^*) \leq 0$. Thus, $\tilde{c}^* + \frac{p(\tilde{c}^*)}{p'(\tilde{c}^*)}$ is an increasing function of $\tilde{c}^*$. Thus, when U increases, $\tilde{c}^*$ should also increase. Thus, we obtain that $c^*(V(\xi_T; \mathcal{S}_{T-1}), p) \leq c^*(U(\xi_T; \mathcal{S}_{T-1}), p)$.

□

Proof of Propositions 3.5.2 and 3.5.3. We denote $p(c)\,(c - U)$ by $\mathtt{f}(c)$. Then, we have

$$\mathtt{f}'(c) = p'(c)[c - U] + p(c).$$

Suppose $\tilde{c}^*$ is the minimizer of $\mathtt{f}(c)$. By setting $\mathtt{f}'(c) = 0$, we have that

$$U = \tilde{c}^* + \frac{p(\tilde{c}^*)}{p'(\tilde{c}^*)}.$$

Since $\mathtt{f}''(c) = p''(c)[c - U] + 2p'(c) \geq 0$ for all $c \leq U$, we have that $\tilde{c}^*$ is the minimizer of $\mathtt{f}(c)$. We can further check that $\tilde{c}^* + \frac{p(\tilde{c}^*)}{p'(\tilde{c}^*)}$ is an increasing function of $\tilde{c}^*$ by showing that the derivative of $\tilde{c}^* + \frac{p(\tilde{c}^*)}{p'(\tilde{c}^*)}$ is positive.

In the first case of Proposition 3.5.2, we have that $U = \tilde{c}^* + \frac{p_0(\tilde{c}^*)+k_1}{p'(\tilde{c}^*)}$. Thus, when k_1 becomes larger, the right hand side $\tilde{c}^* + \frac{p_0(\tilde{c}^*)+k_1}{p'(\tilde{c}^*)}$ becomes larger. Since the right hand side is also an increasing function of $\tilde{c}^*$, we have that to keep the right hand side equal U, $\tilde{c}^*$ needs to be smaller.

In the second case of Proposition 3.5.2, it is easy to see that $U = \tilde{c}^* + \frac{p_0(\tilde{c}^*)}{p_0'(\tilde{c}^*)}$, which is independent of k_1. Finally, Proposition 3.5.3 can be obtained by solving $U = \tilde{c}^* + \frac{p(\tilde{c}^*)}{p'(\tilde{c}^*)}$ directly. □

Proof of Theorem 3.6.1. Recall that after considering T customers, the number of samples for type ξ is $n_T(\xi)$. In Algorithm 2, the prediction for ξ is the mean of the observation, i.e., for product j, the predicted utility is $h^j(\xi) = \frac{1}{n_T(\xi)} \sum_{(\cdot, y_t^j) \in \mathcal{S}_T(\xi)} y_t^j$. Thus, the radius for the confidence interval at time t is $\rho_t(\xi_t)$. Since h_{T-1} is random, the risk of h_{T-1} can be written as

$$\beta \cdot \mathsf{Regret}(h_{T-1}) = \beta \mathbb{E}[\ell(h_{T-1}(\xi), \mathbb{E}[\mathbf{y}|\xi])] = \sum_{\xi \in [m]} \mu(\xi) \beta \mathbb{E}\left[\ell(h_{T-1}(\xi), \mathbb{E}[\mathbf{y}|\xi])\right].$$

The outer expectation $\mathbb{E}\left[\ell(h_{T-1}(\xi), \mathbb{E}[\mathbf{y}|\xi])\right]$ is taken over the randomness of h_{T-1}. By the upper bound of $V_M^O(\xi_t, \hat{\mathbf{y}}, \rho_{t-1})$ in the proof of Theorem 3.3.1, we have that for each type ξ, given h_{T-1},

$$\begin{aligned}\mu(\xi)\beta\ell(h_{T-1}(\xi), \mathbb{E}[\mathbf{y}|\xi]) &\le \mu(\xi)\beta\sqrt{2\min\{z, d-z\}}\rho_{T-1}(\xi)\mathbb{I}\{\nu_S(h_{T-1}(\xi)) \le \rho_{T-1}(\xi)\} \\ &= U_M(\xi, h_{T-1}(\xi), \rho_{T-1}(\xi)).\end{aligned}$$

Thus, the risk of h_{T-1} satisfies:

$$\beta \cdot \mathsf{Regret}(h_{T-1}) = \sum_{\xi \in [m]} \mu(\xi) \beta \mathbb{E}\left[\ell(h_{T-1}(\xi), \mathbb{E}[\mathbf{y}|\xi])\right] \le \sum_{\xi \in [m]} \mathbb{E}\left[U_M(\xi, h_{T-1}(\xi), \rho_{T-1}(\xi))|\xi\right].$$

Thus, to derive the risk bound for h_{T-1}, it suffices to derive the upper bound for $\sum_{\xi \in [m]} \mathbb{E}\left[U_M(\xi, h_{T-1}(\xi), \rho_{T-1}(\xi))|\xi\right]$. We observe that for a given type ξ, $U_M(\xi, h_{T-1}(\xi), \rho_{T-1}(\xi))$ is a non-increasing sequence.

We define $\underline{t}_\xi$ as the time when the incentive $U_M(\xi, h_{T-1}(\xi), \rho_{T-1}(\xi))$ decreases below c_{min}, i.e., $\underline{t}_\xi = \inf_{t \ge 0}\{U_M(\xi, h_{t-1}(\xi), \rho_{t-1}(\xi)) < c_{\mathsf{min}}\}$. (In the special case where $c_{\mathsf{min}} = 0$, $\underline{t}_\xi = \infty$.) Thus, when $t \ge \underline{t}_\xi$, we stop providing any incentives for type ξ, and thus the upper bound for the value of one data point for type ξ, $U_M(\xi, h_{T-1}(\xi), \rho_{T-1}(\xi))$ remains at $U_M(\xi, h_{\underline{t}_\xi - 1}(\xi), \rho_{\underline{t}_\xi - 1}(\xi))$.

First, we consider the case where $T < \underline{t}_\xi$. In this case, we always provide some incentive to customers, and the customer has a probability at least p_{min} to take the survey. Since customers make independent decisions, after T iterations, the probability of $n_T(\xi) = 0$ is at most

$$(1 - \mu(\xi)p_{\mathsf{min}})^T \le (1 - \underline{\mu}p_{\mathsf{min}})^T = e^{T\ln(1 - p_{\mathsf{min}}\underline{\mu})}.$$

Thus, by the union bound, after T iterations, the probability of $n_T(\xi) > 0$ for all $\xi \in [m]$ is at least

$$1 - me^{T\ln(1 - p_{\mathsf{min}}\underline{\mu})}.$$

It implies that with probability $1 - me^{T\ln(1-p_{\min}\underline{\mu})}$, the conditions in Lemma B.1.1 are satisfied. Thus, combining it with the probability in Lemma B.1.1, we obtain that the two upper bounds in Lemma B.1.1 hold with probability at least $1 - me^{-\frac{\underline{\mu}p_{\min}t}{8}} - me^{t\ln(1-p_{\min}\underline{\mu})} - \delta/t^2$.

Next, we consider the case where $t \geq \underline{t}_\xi$. In this case, $U_M(\xi, h_{T-1}(\xi), \rho_{T-1}(\xi))$ remains at $U_M(\xi, h_{\underline{t}_\xi-1}(\xi), \rho_{\underline{t}_\xi-1}(\xi))$. Its naive upper bound is $c_{\min}$, and next, we derive a tighter bound for it by considering the function Ψ.

Recall that $U_M(\xi, h_t(\xi), \rho_t(\xi)) = \sqrt{2\min\{z, d-z\}}\beta\mu(\xi)\rho_t(\xi)\mathbb{I}\{\nu_S(\hat{\mathbf{y}}) \leq \rho_t(\xi)\}$. Define $\rho_{\tau,1} := \max : \{\rho_t(\xi) : \sqrt{2\min\{z, d-z\}}\mu(\xi)\rho_t(\xi) \leq c_{\min}\}$. Define $\rho_{\tau,2} := \max\{\rho_t(\xi) : \nu_S(\bar{\mathbf{y}}) \geq 2\rho_t(\xi)\}$. As shown earlier, $\nu_S(\bar{\mathbf{y}}) \geq 2\rho_t(\xi)$ implies $\nu_S(\hat{\mathbf{y}}) \geq \rho_t(\xi)$. Thus, $\rho_t(\xi) \geq \max\{\rho_{\tau,1}, \rho_{\tau,2}\}$, $\forall t = 1, 2, 3, ...$

Therefore, if $\rho_{\tau,1} \geq \rho_{\tau,2}$, the final risk at ξ is at most $c_{\min}$. If $\rho_{\tau,1} \leq \rho_{\tau,2}$, the final risk at ξ is 0.

Therefore the final upper bound of value of one data point at ξ is equal to

$$c_{\min}\mathbb{I}\Big\{\rho_{\tau,1} \geq \rho_{\tau,2}\Big\} \leq c_{\min}\mathbb{I}\Big\{\frac{c_{\min}}{\sqrt{2\min\{z, d-z\}}\mu(\xi)\beta} \geq \frac{\nu_S(\bar{\mathbf{y}})}{2}\Big\}.$$

Thus, the final risk in total is no more than

$$\begin{aligned}
\sum_{\xi\in[m]} c_{\min}\mathbb{I}\Big\{\rho_{\tau,1} \geq \rho_{\tau,2}\Big\} &\leq \frac{c_{\min}}{\underline{\mu}}\sum_{x\in[m]}\mu(\xi)\mathbb{I}\Big\{\rho_{\tau,1} \geq \rho_{\tau,2}\Big\} \\
&= \frac{c_{\min}}{\underline{\mu}}\mathbb{P}\Big(\rho_{\tau,1} \geq \rho_{\tau,2}\Big) \\
&= \frac{c_{\min}}{\underline{\mu}}\mathbb{P}\Big(\frac{c_{\min}}{\sqrt{2\min\{z, d-z\}}\mu(\xi)\beta} \geq \frac{\nu_S(\bar{\mathbf{y}})}{2}\Big) \\
&\leq \frac{c_{\min}}{\underline{\mu}}\Psi\Big(\frac{\sqrt{2}c_{\min}}{\sqrt{\min\{z, d-z\}\underline{\mu}\beta}}\Big).
\end{aligned}$$

The inequalities are because $\underline{\mu} \leq \mu(\xi)$, $\forall\xi \in [m]$.

Therefore, combining results of the final risk with the upper bounds in Lemma B.1.1, we have that with probability at least $1 - me^{-\frac{\underline{\mu}p_{\min}t}{8}} - me^{t\ln(1-p_{\min}\underline{\mu})} - \delta/t^2$, the total risk of the predictor h_t, is at most $\beta\mathcal{R}(t, c_{\min})$, where

$$\mathcal{R}(T, c_{\min}) := \varphi(T) + \frac{c_{\min}}{\beta\underline{\mu}}\Psi\Big(\frac{\sqrt{2}c_{\min}}{\sqrt{\min\{z, d-z\}\underline{\mu}\beta}}\Big).$$

The function $\varphi(T)$ is defined as

$$\varphi(T) := \eta_{\mathcal{Y}}\min\left\{\sum_{\xi\in[m]} 4\mu(\xi)\sqrt{\min\{z, d-z\}}\sqrt{\frac{d\ln(T/\delta)}{p_{\min}\mu(\xi)T}}, \sum_{\xi\in[m]}\sqrt{d}\Psi\Big(4\eta_{\mathcal{Y}}\sqrt{\frac{2d\ln(T/\delta)}{p_{\min}\mu(\xi)T}}\Big)\right\}.$$

Thus, we obtain the result in Theorem 3.6.1.

□

Proof of Theorem 3.6.2. To derive the upper bound for the cumulative incentives, we first claim that $c^*(U(\xi_T;\mathcal{S}_{T-1}),p) \le U(\xi_T;\mathcal{S}_{T-1})$ for all $T \ge 1$. To prove this, we observe that by the proof of Proposition 3.5.1, the minimum value of $p(c)(c-U(\xi_T;\mathcal{S}_{T-1}))$ should be negative, i.e., $c^*(U(\xi_T;\mathcal{S}_{T-1}),p) - U(\xi_T;\mathcal{S}_{T-1}) < 0$. Therefore, we conclude that $c^*(U(\xi_T;\mathcal{S}_{T-1}),p) \le U(\xi_T;\mathcal{S}_{T-1})$. Consequently, the incentive at each iteration is at most $U_M(\xi, h_{t-1}(\xi), \rho_{t-1}(\xi))$, the expectation of the label cost at iteration t is at most

$$\sum_{\xi\in[m]} \mu(\xi) U_M(\xi, h_{t-1}(\xi), \rho_{t-1}(\xi)).$$

To derive the upper bound for the expected cumulative incentive, we derive upper bounds for $\sum_{\xi\in[m]} \mu(\xi)\mathbb{E}[U_M(\xi, h_{t-1}(\xi), \rho_{t-1}(\xi))|\xi]$ for each iteration.

In Lemma B.1.1, we have two upper bounds for $\mathbb{E}[U_M(\xi, h_{t-1}(\xi), \rho_{t-1}(\xi))|\xi]$ with some high probability. To derive the upper bound for the cumulative incentives, let us first consider the case that (B.3) and (B.5) in Lemma B.1.1 hold for $t = 1, ..., T$. To begin, let us consider (B.3), which implies that for all $t \ge 1$,

$$\begin{aligned}
\sum_{\xi\in[m]} \mu(\xi)\mathbb{E}[U_M(\xi, h_t(\xi), \rho_t(\xi))|\xi] &\le \sum_{\xi\in[m]} 4\mu^2(\xi)\sqrt{\min\{z, d-z\}}\eta_{\mathcal{Y}}\sqrt{\frac{d\ln(t/\delta)}{p_{\min}\mu(\xi)t}} \\
&\le \sum_{\xi\in[m]} 4\mu^2(\xi)\sqrt{\min\{z, d-z\}}\eta_{\mathcal{Y}}\sqrt{\frac{d\ln(t/\delta)}{p_{\min}\underline{\mu}t}} \\
&\le 4\sqrt{\min\{z, d-z\}}\eta_{\mathcal{Y}}\sqrt{\frac{d\ln(t/\delta)}{p_{\min}\underline{\mu}t}} \sum_{\xi\in[m]} \mu(\xi) \\
&= 4\sqrt{\min\{z, d-z\}}\eta_{\mathcal{Y}}\sqrt{\frac{d\ln(t/\delta)}{p_{\min}\underline{\mu}t}}.
\end{aligned}$$

The second inequality is because $\mu(\xi) \ge \underline{\mu}$ for all ξ and the third inequality is because $\mu^2(\xi) \le \mu(\xi)$. Then, summing this upper bound over $t = 1, ..., T$, and using the inequality that $\sum_{t=1}^{T}\sqrt{1/t} \le 2\sqrt{T}$, we have that with the same probability, the expected cumulative incentives are at most

$$\begin{aligned}
\sum_{t=1}^{T}\sum_{\xi\in[m]} \mu(\xi)\mathbb{E}[U_M(\xi, h_t(\xi), \rho_t(\xi))|\xi] &\le \sum_{t=1}^{T} 4\sqrt{\min\{z, d-z\}}\eta_{\mathcal{Y}}\sqrt{\frac{d\ln(T/\delta)}{p_{\min}\underline{\mu}t}} \\
&\le 8\sqrt{\min\{z, d-z\}}\eta_{\mathcal{Y}}\sqrt{\frac{dT\ln(T/\delta)}{p_{\min}\underline{\mu}}}.
\end{aligned}$$

This is the first upper bound for the cumulative label cost. Next, we consider (B.5), which

implies that for all $t \leq T$,

$$\sum_{\xi \in [m]} \mu(\xi) \mathbb{E}[U_M(\xi, h_t(\xi), \rho_t(\xi))|\xi] \leq \sum_{\xi \in [m]} \mu(\xi) \sqrt{d} \eta_{\mathcal{Y}} \Psi\Big(4\eta_{\mathcal{Y}} \sqrt{\frac{2d \ln(t/\delta)}{p_{\mathsf{min}} \mu(\xi) t}}\Big).$$

Using the fact that $\mu(\xi) \geq \underline{\mu}$, this upper bound is no more than

$$\begin{aligned}\sum_{\xi \in [m]} \mu(\xi) \sqrt{d} \eta_{\mathcal{Y}} \Psi\Big(4\eta_{\mathcal{Y}} \sqrt{\frac{2d \ln(t/\delta)}{p_{\mathsf{min}} \underline{\mu} t}}\Big) &= \sqrt{d} \eta_{\mathcal{Y}} \Psi\Big(4\eta_{\mathcal{Y}} \sqrt{\frac{2d \ln(t/\delta)}{p_{\mathsf{min}} \underline{\mu} t}}\Big) \sum_{\xi \in [m]} \mu(\xi) \\ &\leq \sqrt{d} \eta_{\mathcal{Y}} \Psi\Big(4\eta_{\mathcal{Y}} \sqrt{\frac{2d \ln(t/\delta)}{p_{\mathsf{min}} \underline{\mu} t}}\Big).\end{aligned}$$

Summing this upper bound over $t = 1, ..., T$, we have that with the same probability, the expected cumulative incentives are at most

$$\sum_{t=1}^{T} \sum_{\xi \in [m]} \mu(\xi) \mathbb{E}[U_M(\xi, h_t(\xi), \rho_t(\xi))|\xi] \leq \sqrt{d} \eta_{\mathcal{Y}} \sum_{t=1}^{T} \Psi\Big(4\eta_{\mathcal{Y}} \sqrt{\frac{2d \ln(t/\delta)}{p_{\mathsf{min}} \underline{\mu} t}}\Big).$$

Taking the minimum of these two upper bounds, we have that conditional on (B.3) and (B.5) are true, the expected cumulative incentive is at most

$$\min \left\{ 8\sqrt{\min\{z, d-z\}} \eta_{\mathcal{Y}} \sqrt{\frac{dT \ln(T/\delta)}{p_{\mathsf{min}} \underline{\mu}}}, \sqrt{d} \eta_{\mathcal{Y}} \sum_{t=1}^{T} \Psi\Big(4\eta_{\mathcal{Y}} \sqrt{\frac{2d \ln(T/\delta)}{p_{\mathsf{min}} \underline{\mu} t}}\Big) \right\}. \tag{B.13}$$

If (B.3) or (B.5) is not true for iteration t, then the label cost is at most c_{max}. The probability that (B.3) or (B.5) is not true is at most $me^{-\frac{\underline{\mu} p_{\mathsf{min}} t}{8}} + me^{t \ln(1 - p_{\mathsf{min}} \underline{\mu})} + \delta/t^2$. Thus, if (B.3) or (B.5) is not true for some iterations before iteration T, then the cumulative cost is at most

$$c_{\mathsf{max}} \sum_{t=1}^{T} \left[me^{-\frac{\underline{\mu} p_{\mathsf{min}} t}{8}} + me^{t \ln(1 - p_{\mathsf{min}} \underline{\mu})} + \delta/t^2 \right].$$

The summation of the first and second terms is $c_{\mathsf{max}} \sum_{t=1}^{T} \left[me^{-\frac{\underline{\mu} p_{\mathsf{min}} t}{8}} + me^{t \ln(1 - p_{\mathsf{min}} \underline{\mu})} \right]$. As the integration $\int_t te^{-t} < \infty$, it is obvious that this summation is finite. We use a constant $c_q > 0$ to denote the upper bound for $c_{\mathsf{max}} \sum_{t=1}^{T} \left[me^{-\frac{\underline{\mu} p_{\mathsf{min}} t}{8}} + me^{t \ln(1 - p_{\mathsf{min}} \underline{\mu})} \right]$. Thus, the cumulative cost is at most

$$c_q + c_{\mathsf{max}} \sum_{t=1}^{T} \left[\delta/t^2\right] \leq c_q + c_{\mathsf{max}} \sum_{t=1}^{\infty} \left[\delta/t^2\right] \leq c_q + c_{\mathsf{max}}\ \delta \cdot \frac{\pi^2}{6} \leq c_q + 2\delta.$$

Finally, combining this bound with the previous bound (B.13), we have that the expectation of the cumulative incentives at iteration T is at most

$$c_q + 2\delta + \min\left\{8\sqrt{\min\{z, d-z\}}\eta_{\mathcal{Y}}\sqrt{\frac{dT\ln(T/\delta)}{p_{\mathsf{min}}\underline{\mu}}}, \sqrt{d}\eta_{\mathcal{Y}}\sum_{t=1}^{T}\Psi\Big(4\eta_{\mathcal{Y}}\sqrt{\frac{2d\ln(T/\delta)}{p_{\mathsf{min}}\underline{\mu}t}}\Big)\right\},$$

which is the result in Theorem 3.6.2.

Next, we consider the case when $c_{\mathsf{min}} > 0$. For each type ξ, by utilizing the notations of $\underline{t}_\xi$ in the proof of Theorem 3.6.1, we have that when $t \geq \underline{t}_\xi$, we stop exploring type ξ and thus the cumulative label cost is finite. As an interest, we derive an upper bound for this finite label cost. Since $\rho_T(\xi) = 2\eta_{\mathcal{Y}}\sqrt{\frac{d\ln(T/\delta)}{n_T(\xi)}}$, we have

$$\begin{aligned}
&\mathbb{P}\big(\beta\sqrt{2\min\{z, d-z\}}\mu(\xi_t)\rho_T(\xi_t) \geq c_{\mathsf{min}}\big)\\
&= \mathbb{P}\left(2\beta\sqrt{2\min\{z, d-z\}}\mu(\xi_t)\eta_{\mathcal{Y}}\sqrt{\frac{d\ln(T/\delta)}{n_T(\xi_t)}} \geq c_{\mathsf{min}}\right)\\
&= \mathbb{P}\left(\frac{8\beta^2\min\{z, d-z\}d\mu^2(\xi_t)\eta_{\mathcal{Y}}^2\ln(T/\delta)}{c_{\mathsf{min}}^2} \geq n_T(\xi_t)\right)\\
&= \sum_{\xi\in[m]}\mu(\xi)\mathbb{P}\left(\frac{8d\beta^2\min\{z, d-z\}\mu^2(\xi)\eta_{\mathcal{Y}}^2\ln(T/\delta)}{c_{\mathsf{min}}^2} \geq n_T(\xi)\right)
\end{aligned}$$

Since the probability of taking the survey is at least $p_{\mathsf{min}}\mu(\xi)$, by Chernoff's inequality, we have that

$$\mathbb{P}\left(\frac{8d\beta^2\min\{z, d-z\}\mu^2(\xi)\eta_{\mathcal{Y}}^2\ln(T/\delta)}{c_{\mathsf{min}}^2} \geq n_T(\xi)\right) \leq e^{-\alpha(\xi,T)^2/(2T)},$$

where $\alpha(\xi, T) := \max\big\{0, p_{\mathsf{min}}\mu(\xi)T - 8\beta^2\min\{z, d-z\}\mu^2(\xi)\eta_{\mathcal{Y}}^2\ln(T/\delta)c_{\mathsf{min}}^{-2}\big\}$.

Thus, when $c_{\mathsf{min}} > 0$, the cumulative label cost by time t is at most

$$\begin{aligned}
c_{\mathsf{max}}\sum_{t=1}^{T}\sum_{\xi\in[m]}\mu(\xi)\mathbb{P}\left(\frac{8d\beta^2\min\{z, d-z\}\mu^2(\xi)\eta_{\mathcal{Y}}^2\ln(T/\delta)}{c_{\mathsf{min}}^2} \geq n_T(\xi)\right)&\\
\leq c_{\mathsf{max}}\sum_{t=1}^{T}\sum_{\xi\in[m]}\mu(\xi)e^{-\alpha(\xi,T)^2/(2T)}.&
\end{aligned}$$

□

Proof of Theorem 3.6.3. We first prove Theorem 3.6.3.(1) in Theorem 3.6.3. Since $c_{\mathsf{min}} = c_{\mathsf{max}}$, our offered incentive is either 0 or c_{min}. When $U_M(\xi, h_{t-1}(\xi) \geq c_{\mathsf{min}}$, Algorithm 2 provides c_{min} to the customer, which is the same as the supervised learning algorithm. Thus, it suffices to consider the case where $U_M(\xi, h_{t-1}(\xi) < c_{\mathsf{min}}$. In this case, Algorithm 2 does

not provide any incentive, so the change of the comprehensive cost is zero. However, for the supervised learning algorithm, if the customer does not accept the incentive, the change of the comprehensive cost is zero. If the customer accepts the incentive, the risk reduction $R(h_{t-1})-R(h_t)$ is at most $\mu(\xi_t)\beta V_M^O(\xi_t, \hat{\mathbf{y}}, \rho_{t-1})$, which is defined in the proof of Theorem 3.3.1. By the proof of Theorem 3.3.1, $U_M(\xi, h_{t-1}(\xi), \rho_{t-1}(\xi))$ is no less than $\mu(\xi_t)\beta V_M^O(\xi_t, \hat{\mathbf{y}}, \rho_{t-1})$, so the risk reduction $R(h_{t-1}) - R(h_t)$ is smaller than $U_M(\xi, h_{t-1}(\xi), \rho_{t-1}(\xi))$. Therefore, the change of the comprehensive cost is at least $c_{\mathsf{min}} - U_M(\xi, h_{t-1}(\xi), \rho_{t-1}(\xi)) > 0$. Thus, by combining these two cases, the change of the supervised learning algorithm is larger than or equal to the change of Algorithm 2. Thus, Algorithm 2 always achieves a smaller or equivalent comprehensive cost than the supervised learning algorithm.

Next, we prove Theorem 3.6.3.(2). We denote $p(c)\,(c-U)$ by $\mathtt{f}_U(c)$. By the proof of Proposition 3.5.2, we have

$$\mathtt{f}'_U(c) = p'(c)[c-U] + p(c).$$

Suppose $\tilde{c}^*$ is the minimizer of $\mathtt{f}_U(c)$. By setting $\mathtt{f}'_U(c) = 0$, we have that

$$U = \tilde{c}^* + \frac{p(\tilde{c}^*)}{p'(\tilde{c}^*)}.$$

Since $\mathtt{f}''_U(c) = p''(c)[c-U] + 2p'(c) \geq 0$ for all $c \leq U$, we have that $\tilde{c}^*$ is the minimizer of $\mathtt{f}_U(c)$. Let us consider two cases: (1) $\tilde{c}^* \geq c_{\mathsf{max}}$, and (2) $\tilde{c}^* < c_{\mathsf{max}}$.

In the first case, since $\mathtt{f}''_U(c) \geq 0$, for any $c \in [c_{\mathsf{min}}, c_{\mathsf{max}}]$, we have $\mathtt{f}'_U(c) \leq \mathtt{f}'_U(c_{\mathsf{max}}) \leq \mathtt{f}'_U(\tilde{c}^*) = 0$. Thus, the optimal incentive $c^*(U_M(\xi, h_{t-1}(\xi), \rho_{t-1}(\xi)), p)$ is c_{max}. It is the same as the supervised learning algorithm that offers c_{max} all the time.

In the second case, by Proposition 3.5.1, we have that $c^*(V(\xi, h_{t-1}(\xi), \rho_{t-1}(\xi)), p) \leq c^*(U_M(\xi, h_{t-1}(\xi), \rho_{t-1}(\xi)), p) \leq c_{\mathsf{max}}$. Since $\mathtt{f}'_V(c) \geq 0$ when $c \geq c^*(V(\xi, h_{t-1}(\xi), \rho_{t-1}(\xi)), p)$, we have that $\mathtt{f}_V(c^*(U_M(\xi, h_{t-1}(\xi), \rho_{t-1}(\xi)), p)) \leq \mathtt{f}_V(c_{\mathsf{max}})$. Thus, the expected change of the comprehensive cost of Algorithm 2 is no larger than the supervised learning that offers c_{max} all the time.

Combining these two cases, we conclude that the expected comprehensive cost of Algorithm 2 is no larger than the expected comprehensive cost of the fixed incentive at c_{max}.

Lastly, we prove the third argument. Since $c_{\mathsf{min}} > 0$, by Theorem 3.6.2, we have that the cumulative label cost of Algorithm 2 is finite. Thus, the comprehensive cost of Algorithm 2 is finite. In contrast, for the supervised learning, since the designer offers a fixed incentive between $[c_{\mathsf{min}}, c_{\mathsf{max}}]$, the expectation of cumulative label cost is at least $Tp_{\mathsf{min}}c_{\mathsf{min}} = \mathcal{O}(T)$. Thus, there exists a time point $T_s > 0$: when the time $T > T_s$, the comprehensive cost $\mathscr{C}(\mathbf{c}, g)$ of Algorithm 2 is no more than the cost of supervised learning.

□

Proof of Proposition 3.6.1. We first prove the bound for the cumulative label cost. By Theorem 3.6.2, the cumulative label cost after surveying T customers is at most

$$c_q + 2\delta + \sqrt{d}\eta_{\mathcal{Y}} \sum_{t=1}^{T} \Psi\Big(4\eta_{\mathcal{Y}} \sqrt{\frac{2d\ln(T/\delta)}{p_{\mathsf{min}}\underline{\mu}t}}\Big).$$

Since $\Psi(\rho) \leq \tilde{\mathcal{O}}(\rho^{\kappa})$, the cumulative label cost is at most

$$\tilde{\mathcal{O}}\left(\sum_{t=1}^{T}\Psi\Big(4\eta_{\mathcal{Y}}\sqrt{\frac{2d\ln(T/\delta)}{p_{\mathsf{min}}\underline{\mu}t}}\Big)\right) \leq \tilde{\mathcal{O}}\left(\sum_{t=1}^{T}\Big(4\eta_{\mathcal{Y}}\sqrt{\frac{2d\ln(t/\delta)}{p_{\mathsf{min}}\underline{\mu}t}}\Big)^{\kappa}\right) \leq \tilde{\mathcal{O}}\left(\sum_{t=1}^{T}t^{-\kappa/2}\right).$$

Since $\tilde{\mathcal{O}}\left(\sum_{t=1}^{T}t^{-\kappa/2}\right) \leq \tilde{\mathcal{O}}\left(\int_{t=1}^{T}t^{-\kappa/2}dt\right) \leq \tilde{\mathcal{O}}(T^{1-\kappa/2})$, we obtain that the cumulative label cost after iteration T is at most $\tilde{\mathcal{O}}(T^{1-\kappa/2})$.

Next, we prove the bound for the risk of the model. By Theorem 3.6.1, the risk of the model h_T is at most $\mathcal{R}(T, c_{\mathsf{min}})$. When Assumption 2.5.1 holds, we have $\varphi(T) \leq \tilde{\mathcal{O}}(\Psi(T^{-\kappa/2}))$. Thus, we have that $\mathcal{R}(T, c_{\mathsf{min}}) \leq \tilde{\mathcal{O}}(T^{-\kappa/2}) + \frac{c_{\mathsf{min}}}{\beta\underline{\mu}}\Psi\Big(\frac{\sqrt{2}c_{\mathsf{min}}}{\beta\underline{\mu}\sqrt{\min\{z,d-z\}}}\Big)$. Since the risk of the predictor h_T is at most $\beta\mathcal{R}(T, c_{\mathsf{min}})$, we obtain the upper bound for the risk in Proposition 3.6.1. □

Proof of Theorem 3.6.4. First, we prove Theorem 3.6.4.(1). It follows Theorem 3.4.1 and Theorem 3.6.1. Similar to Theorem 3.6.1, we first consider the case when $t \leq \underline{t}_{\xi}$. It suffices to derive upper bounds for $\mathbb{E}[U_M^A(\xi_t, \hat{\mathbf{y}}, H_t)|\xi_t] = \mathbb{E}[\min\Big\{\varkappa\sqrt{2\min\{z, d-z\}}\mu(\xi_t)\rho_t\mathbb{I}\{\nu_S(\hat{\mathbf{y}}) \leq \rho_t\}, \eta_p\Big\}|\xi_t]$. The first upper bound in $\varphi(T)$ in Theorem 3.6.4.(1) is immediately obtained when multiplying (B.3) in the proof Theorem 3.6.1 by κ. The second upper bound is obtained by replacing the maximum decision loss $\sqrt{d}\eta_{\mathcal{Y}}$ in (B.5) with η_p.

Next, we consider the upper bound of the risk when $t > \underline{t}_{\xi}$. Similar to Theorem 3.6.1, since $U_M^A(\xi_t, \hat{\mathbf{y}}, H_t) = \min\Big\{\varkappa\sqrt{2\min\{z, d-z\}}\mu(\xi_t)\rho_t\mathbb{I}\{\nu_S(\hat{\mathbf{y}}) \leq \rho_t\}, \eta_p\Big\}$, we define $\rho_{\tau,1} := \max : \{\rho_t(\xi) : \varkappa\sqrt{2\min\{z, d-z\}}\mu(\xi)\rho_t(\xi) \leq c_{\mathsf{min}}/\beta\}$. Define $\rho_{\tau,2} := \max\{\rho_t(\xi) : \nu_S(\bar{\mathbf{y}}) \geq 2\rho_t(\xi)\}$. Recall that $\nu_S(\bar{\mathbf{y}}) \geq 2\rho_t(\xi)$ implies $\nu_S(\hat{\mathbf{y}}) \geq \rho_t(\xi)$. Thus, $\rho_t(\xi) \geq \max\{\rho_{\tau,1}, \rho_{\tau,2}\}$, $\forall t = 1, 2, 3, ...$

Therefore, if $\rho_{\tau,1} \geq \rho_{\tau,2}$, the final risk at ξ is at most c_{min}/β. If $\rho_{\tau,1} \leq \rho_{\tau,2}$, the final risk at ξ is 0.

Therefore the final risk at ξ is equal to

$$\frac{c_{\mathsf{min}}}{\beta}\mathbb{I}\Big\{\rho_{\tau,1} \geq \rho_{\tau,2}\Big\} \leq \frac{c_{\mathsf{min}}}{\beta}\mathbb{I}\Big\{\frac{c_{\mathsf{min}}}{\sqrt{2\min\{z, d-z\}}\mu(\xi)\beta} \geq \frac{\nu_S(\bar{\mathbf{y}})}{2}\Big\}$$

Thus, the final risk in total is no more than

$$\begin{aligned}\sum_{\xi\in[m]}\frac{c_{\mathsf{min}}}{\beta}\mathbb{I}\Big\{\rho_{\tau,1}\geq\rho_{\tau,2}\Big\}&\leq\frac{c_{\mathsf{min}}}{\beta\underline{\mu}}\sum_{x\in[m]}\mu(\xi)\mathbb{I}\Big\{\rho_{\tau,1}\geq\rho_{\tau,2}\Big\}\\&=\frac{c_{\mathsf{min}}}{\beta\underline{\mu}}\mathbb{P}\Big(\rho_{\tau,1}\geq\rho_{\tau,2}\Big)\\&=\frac{c_{\mathsf{min}}}{\beta\underline{\mu}}\mathbb{P}\Big(\frac{c_{\mathsf{min}}}{\varkappa\sqrt{2\min\{z,d-z\}}\mu(\xi)\beta}\geq\frac{\nu_S(\bar{\mathbf{y}})}{2}\Big)\\&\leq\frac{c_{\mathsf{min}}}{\beta\underline{\mu}}\Psi\Big(\frac{\sqrt{2}c_{\mathsf{min}}}{\sqrt{\min\{z,d-z\}}\varkappa\underline{\mu}\beta}\Big).\end{aligned}$$

The inequalities hold because $\underline{\mu}\leq\mu(\xi)$, $\forall\xi\in[m]$.

Next, we provide the proof of Theorem 3.6.4.(2). We change (B.13) in the proof of Theorem 3.6.2 to the setting of the assortment optimization problem. Specifically, since the upper bound of the value of one data point is multiplied by κ, we multiply the first term in (B.13) by κ as well. For the second term, we replace the maximum regret $\sqrt{d}\eta_Y$ with η_p. Next, by defining the same value of c_q and following a similar procedure in the proof of Theorem 3.6.2, we can obtain the bounds for the expectation of the cumulative label cost in Theorem 3.6.4.(2).

Finally, to prove Theorem 3.6.4.(3), we observe that the above analysis deviates from the results in the product selection problem only by a multiplier $\varkappa$. Thus, Theorem 3.6.3 and Proposition 3.6.1 still hold for the assortment optimization problem. □

B.4 Proofs in Sections 3.7, 3.8 and 3.9

Proof of Theorem 3.7.1. We first prove Theorem 3.7.1.(1).

Recall in Algorithm 3, given one type ξ of customer, we let $U(\xi;\mathcal{S}_{t-1})\leftarrow\mu(\xi)\cdot\sqrt{2\min\{z,d-z\}}\,\beta\rho\mathbb{P}_{x_t\sim\mathcal{D}_\xi}(\nu_S(h_{t-1}(\xi,x_t))\leq\rho)$. Thus, we consider two cases: (1) $U(\xi;\mathcal{S}_{t-1})\geq\mu(\xi)c_{\mathsf{min}}$, and (2) $U(\xi;\mathcal{S}_{t-1})<\mu(\xi)c_{\mathsf{min}}$.

In the first case, we observe that the incentive in Algorithm 3, $U(\xi;\mathcal{S}_{t-1})$, is universal for all $x\in\mathcal{X}^\xi$. Thus, at each iteration, given one type ξ, the probability of taking the survey is the same for all features $x\in\mathcal{X}^\xi$. Thus, the samples within the training set $\mathcal{S}_{t-1}(\xi)$ are i.i.d. Since $p(c_{\mathsf{min}})\geq p_{\mathsf{min}}$, at each iteration, the probability of acquiring the labels of one type ξ in training set $\mathcal{S}_{t-1}(\xi)$ is at least $\mu(\xi)p(c_{\mathsf{min}})\geq p_{\mathsf{min}}\underline{\mu}$. Thus, by Chernoff's bound, after t iterations, when $U(\xi;\mathcal{S}_{t-1})\geq c_{\mathsf{min}}$, the number of samples in training set $\mathcal{S}_{t-1}(\xi)$ is least $0.5p_{\mathsf{min}}\underline{\mu}t$ with probability at least $1-e^{-\frac{\underline{\mu}p_{\mathsf{min}}t}{8}}$.

Recall that in Algorithm 3, given one type ξ, we let $\rho\leftarrow\Phi(|\mathcal{S}_{t-1}(\xi)|,\xi,\delta)$. Since in Theorem 3.7.1, $\rho_t\leftarrow\max_{\xi\in[m]}\Phi(\lfloor 0.5p_{\mathsf{min}}\underline{\mu}t\rfloor,\xi,\delta)$, we have that for any type ξ, we have $\rho\leq\rho_t$ with probability at least $1-e^{-\frac{\underline{\mu}p_{\mathsf{min}}t}{8}}$. By the definition of Φ, we have that ρ_t is the prediction error for the $h_{t-1}(\xi,x_t)$ with probability at least $1-e^{-\frac{\underline{\mu}p_{\mathsf{min}}t}{8}}$. In other words, with

probability at least $1 - e^{-\frac{\mu p_{\min} t}{8}}$, for any type ξ, we have

$$\sup_{x \in \mathcal{X}^\xi} \{\|h_{t-1,\xi}(x) - h^*_\xi(x)\|\} \leq \rho_t.$$

We denote $h_{t-1}(\xi, x_t)$ by $\hat{y}$, then similar to Theorem 3.3.1, the risk at one feature x is at most $U_M(\xi, \mathbf{y}, \rho) = \sqrt{2\min\{z, d-z\}}\rho_t \beta \mathbb{I}\{\nu_S(\hat{\mathbf{y}}) \leq \rho_t\}$. Thus, the risk for all feature $x \in \mathcal{X}^\xi$ is at most

$$\begin{aligned}\int_{x \in \mathcal{X}^\xi} \mu_\xi(x)\sqrt{2\min\{z, d-z\}}\beta\rho_t \mathbb{I}\left\{\nu_S(h_{t-1,\xi}(x)) \leq \rho_t\right\} dx \\ = \sqrt{2\min\{z, d-z\}}\beta\rho_t \mathbb{P}(\nu_S(\Theta x) \leq \rho_t).\end{aligned}$$

Next, we consider the second case $U(\xi; \mathcal{S}_{t-1}) < \mu(\xi)c_{\mathsf{min}}$. Given a type ξ, if $U(\xi; \mathcal{S}_{t-1}) < c_{\mathsf{min}}$, then we stop exploring type ξ, because $U(\xi; \mathcal{S}_{t-1})$ is a non-increasing sequence. Therefore, the risk for type ξ remains at $\mu(\xi)c_{\mathsf{min}}$.

Combining the upper bounds for these two cases, we have that the risk of h_T at type ξ is at most $U(\xi_t; \mathcal{S}_{t-1}) + \mu(\xi)c_{\mathsf{min}}$. Since the predictions across different types are independent, we have that the risk of h_T for all types is at most

$$\sum_{\xi \in [m]} [U(\xi_t; \mathcal{S}_{t-1}) + \mu(\xi)c_{\mathsf{min}}] = \sum_{\xi \in [m]} [U(\xi_t; \mathcal{S}_{t-1})] + c_{\mathsf{min}}.$$

To prove the risk bound in Theorem 3.7.1.(1), it suffices to show that $\frac{1}{\beta}\sum_{\xi \in [m]} [U(\xi_t; \mathcal{S}_{t-1})] \leq \varphi(t)$.

To prove $\varphi(t)$ is an upper bound for $\frac{1}{\beta}U(\xi_t; \mathcal{S}_{t-1})$, we first relax $\mathbb{P}_{x_t \sim \mathcal{D}_\xi}(\nu_S(h_{t-1}(\xi, x_t)) \leq \rho)$ to 1. Then, we obtain

$$\frac{U(\xi_t; \mathcal{S}_{t-1})}{\beta} \leq \sqrt{2\min\{z, d-z\}}\rho_T,$$

which is the first part of the upper bound in $\varphi(t)$.

To show that the second part in φ is also an upper bound for $\frac{1}{\beta}\sum_{\xi \in [m]} [U(\xi_t; \mathcal{S}_{t-1})]$, we relax $\sqrt{2\min\{z, d-z\}}\rho_T$ to $\sqrt{d}\eta_{\mathcal{Y}}$, which is the largest possible satisfaction loss. Next, to derive the upper bound for the risk, we provide an upper bound for $\mathbb{P}(\nu_S(h_{t-1}(\xi, x_t)) \leq \rho_t)$.

Note that this prediction error ρ_T does not depend on the choice of type ξ, and is a deterministic value. Hence, by the proof of Theorem 3.6.1, $\mathbb{P}(\nu_S(h_{t-1}(\xi, x_t)) \leq \rho_t) \leq \mathbb{P}(\nu_S(h^*_\xi(\xi, x_t)) \leq 2\rho_t) = \Psi_\xi(2\rho_t)$. Thus, we have that

$$\frac{U(\xi_t; \mathcal{S}_{t-1})}{\beta} \leq \sqrt{d}\eta_{\mathcal{Y}} \sum_{\xi \in [m]} \mu(\xi)\Psi_\xi(2\rho_t),$$

which is the second part in $\phi(t)$.

Next, we prove Theorem 3.7.1.(2). Since $\rho \leq \rho_t$, by the definition of $U(\xi_t; \mathcal{S}_{t-1})$, we have that the maximum incentive at time t is $\beta\sqrt{2\min\{z, d-z\}}\rho_t\mathbb{P}(\nu_S(h_{t-1}(\xi, x_t)) \leq \rho_t)$. Thus, to the expectation of the cumulative incentive is at most $\sum_{t=1}^{T}\beta\sqrt{2\min\{z, d-z\}}\rho_t\mathbb{P}(\nu_S(h_{t-1}(\xi, x_t)) \leq \rho_t)$. To prove Theorem 3.7.1.(2), it suffices to derive bounds on this summation.

To derive the upper bound in Theorem 3.7.1.(2), we relax $\mathbb{P}(\nu_S(h_{t-1}(\xi, x_t)) \leq \rho_t)$ to 1, so the upper bound is

$$\sum_{t=1}^{T}\beta\sqrt{2\min\{z, d-z\}}\rho_t.$$

This is the first part in Theorem 3.7.1.(2).

To obtain the second part in Theorem 3.7.1.(2), we relax $\sqrt{2\min\{z, d-z\}}\rho_t$ to $\sqrt{d}\eta_{\mathcal{Y}}$, and utilize the conclusion in the proof of Theorem 3.7.1.(2) that $\mathbb{P}(\nu_S(h_{t-1}(\xi, x_t)) \leq \rho_t) \leq \Psi(2\rho_t)$. Thus, the expectation of the total incentive is at most:

$$\sqrt{d}\eta_{\mathcal{Y}}\sum_{t=1}^{T}\Psi(2\rho_t),$$

which is the second part in Theorem 3.7.1.(2).

Next, we consider the case when $c_{\mathsf{min}} > 0$. Since the maximum incentive is c_{max}, the maximum incentive given at time t is at most $c_{\mathsf{max}}\mathbb{P}\big(\sqrt{2\min\{z, d-z\}}\rho_{n_t}\mathbb{P}(\nu_S(\Theta x) \leq \rho_{n_t}) \geq c_{\mathsf{min}}\big)$. Thus, when ρ_{n_t} is small enough such that $\sqrt{2\min\{z, d-z\}}\rho_{n_t}\mathbb{P}(\nu_S(\Theta x) \leq \rho_{n_t}) < c_{\mathsf{min}}$, we will stop providing incentives. Thus, n_t cannot go to infinity, and the cumulative label cost is finite. We can also derive a closed form upper bound for n_t with a lengthy proof similar to the proof of Theorem 3.6.2, but no interesting insights can be drawn.

Lastly, to prove Theorem 3.7.1.(3), we observe that the orders on T in the previous two arguments are the same as the product selection problem with finite support. Thus, we can immediately obtain Theorem 3.7.1.(3). □

Proof of Theorem 3.8.1. We again notice that the objective function (P4) is a linear function where the coefficients of the decision are $p_i\phi(i; \bar{\mathbf{y}})$. When Assumption 3.8.1 holds and the prediction errors for $\bar{y}$ is ρ, the estimation error of the coefficients in (P4) is at most $\eta_p\eta_b\rho$. By (3.9), the results of Theorem 3.8.1 can be immediately obtained when we replace $\varkappa$ in Theorem 3.6.4 with $\eta_p\eta_b$. □

Proof of Lemma 3.9.1. Since the estimations of different rows in Θ are independent, we first focus on the estimation of the first row, whose estimation and true value are denoted by $\hat{\theta}_{(1)}$ and $\theta^*_{(1)}$ respectively. Since $\hat{\theta}_{(1)}$ is the minimizer of the empirical squared loss, we have that $\hat{\theta}_{(1)} = \mathbf{\Lambda}^{-1}\mathbf{X}_t^T\mathbf{Y}_t$.

Then, by the fact that $\mathbf{Y}_t = \mathbf{X}_t\theta^*_{(1)} + \epsilon$, we have that the estimation error is at most

$$\begin{aligned}\|\hat{\theta}_{(1)} - \theta^*_{(1)}\| &= \|\mathbf{\Lambda}_t^{-1}\mathbf{X}_t^T\mathbf{Y}_t - \theta^*_{(1)}\| \\ &= \|\mathbf{\Lambda}^{-1}\mathbf{X}_t^T\epsilon\| \\ &\leq \|\mathbf{\Lambda}_t^{-1}\|_F\|\mathbf{X}_t^T\epsilon\| \\ &\leq \frac{\sqrt{m_\xi}}{\lambda_{\min}(\mathbf{\Lambda}_t)}\|\mathbf{X}_t^T\epsilon\|.\end{aligned}$$

Since $\lambda_{\min}(\mathbf{\Lambda}_t) \geq \frac{n_t 2\underline{\lambda}}{2} = n_t\underline{\lambda}$ with high probability, we have that $\|\hat{\theta}_{(1)} - \theta^*_{(1)}\| \leq \frac{\sqrt{m_\xi}}{\underline{\lambda}}\|\frac{\mathbf{X}_t^T\epsilon}{n_t}\|$.

From the standard Gaussian tail bounds, we have that with probability at least $1-2e^{-\frac{n\delta^2}{2}}$, we have

$$\left\|\frac{\mathbf{X}_t^T\epsilon}{n_t}\right\| \leq \sqrt{d}\left\|\frac{\mathbf{X}_t^T\epsilon}{n_t}\right\|_\infty \leq \sqrt{d}\eta_{\mathcal{X}}\sigma_\epsilon(\sqrt{\frac{2\ln(d)}{n_t}} + \delta).$$

Thus, by combining the above results, we have that with probability at least $1 - \sqrt{2}m_\xi e^{-\underline{\lambda}n_t/(2\eta_{\mathcal{X}}^2)} - 2e^{-\frac{n\delta^2}{2}}$, $\|\hat{\theta}_{(1)} - \theta^*_{(1)}\| \leq \frac{\sqrt{m_\xi}}{\underline{\lambda}}\sqrt{d}\eta_{\mathcal{X}}\sigma_\epsilon(\sqrt{\frac{2\ln(d)}{n_t}} + \delta)$. Thus, with the same probability, the prediction error for any $\hat{\theta}_{(1)}x_t$ on y^1 is at most $\|\hat{\theta}_{(1)} - \theta^*_{(1)}\| \leq \frac{\sqrt{m_\xi}}{\underline{\lambda}}\sqrt{d}\eta_{\mathcal{X}}^2\sigma_\epsilon(\sqrt{\frac{2\ln(d)}{n_t}} + \delta)$.

The prediction errors on the other rows have the same bound as the first row, so with probability at least $1 - \sqrt{2}dm_\xi e^{-\underline{\lambda}n_t/(2\eta_{\mathcal{X}}^2)} - 2de^{-\frac{n\delta^2}{2}}$, the whole prediction error for $\hat{\Theta}x$ is at most

$$\frac{\sqrt{m_\xi}}{\underline{\lambda}}d\eta_{\mathcal{X}}^2\sigma_\epsilon(\sqrt{\frac{2\ln(d)}{n_t}} + \delta).$$

Thus, by resetting the probability $1 - \sqrt{2}dm_\xi e^{-\underline{\lambda}n_t/(2\eta_{\mathcal{X}}^2)} - 2de^{-\frac{n\delta^2}{2}}$ to $1-\delta$, we can achieve a function $\Phi(n, \xi, \delta) \leq \tilde{\mathcal{O}}\left(n^{-1/2}\sqrt{\ln(\frac{1}{\delta})}\right)$.

To conclude this section, we provide another example of the Φ function. Suppose we are considering a decision tree hypothesis class and the density of features is bounded below by a positive constant. Then, by Hu, Kallus, and Mao, 2022, we can also obtain that $\Phi(n, \xi, \delta) \leq \tilde{\mathcal{O}}\left(n^{-1/2}\sqrt{\ln(\frac{1}{\delta})}\right)$. □

B.5 Numerical Experiments: Survey Details

In this appendix, we provide the details of the campus survey in the numerical experiments. The first 37 columns of the survey are shown in Table B.1. The answer to each question is Yes or No. We adopt the following rules to relate the columns to the six groups. The "art and culture" group is related to columns [1,17,26]; "science and tech" group is

related to columns [18,19]; The 'Social welfare and diversity' group are related to columns [0,2,3,4,5,8,9,10,23,25,27,28,29,30,33,34,35]; The "entrepreneurship" group is related to column [22]; The "sports" group is related to column [7,20,32]. The rest of the columns are related to the "others" group. For each column, if the answer of the student is "Yes", we add 1 point to her rating of the related group. Thus, the rating of each group is an integer number from 0 to 17.

Columns index	Question
0	Q1-Volunteered For Animal welfare
1	Q1-Volunteered For Arts/Culture/Heritage
2	Q1-Volunteered For Children/Youth
3	Q1-Volunteered For Community building
4	Q1-Volunteered For Diversity & Inclusion
5	Q1-Volunteered For Environmental sustainability
6	Q1-Volunteered For Families
7	Q1-Volunteered For Health/Well-being (e.g ment...
8	Q1-Volunteered For Seniors
9	Q1-Volunteered For Poverty reduction
10	Q1-Volunteered For Education
11	Q1-Volunteered For Others
12	Q2-Participated in Societies and Interest Groups
13	Q2-Participated in Clubs
14	Q2-Participated in Halls, JCRCs and/or Residen...
15	Q2-Participated in University organised events
16	Q2-Participated in Others
17	Q3-Interested in Arts & Culture
18	Q3-Interested in Science & Technology
19	Q3-Interested in Research and independent study
20	Q3-Interested in Sports
21	Q3-Interested in Other competitions (eg case, ...
22	Q3-Interested in Entrepreneurship
23	Q3-Interested in Volunteering
24	Q3-Interested in Others
25	Q4-Passionate about Animal welfare
26	Q4-Passionate about Arts/Culture/Heritage
27	Q4-Passionate about Children/Youth
28	Q4-Passionate about Community building
29	Q4-Passionate about Diversity & Inclusion (e.g...
30	Q4-Passionate about Environmental sustainability
31	Q4-Passionate about Families
32	Q4-Passionate about Health/Well-being (e.g men...
33	Q4-Passionate about Seniors
34	Q4-Passionate about Poverty reduction
35	Q4-Passionate about Education
36	Q4-Passionate about None of the above
37	Q4-Passionate about Others

Table B.1: Questions in survey

Appendix C

Proof for Chapter 3

C.1 Proofs in Sections 4.3

Proof of Proposition 4.3.1. Suppose the set of available product sets $\mathbb{S}_c$ is not a cover, then there exists a pair of products (i, j), such that products i and j are not included in any available product set within $\mathbb{S}_c$. Suppose $\hat{\rho}_{i,k}$ $k \neq j$ is the estimation result of Problem (4.2). Thus, for any value of $\hat{\rho}_{i,j}$ between $(0, 1)$, let $\hat{\rho}_{i,k} \leftarrow (1 - \hat{\rho}_{i,j})\hat{\rho}_{i,k}$, the objective value of Problem (4.2) remains the same, (because the likelihood function in Objective (4.2) does not depend on the scales). Thus, the estimation of $\hat{\boldsymbol{\rho}}$ is not unique.

□

Proof of Theorem 4.3.1. For any $S \in \mathbb{S}_c$, $|S| \leq n$ and $\Theta_{ij}^{*S} \geq \rho_{ij}, \forall i, j \in M$. Since $\hat{\boldsymbol{\rho}}$ is the solution of equality (4.9), we have $(\Theta_{ij}^{*S} - \hat{\Theta}_{ij}^{S})^2 \geq (\rho_{ij}^{*} - \hat{\rho}_{ij})^2$. Thus, the estimation error of each entry ρ_{ij} is no more than the estimation error in the submatrix Θ_{ij}. Since we recover each entry ρ_{ij} by re-scaling Θ_{ij}, the final error bounds for the squared Frobenius norm is at most the sum of the squared Frobenius norm under each submatrix. Thus, when each submatrix satisfies the estimation error bound, the estimation error for the entire matrix satisfies $\|\hat{\boldsymbol{\rho}} - \boldsymbol{\rho}^*\|_F \leq \sqrt{\sum_{S \in \mathbb{S}_c} \mathrm{err}_S^2}$. Since each estimation error bound holds with probability at least $1 - 4(2|S|)^{-\tau/c_1}$, by the union bound, the estimation error bound for the entire matrix holds with probability at least $1 - 4\sum_{S \in \mathbb{S}_c}(2|S|)^{-\tau/c_1}$.

□

Proof of Lemma 4.3.2. The first order derivative of $\ell_{\text{purchase}}(\alpha_i)$ is

$$-p_i W_{ii} + (N_i^B - W_{ii}) \frac{p_i \sum_{j \in S} \rho_{ij} e^{-\alpha_i p_i}}{1 - e^{-\alpha_i p_i}}.$$

The second order derivative of $\ell_{\text{purchase}}(\alpha_i)$ is

$$(N_i^B - W_{ii}) \frac{-(p_i)^2 e^{-\alpha_i p_i}}{(1 - e^{-\alpha_i p_i})^2} < 0.$$

Therefore, the function $\ell_{\text{purchase}}(\alpha_i)$ is concave for any product i.

□

Proof of Lemmas 4.3.1. The gradients of the $\mathcal{L}(\boldsymbol{\rho})$ in this lemma can be verified by the chain rule and derivative rules.

□

C.1.1 Proof of Lemma 4.3.3

The proof of Lemma 4.3.3 is inspired by Theorem 3 in Kallus and Udell, 2020. They consider the estimation of low-rank matrix, where the transition probability is an exponential function (MNL model) of the entries in each row. In contrast, in our setting, the low-rank matrix itself is the transition matrix. In their setting, the error bound grows in $\tilde{\mathcal{O}}(n)$, where n is the number of products. In our setting, the error bound grows in $\tilde{\mathcal{O}}(\ln(n))$, which is consistent with Theorem 1 in Z. Zhu et al., 2021, under some different settings and assumptions.

To begin the proof, we first introduce some necessary notations. Due to the simplicity of expression, throughout the proof, we neglect the index S for N_S and Θ^S, and use N and Θ instead in the proof. Since we consider a given availability S, for the simplicity of proof, we use n to denote the cardinality $|S|$. Let e_l be the l^{th} unit vector, i.e., the l^{th} element is one while other elements are zero. Let e_0 be the vector of all zeros. Suppose there are n products, then $e_l \in \mathbb{R}^n$. N is the total number of click transitions we observed. Recall that S denotes the set of available products and $\bar{S} = S \cup \{0\}$. We use $\|\cdot\|_{\max}$ to denote the maximum absolute value of the entries in the matrix, i.e., $\|\Delta\|_{\max} = \max_{i,j\in\bar{S}} : |\Delta_{i,j}|$. The dot product of two matrix $\mathbf{A} \cdot \mathbf{B}$ is $\sum_{i,j\in\bar{S}} A_{ij}B_{ij}$. Recall that in Assumption 4.3.1, we assume $\beta_0 \leq \Theta^S_{ij} \leq \frac{\beta_2}{|S|}$, $\forall i \in S, j \in \bar{S}$.

Then, we define:

- The error to bound $\Delta = \hat{\Theta} - \Theta^*$. We have that $\|\Delta\|_{\max} \leq 1$.
- The click indicator $X_{ij} = e_{i_t} e_{j_t}^T$. It means for the t^{th} click transition, it starts from product i_t, and click product j_t.
- The estimation error of the t^{th} observation $Y_t(\Delta) = \Delta^2_{i_t j_t}$.

Using these notations, the negative log likelihood function, its gradient, and its Hessian

matrix can be written as

$$\ell(\Theta) = \frac{1}{N}\sum_{t=1}^{N}\Big(\ln(\sum_{j\in S} X_{i_t j}\cdot\Theta) - \ln(X_{i_t j}\cdot\Theta)\Big) \tag{C.1}$$

$$\nabla\ell(\Theta) = \frac{1}{N}\sum_{t=1}^{N}\Big(\frac{\sum_{j\in S} X_{i_t j}}{\sum_{j\in S} X_{i_t j}\cdot\Theta} - \frac{X_{i_t j_t}}{X_{i_t j_t}\cdot\Theta}\Big) \tag{C.2}$$

$$\nabla^2\ell(\Theta) = \frac{1}{N}\sum_{t=1}^{N}\Big(\frac{X_{i_t j_t}^{\otimes 2}}{(X_{i_t j_t}\cdot\Theta)^2} - \frac{(\sum_{j\in S} X_{i_t j})^{\otimes 2}}{(\sum_{j\in S} X_{i_t j}\cdot\Theta)^2}\Big), \tag{C.3}$$

where $A^{\otimes 2} = A\otimes A$ is the symmetric linear operator on matrices defined by $(A\otimes A)(B) = (A\cdot B)A$.

Then, we first define the Bregman divergence

$$D_{\Theta^*}(\Delta) = \ell(\Theta^*+\Delta) - \ell(\Theta^*) - \nabla\ell(\Theta^*)\Delta.$$

We define the quadratic function

$$\ell_{\text{quad}}(\Delta) = \frac{1}{N}\sum_{t=1}^{N} Y_t(\Delta) = \frac{1}{N}\sum_{t=1}^{N}\Delta_{i_t j_t}^2.$$

The general idea of the proof of Lemma 4.3.3 is as follows. Lemma C.1.1 shows that $\ell_{\text{quad}}(\Delta)$ provides a lower bound on the Bregman divergence function. Lemma C.1.2 also shows that the log-likelihood function $\ell(\Theta)$ is restricted convex in the feasible space, and Lemma C.1.2 is built on Lemma C.1.3. The upper bound in Lemma C.1.4 shows that when the estimation error Δ is close to zero, the gradient of the log-likelihood function is also close to zero. Then, together with the results of Lemma 5 and 6 in Kallus and Udell, 2020, we show that when the regularization penalty γ is set as a proper value, the Frobenius norm of estimation error is close to zero.

Lemma C.1.1. *For any $\Delta := \hat{\Theta} - \Theta^*$, it holds that*

$$D_{\Theta^*}(\Delta) \geq \frac{1}{4}\ell_{quad}(\Delta).$$

Proof of Lemma C.1.1. Define $v_{tj} = X_{i_t j}(\Theta^* + s\Delta)$, where $s\in[0,1]$. Because $\|\Delta\|_{\max} \leq$

1, we have $v_{tj} \leq 2$. By Taylor's theorem, there is some $s \in [0,1]$ such that

$$\begin{aligned} D_{\Theta^*}(\Delta) &= \ell(\Theta^* + \Delta) - \ell(\Theta^*) - \nabla\ell(\Theta^*)\Delta \\ &= \nabla^2\ell(\Theta^* + s\Delta)[\Delta, \Delta] \\ &= \frac{1}{N}\sum_{t=1}^{N}\Big(\frac{(X_{t_i t_j}\Delta)^2}{v_{tj}^2} - \frac{\big((\sum_{j\in S} X_{i_t j})\Delta\big)^2}{(\sum_{j\in S} v_{tj})^2}\Big) \\ &= \frac{1}{N}\sum_{t=1}^{N}\frac{(X_{t_i t_j}\Delta)^2}{v_{tj}^2} \\ &\geq \frac{1}{N}\sum_{t=1}^{N}\frac{1}{4}(X_{t_i t_j}\Delta)^2 \\ &= \frac{1}{N}\frac{1}{4}\sum_{t=1}^{N}\Delta_{i_t j_t}^2 = \frac{1}{4}\ell_{\text{quad}}(\Delta). \end{aligned}$$

The fourth equality holds because $(\sum_{j\in S} X_{i_t j})\Delta = 0$. To see why it holds, Constraints (4.2a) require that $\sum_j \hat{\Theta}_{ij} = 1$, and thus $(\sum_{j\in S} X_{i_t j})\Delta = (\sum_{j\in S} X_{i_t j})(\hat{\Theta} - \Theta^*) = 0$.

□

Then, we will show that $\ell_{\text{quad}}(\Delta)$ is strongly convex when restricted to the matrices Δ, with high probability.

Lemma C.1.2. *Fix a parameter $\tau \geq 1$. Let*

$$\mathcal{A}^* = \left\{\Delta : \|\Delta\|_{\max} \leq 1, \|\Delta\|_* \leq \frac{1}{(9\max\{\tau, 16\})^{1/4}}\sqrt{\frac{\beta_1^5 N}{\ln(2n)}}\|\Delta\|_F^2\right\}.$$

We have that

$$\mathbb{P}\Big(\ell_{quad}(\Delta) \geq \beta_1^2\|\Delta\|_F^2, \forall\Delta \in \mathcal{A}^*\Big) \geq 1 - 3(2n)^{-\tau}.$$

Lemma C.1.2 is built on Lemma C.1.3, which is stated as follows.

Lemma C.1.3. *Let $\mathcal{A}_{\Gamma,\nu} = \left\{\Delta : \|\Delta\|_{\max} \leq 1, \|\Delta\|_F \leq \Gamma, \|\Delta\|_* \leq \frac{\nu\beta_1^2}{96\sqrt{2}}\sqrt{\frac{3N}{\ln(n)}}\Gamma^2\right\}$. Define the maximum deviation from strong convexity*

$$\mathcal{M}_{\Gamma,\nu} = \sup_{\Delta\in\mathcal{A}_{\Gamma,\nu}}\Big(\beta_1^2\|\Delta\|_F^2 - \ell_{quad}(\Delta)\Big).$$

Then, we have

$$\mathbb{P}\Big(\mathcal{M}_{\Gamma,\tau} \geq \nu\beta_1^2\Gamma^2\Big) \leq \exp\Big(-\frac{8}{9}\frac{Nn^2\nu^2\Gamma^4\beta_1^4}{\beta_2^2}\Big).$$

Next, we first provide the proof of Lemma C.1.3, and then we prove Lemma C.1.2.

Proof of Lemma C.1.3. Because $\Theta_{ij}^2 \geq \beta_1^2$, $\forall i, j \in [n]$, we have

$$\mathbb{E}[Y_t(\Delta)] = \mathbb{E}[\Delta_{i_t j_t}^2] = \sum_{i,j \in [n]} \mathbb{P}(i = i_t, j = j_i) \Delta_{i_t, j_t}^2 \geq \beta_1^2 \|\Delta\|_F^2.$$

Define

$$\tilde{\mathcal{M}}_{\Gamma,\nu} = \sup_{\Delta \in \mathcal{A}_{\Gamma,\nu}} \frac{1}{N} \sum_{t=1}^{N} \Big(\mathbb{E}[\Delta_{i_t,j_t}^2] - \Delta_{i_t,j_t}^2 \Big).$$

Therefore, we have $\tilde{\mathcal{M}}_{\Gamma,\nu} \geq \mathcal{M}_{\Gamma,\nu}$. Let Δ'_{i,j_t} be an iid replicate of Δ_{i,j_t}, and let ϵ_t be the iid Rademacher random variables. Then, we have

$$\begin{aligned}
\mathbb{E}[\tilde{\mathcal{M}}_{\Gamma,\nu}] &= \mathbb{E}\left[\sup_{\Delta \in \mathcal{A}_{\Gamma,\nu}} \frac{1}{N} \sum_{t=1}^{N} \Big(\mathbb{E}[(\Delta'_{i_t,j_t})^2] - \Delta_{i_t,j_t}^2 \Big) \right] \\
&\leq \mathbb{E}\left[\sup_{\Delta \in \mathcal{A}_{\Gamma,\nu}} \frac{1}{N} \sum_{t=1}^{N} \Big((\Delta'_{i_t,j_t})^2 - \Delta_{i_t,j_t}^2 \Big) \right] \\
&= \mathbb{E}\left[\sup_{\Delta \in \mathcal{A}_{\Gamma,\nu}} \frac{1}{N} \sum_{t=1}^{N} \epsilon_t \Big((\Delta'_{i_t,j_t})^2 - \Delta_{i_t,j_t}^2 \Big) \right] \\
&\leq 2\mathbb{E}\left[\sup_{\Delta \in \mathcal{A}_{\Gamma,\nu}} \frac{1}{N} \sum_{t=1}^{N} \epsilon_t \Delta_{i_t,j_t}^2 \right] \\
&= 2\mathbb{E}\left[\sup_{\Delta \in \mathcal{A}_{\Gamma,\nu}} \frac{1}{N} \sum_{t=1}^{N} \epsilon_t \|e_{i_t}^T \Delta e_{j_t}\|_2^2 \right].
\end{aligned}$$

Define $W_t = \epsilon_t e_{i_t} e_{j_t}^T$, where ϵ_t is the iid Rademacher random variables. By the Lemma 7 of Bertsimas and Kallus, 2020 and by Holder's inequality, we have

$$\mathbb{E}[\tilde{\mathcal{M}}_{\Gamma,\nu}] \leq 4\mathbb{E}\Big[\sup_{\Delta \in \mathcal{A}_{\Gamma,\nu}} \frac{1}{N} \sum_{t=1}^{N} W_t \Delta \Big] \leq 4\mathbb{E}\Big[\|\frac{1}{N} \sum_{t=1}^{N} W_t\|_2 \Big] \sup_{\Delta \in \mathcal{A}_{\Gamma,\nu}} \|\Delta\|_*.$$

Note that $\|W_t\|_2 = 1$, $\mathbb{E}[\|W_t W_t^T\|_2] \leq \frac{1}{n}$, and $\mathbb{E}[\|W_t^T W_t\|_2] \leq \frac{1}{n}$. By the matrix Bernstein inequality, we have that

$$\mathbb{P}\Big(\|\frac{1}{N} \sum_{t=1}^{N} W_t\|_2 \geq \delta \Big) \leq 2ne^{\frac{-3N\delta^2}{\frac{6}{n}+2\delta}} \leq 2n \max\{e^{-\frac{3\delta^2 nN}{8}}, e^{-\frac{3\delta N}{8}}\}.$$

Suppose $N \geq \frac{32}{3} n \ln(n)$. (In other words, we consider the case that the length of click trajectories is at least proportional to the number of products, which is the common setting in

low-rank estimation, e.g., Kallus and Udell, 2020 and Z. Zhu et al., 2021.) Given $\delta_1 = \sqrt{\frac{32\ln(n)}{3nN}}$, we have $e^{-\frac{3\delta^2 nN}{8}} \leq e^{-\frac{3\delta N}{8}}$, and thus, $\mathbb{P}(\|\frac{1}{n}\sum_{t=1}^{N} W_t\|_2 \geq \delta_1) \leq \frac{1}{n^2}$. As

$$\|\frac{1}{N}\sum_{t=1}^{N} W_t\|_2 \leq \frac{\sum_{t=1}^{N}\|W_t\|_2}{N} \leq 1,$$

we have

$$\begin{aligned}\mathbb{E}\Big[\|\frac{1}{n}\sum_{t=1}^{N} W_t\|_2\Big] &\leq \delta_1 \mathbb{P}\Big(\|\frac{1}{n}\sum_{t=1}^{N} W_t\|_2 \leq \delta_1\Big) + 1\cdot\mathbb{P}\Big(\|\frac{1}{n}\sum_{t=1}^{N} W_t\|_2 \geq \delta_1\Big)\\ &= \sqrt{\frac{32\ln(n)}{3nN}} + \frac{1}{n^2}\\ &\leq 8\sqrt{\frac{2\ln(n)}{3nN}}.\end{aligned}$$

Then, we have

$$\mathbb{E}[\tilde{\mathcal{M}}_{\Gamma,\nu}] \leq 32\sqrt{\frac{2\ln(n)}{3nN}}\frac{\nu\beta_1^2}{96\sqrt{2}}\sqrt{\frac{3N}{\ln(n)}}\Gamma^2 \leq \frac{\nu\beta_1^2\Gamma^2}{3}.$$

Let $\tilde{\mathcal{M}}'_{\Gamma,\nu}$ be a replicate of $\tilde{\mathcal{M}}_{\Gamma,\nu}$, where only i_t and j_t are different. Then the difference $|\tilde{\mathcal{M}}_{\Gamma,\nu} - \tilde{\mathcal{M}}'_{\Gamma,\nu}|$ is bounded by $\frac{1}{N}\sup_{\Delta\in\mathcal{A}_{\Gamma,\nu}}(\Delta_{ij}^2 - \Delta_{i'j'}{}^2) \leq \frac{1}{N}\frac{\beta_2^2}{n^2}$.

Hence, by McDiarmid's inequality, we have

$$\mathbb{P}(\tilde{\mathcal{M}}_{\Gamma,\nu} \geq \nu\beta_1^2\Gamma^2) \leq \mathbb{P}\Big(\tilde{\mathcal{M}}_{\Gamma,\nu} - \mathbb{E}[\tilde{\mathcal{M}}_{\Gamma,\nu}] \geq \frac{2\nu\beta_1^2\Gamma^2}{3}\Big) \leq \exp\Big(-\frac{8}{9}Nn^2\nu^2\Gamma^4\beta_1^4\frac{1}{\beta_2^2}\Big).$$

As $\tilde{\mathcal{M}}_{\Gamma,\nu} \geq \mathcal{M}_{\Gamma,\nu}$, we obtain Lemma C.1.3.

□

Next, we use Lemma C.1.3 to prove Lemma C.1.2.

Proof of Lemma C.1.2. This proof is similar to the proof of Lemma 3 in Kallus and Udell, 2020, except that we assign different values to τ', η, ν, and κ (κ is denoted by β in Kallus and Udell, 2020). Particularly, we set $\tau' = \max\{\tau, 16\}$, $\eta = (9\tau')^{1/4}\sqrt{\frac{\ln(2n)}{\beta_1^5 N}}$, $\nu = (\tau')^{1/4}$ and $\kappa = \sqrt{2\nu}$. Since $\|\cdot\|_* \geq \|\cdot\|_F$, we have that $\forall\Delta \in \mathcal{A}^*$, $\|\Delta\|_F \geq \eta$. Then, we have $\tau' \geq 16$, $\nu \geq 2$, and $\kappa \geq 2 > 1$. Let $\mathcal{A}_{lN} = \mathcal{A}^* \cap \{\eta\kappa^{(l-1)N} \leq \|\Delta\|_F \leq \eta\kappa^{lN}\}$. Thus, we have that $\mathcal{A}^* = \cup_{l=1\ldots\infty}\mathcal{A}_{lN}$ Then, if the event in Lemma C.1.2 is invalid and $\Delta \in \mathcal{A}_l$, we have

$\ell_{\text{quad}}(\Delta) \leq \frac{\beta_1^2}{2}\|\Delta\|_F^2$. Then, we have $\beta_1^2\|\Delta\|_F^2 - \ell_{\text{quad}}(\Delta) \geq \frac{\beta_1^2}{2}\|\Delta\|_F^2 \geq \frac{\beta_1^2}{2}(\eta\kappa^{(l-1)N})^2$. Then, the probability that the event is invalid is bounded by

$$\begin{aligned}
&\min\Big\{1, \sum_{l=1}^{\infty} \mathbb{P}(\mathcal{M}_{\kappa^{lN}\eta,1/2} \geq \frac{\beta_1^2}{2}(\eta\kappa^{(l-1)N})^2\Big\} \\
\leq &\min\Big\{1, \sum_{l=1}^{\infty} \exp(-\frac{2}{9}Nn^2\frac{\beta_1^4}{\beta_2^2}(\eta\kappa^{(l-1)N})^4)\Big\} \\
\leq &\min\Big\{1, \exp(-\frac{2}{9}Nn^2\frac{\beta_1^4}{\beta_2^2}\eta^4) + \sum_{l=2}^{\infty} \exp(-\frac{2}{9}Nn^2\frac{\beta_1^4}{\beta_2^2}(\eta\kappa^{(l-1)N})^4)\Big\} \\
= &\min\Big\{1, \exp(-\frac{2}{9}Nn^2\frac{\beta_1^4}{\beta_2^2}\eta^4) + \sum_{l=1}^{\infty} \exp(-\frac{2}{9}Nn^2\frac{\beta_1^4}{\beta_2^2}(\eta\kappa^{lN})^4)\Big\}.
\end{aligned}$$

Since $N \leq \frac{2c_1n^2}{9\beta_2^2}$, we have that

$$\exp\Big(-\frac{2}{9}Nn^2\frac{\beta_1^4}{\beta_2^2}\eta^4\Big) = \exp\Big(-\frac{2\tau' n^2(\ln(2n))^2}{\beta_1^6\beta_2^2 N}\Big) \leq \exp\Big(-\frac{2\tau'(\ln(2n))^2}{c_1}\Big).$$

Since $\kappa \geq 2$, we also have that

$$\begin{aligned}
\sum_{l=1}^{\infty} \exp\Big(-\frac{2}{9}Nn^2\frac{\beta_1^4}{\beta_2^2}(\eta\kappa^{lN})^4\Big) &\leq \sum_{l=1}^{\infty} \exp\Big(-\frac{2}{9}Nn^2\frac{\beta_1^4}{\beta_2^2}(\eta 2^{lN})^4\Big) \\
&\leq \sum_{l=1}^{\infty} \exp\Big(-\frac{2}{9}Nn^2\frac{\beta_1^4}{\beta_2^2}(\eta N^{2l})^4\Big) \\
&\leq \sum_{l=1}^{\infty} \exp\Big(-\frac{2}{9}Nn^2\frac{\beta_1^4}{\beta_2^2}\eta^4 N^4 l^4\Big) \\
&\leq \sum_{l=1}^{\infty} \exp\Big(-\frac{2}{9}Nn^2\frac{\beta_1^4}{\beta_2^2}\eta^4 Nl\Big) \\
&\leq \Big(\exp\Big(\frac{2}{9}Nn^2\frac{\beta_1^4}{\beta_2^2}\eta^4 N\Big) - 1\Big)^{-1} \\
&\leq 2\exp\Big(-\frac{2}{9}Nn^2\frac{\beta_1^4}{\beta_2^2}\eta^4 N\Big) \\
&= 2\exp\Big(-\frac{2\tau' n^2((\ln(2n))^2)}{\beta_1^6\beta_2^2}\Big) \\
&\leq 2\exp\Big(-\tau'(\ln(2n))^2\Big).
\end{aligned}$$

Since we only require the lower bound of c_1, without loss of generality, we assume $\frac{1}{c_1} \leq 1$. Therefore, the probability that the event is invalid is bounded by

$$\begin{aligned}\min\Big\{1, \exp\Big(-\frac{2\tau'(\ln(2n))^2}{c_1}\Big) + 2\exp\Big(-\tau'((\ln(2n))^2)\Big)\Big\} &\leq \min\Big\{1, 3\exp(-\frac{\tau'}{c_1}((\ln(2n))^2))\Big\} \\ &\leq 3\exp(-\frac{\tau'}{c_1}(\ln(2n))^2) \\ &\leq 3(2n)^{-\frac{\tau'}{c_1}} \leq 3(2n)^{-\tau/c_1}.\end{aligned}$$

Therefore, we conclude that

$$\mathbb{P}\Big(\ell_{\text{quad}}(\Delta) \geq \beta_1^2\|\Delta\|_F^2, \exists \Delta \in \mathcal{A}^*\Big) \geq 1 - 3(2n)^{-\tau/c_1}.$$

□

Then, Lemma C.1.4 provides an upper bound on the gradient of the log-likelihood function at Δ^*.

Lemma C.1.4. *For a fixed a parameter $\tau \geq 1$, with probability at least $1 - (2n)^{-\tau}$,*

$$\|\nabla\ell(\Delta^*)\|_2 \leq \sqrt{\frac{8\tau\ln(2n)}{N\beta_1}}.$$

Proof of Lemma C.1.4. Define

$$G_t = \frac{\sum_{j\in S} X_{i_t j}}{\sum_{j\in S} X_{i_t j}\Theta} - \frac{X_{i_t j_t}}{X_{i_t j_t}\Theta}.$$

Then, $\nabla\ell(\Delta^*) = \frac{1}{N}\sum_{t=1}^N G_t$. Because t_j is drawn according to Θ^*, we have $\mathbb{E}[G_t] = 0$. Note that $\sum_{j\in S} X_{ij}\Theta = 1, \forall i$. As $\frac{1}{\Theta_{ij}^2} \leq \frac{1}{\beta_1^2}$, $\forall i, j = 1...n$, we have $\|G_t\|_2 = \sqrt{\lambda_{\max}(G_t^T G_t)}$. Since each entry of $G_t^T G_t$ is less than

$$\max\Big\{1, 1 - \frac{1}{\Theta_{ij}}, \big(1 - \frac{1}{\Theta_{ij}}\big)^2\Big\} \leq \frac{4}{\beta_1^2}.$$

We have that

$$\|G_t\|_2 \leq \sqrt{\frac{4n}{\beta_1^2}} \leq \frac{2\sqrt{n}}{\beta_1}.$$

Next, we bound $\|\mathbb{E}[G_t G_t^T]\|_2$ and $\|\mathbb{E}[G_t^T G_t]\|_2$.

We have

$$G_t G_t^T = e_{i_t} e_{i_t}^T\Big(n - 1 + (1 - \frac{1}{\Theta_{i_t j_t}^2})^2\Big),$$

so $\mathbb{E}[G_t G_t^T]$ is a diagonal matrix, and each entry on the diagonal is

$$\Theta_{ij}\Big(n-1+(1-\frac{1}{\Theta_{ij}^2})^2\Big) = n\Theta_{ij} + \frac{1}{\Theta_{ij}} - 2.$$

Since $\Theta_{ij} \leq \frac{\beta_2}{n}$, we have that each entry on the diagonal is less than $\beta_2 - 2 + \frac{1}{\beta_1}$. Thus,

$$\|\mathbb{E}[G_t G_t^T]\|_2 \leq \Theta_{ij} \leq \beta_2 - 2 + \frac{1}{\beta_1}.$$

Let $y_{tj} = \mathbb{I}[j = j_t]$, and we have

$$G_t^T G_t = \sum_{j,k=1\ldots n} e_j e_k^T \Big((1 - y_{tj}\frac{1}{\Theta_{i_t j}})(1 - y_{tk}\frac{1}{\Theta_{i_t k}})\Big).$$

Then, for a given i, for the off-diagonal entries, they are 1 with probability $1 - \Theta_{ij}$ and they are $1 - \frac{1}{\Theta_{ij}}$ with probability Θ_{ij}. Thus, for the off-diagonal entries in $\mathbb{E}[G_t^T G_t]$, their value is

$$1 - \Theta_{ij} + (1 - \frac{1}{\Theta_{ij}})\Theta_{ij} = 0.$$

For the entries on the diagonal of $\mathbb{E}[G_t^T G_t]$, their values are

$$1 - \Theta_{ij} + \Theta_{ij}(1 - \frac{1}{\Theta_{ij}})^2 = \frac{1}{\Theta_{ij}} - 1 \leq \frac{1}{\beta_1} - 1 \leq \frac{2}{\beta_1}.$$

Therefore, we have,

$$\|\mathbb{E}[G_t^T G_t]\|_2 \leq \frac{2}{\beta_1}.$$

Thus, by $\beta_2 \leq \frac{1}{\beta_1}$, we have that

$$\max\{\|\mathbb{E}[G_t G_t^T]\|_2, \|\mathbb{E}[G_t^T G_t]\|_2\} \leq \beta_2 + \frac{2}{\beta_1} < \frac{3}{\beta_1}.$$

Therefore, by the matrix Bernstein inequality, we have

$$\mathbb{P}\left(\|\frac{1}{N}\sum_{t=1}^{N} G_t\|_2 \geq \delta\right) \leq 2n \max\left\{e^{-\frac{\delta^2 N\beta_1}{\beta}}, e^{-\frac{3\delta\beta_1 N}{16\sqrt{n}}}\right\}.$$

Suppose N satisfies $N \geq \frac{32}{9}\frac{n\tau}{\beta}\ln(2n)$. (In other words, we consider the case that the length of click trajectories is at least proportional to the number of products.) Then, the first term in $\max\{\cdot,\cdot\}$ dominates. We have that with probability at least $1 - (2n)^{-\tau}$, it holds that

$$\|\frac{1}{N}\sum_{t=1}^{N} G_t\|_2 \leq \sqrt{\frac{8\tau\ln(2n)}{N\beta_1}}.$$

□

By the Lemma 5 and Lemma 6 in Kallus and Udell, 2020, we have

$$D_{\Theta^*}(\Delta) \leq (\|\nabla \ell(\Theta^*)\|_2 + \gamma)\|\Delta\|_*,$$

and if $\|\nabla \ell(\Theta^*)\| \leq \gamma/2$, it holds that

$$\|\Delta\|_* \leq 16 \max\{\sqrt{r}\|\Delta\|_F, \|\bar{\Theta}_r^*\|_*\}.$$

Now, we provide the proof of Lemma 4.3.3.

Proof of Lemma 4.3.3. Setting $\gamma = \frac{1}{2}\sqrt{\frac{8\tau \ln(2n)}{N\beta_1}}$, we have

$$\|\nabla \ell(\Theta^*)\|_2 \leq \gamma/2 \leq \gamma,$$

then

$$D_{\Theta^*}(\Delta) \leq 2\gamma\|\Delta\|_*.$$

By Lemma C.1.1 and Lemma C.1.2, we have

$$\frac{\beta_1^2}{4}\|\Delta\|_F^2 \leq D_{\Theta^*}(\Delta) \leq \sqrt{\frac{8\tau \ln(2n)}{N\beta_1}}\|\Delta\|_*.$$

Then, we have

$$\|\Delta\|_F^2 \leq \frac{8}{\beta_1^2}\sqrt{\frac{2\tau \ln(2n)}{N\beta_1}}\|\Delta\|_*. \tag{C.4}$$

Next, we show that (C.4) holds even if $\Delta \notin \mathcal{A}^*$. Suppose so, then

$$\|\Delta\|_* > \frac{1}{(9\max\{\tau, 16\})^{1/4}}\sqrt{\frac{\beta_1^5 N}{\ln(2n)}}\|\Delta\|_F^2.$$

Then, by rewriting both sides and introducing redundant terms greater than 1, we can recover (C.4). Therefore, for all $\Delta \in \mathcal{A}^*$, (C.4) holds with high probability.

Then, if $\sqrt{r}\|\Delta\|_F \geq \|\bar{\Theta}_r^*\|_*$, we have $\|\Delta\|_* \leq 16\sqrt{r}\|\Delta\|_F$. Then, we have

$$\|\Delta\|_F \leq \frac{128}{\beta_1^2}\sqrt{\frac{2\tau r \ln(2n)}{N\beta_1}}.$$

□

C.2 Proofs in Section 4.4

Proof of Convergence of Algorithm 6. We prove the convergence of Algorithm 6 by showing that each iteration is a contraction. This proof is in the same vein as the proof of Lemma 14 in Dong, Simsek, and Topaloglu, 2019.

Define function $f_i(\cdot)$ as

$$f_i(\mathbf{r}^t) = \max_{p_i} \Big\{ \mu(p_i)(p_i - c_i) + \big(1 - \mu(p_i)\big) \sum_{j\in[n]} \rho_{ij} r_j^t \Big\}.$$

Then, for any two feasible vectors of stationary revenue $\mathbf{r}_1^t, \mathbf{r}_2^t \in \mathbb{R}^{n+1}$, suppose that $p_{i,1}$ be the optimal solution to problem $f_i(\mathbf{r}_1^t)$. Then, because $p_{i,1}$ may not be optimal under $f_i(\mathbf{r}_2^t)$, we have that

$$f_i(\mathbf{r}_2^t) \geq \mu(p_{i,1})(p_{i,1} - c_i) + \big(1 - \mu(p_{i,1})\big) \sum_{j\in[n]} \rho_{ij} r_{2,j}^t.$$

Since

$$f_i(\mathbf{r}_1^t) = \mu(p_{i,1})(p_{i,1} - c_i) + \big(1 - \mu(p_{i,1})\big) \sum_{j\in[n]} \rho_{ij} r_{1,j}^t,$$

we have that

$$f_i(\mathbf{r}_1^t) - f_i(\mathbf{r}_2^t) \leq \big(1 - \mu(p_{i,1})\big) \sum_{j\in[n]} \rho_{ij} \big[r_{1,j}^t - r_{2,j}^t\big] \leq \sum_{j\in[n]} \rho_{ij} \|\mathbf{r}_1^t - \mathbf{r}_2^t\|_\infty.$$

Because $\sum_{j\in[n]} \rho_{ij} \leq 1$, we further get $f_i(\mathbf{r}_1^t) - f_i(\mathbf{r}_2^t) \leq \|\mathbf{r}_1^t - \mathbf{r}_2^t\|_\infty$. Switching the role of $\mathbf{r}_1^t$ and $\mathbf{r}_2^t$, we can also get $f_i(\mathbf{r}_2^t) - f_i(\mathbf{r}_1^t) \leq \|\mathbf{r}_1^t - \mathbf{r}_2^t\|_\infty$. Therefore,

$$|f_i(\mathbf{r}_1^t) - f_i(\mathbf{r}_2^t)| \leq \|\mathbf{r}_1^t - \mathbf{r}_2^t\|_\infty,$$

and thus each iteration of Algorithm 6 is a contraction. By Theorem 6.2.3.a in Puterman, 2014, we have that there exist $\mathbf{r}^*$, such that $r_i^* = f_i(r^*)$ for all $i \in [n]$. (The uniqueness of r_i^* is proved in Dong, Simsek, and Topaloglu, 2019.) Therefore, Algorithm 6 converges to the $\mathbf{r}^*$. □

Proof of Proposition 4.4.1. We use $r_i(t)$ to denote the value of r_i in Algorithm 6 at iteration t. First, by the conclusion in Dong, Simsek, and Topaloglu, 2019, we have that $\arg\max_{p_i}\{\mu(p_i)(p_i - c_i) + (1 - \mu(p_i)) \sum_{j\in[n]} \rho_{ij} r_j\}$ is increasing in $\sum_{j\in[n]} \rho_{ij} r_j$. We use r_i and p_i to denote the outputs of Algorithm 6 under the transition matrix $\boldsymbol{\rho}$. Then, to get p_i' under $\boldsymbol{\rho}'$, we run Algorithm 6 under $\boldsymbol{\rho}'$, by initializing $r_i'(0) = r_i$.

Then, because the previous algorithm converges to the optimal price, we run the algorithm on transition matrix $\boldsymbol{\rho}'$ by initializing $r_i'(0) = r_i$. Then, we use the induction to prove that $r_i'(t+1) \geq r_i'(t)$ for any i.

At the first iteration, for $i \in [K]$,

$$\begin{aligned} r'_i(1) &= \max_{p_i} \left\{ \mu(p_i)(p_i - c_i) + (1 - \mu(p_i)) \sum_{j \in [n]} \boldsymbol{\rho}'_{ij} r'_j(0) \right\} \\ &= \max_{p_i} \left\{ \mu(p_i)(p_i - c_i) + (1 - \mu(p_i)) \sum_{j \in [n]} \rho'_{ij} r_j \right\} \\ &\geq \max_{p_i} \left\{ \mu(p_i)(p_i - c_i) + (1 - \mu(p_i)) \sum_{j \in [n]} \rho_{ij} r_j \right\} = r_i. \end{aligned}$$

For $i \notin [K]$,

$$r'_i(1) = \sum_{j \in [n]} \rho'_{ij} r'_j(0) = \sum_{j \in [n]} \rho'_{ij} r_j \geq \sum_{j \in [n]} \rho_{ij} r_j = r_i.$$

Suppose $r'_i(t+1) \geq r'_i(t)$, $\forall i \in [n]$ and $t \leq t_0$.
For $t = t_0 + 1$ and $i \in [K]$, we have

$$\begin{aligned} r'_i(t_0 + 2) &= \max_{p_i} \left\{ \mu(p_i)(p_i - c_i) + (1 - \mu(p_i)) \sum_{j \in [n]} \rho'_{ij} r'_j(t_0 + 1) \right\} \\ &\geq \max_{p_i} \left\{ \mu(p_i)(p_i - c_i) + (1 - \mu(p_i)) \sum_{j \in [n]} \rho'_{ij} r'_j(t_0) \right\} \\ &= r'_i(t_0 + 1). \end{aligned}$$

For $t = t_0 + 1$ and $i \notin [K]$, we have

$$r'_i(t_0 + 2) = \sum_{j \in [n]} \rho'_{ij} r'_j(t_0 + 1) \geq \sum_{j \in [n]} \rho'_{ij} r'_j(t_0) = r'_i(t + 1).$$

The inequality is by the assumption of induction. Therefore, $r'_i(t_0 + 2) \geq r'_i(t_0 + 1)$, $\forall i \in [n]$. Then, by induction, we have that $r'_i(t + 1) \geq r'_i(t)$, $\forall i \in [n]$ and $\forall t$. Therefore, taking t to infinity, we have $r'_i \geq r_i$. Therefore,

$$\sum_{j \in [n]} \rho'_{ij} r'_j \geq \sum_{j \in [n]} \rho'_{ij} r_j \geq \sum_{j \in [n]} \rho_{ij} r_j, \text{ for any product } i.$$

Then, we have $p'_i \geq p_i$, $\forall i \in [n]$. □

Proof of Proposition 4.4.2. Without loss of generality, suppose that $\rho_{i0} = \rho_{j0} = \rho^0$, $\forall i, j \in [n]$ and $c_i = c_j = c^0$, $\forall i, j \in [\bar{n}]$.

Let us first consider the optimal price under the attraction matrix $\tilde{\boldsymbol{\rho}}$ where all rows are the same, and $\tilde{\rho}_{i0} = \rho^0, \forall i \in [n]$. Suppose the optimal prices under $\tilde{\boldsymbol{\rho}}$ is $\tilde{\mathbf{p}}$, and the stationary revenue when clicking product i is $\tilde{v}_i$. Because all the products are the same, $\tilde{v}_i = \tilde{v}_j$, $\forall i, j \in [n]$, which is denoted by $\tilde{v}^0$. Hence, we have that

$$\begin{aligned}\tilde{p}_i &= \arg\max_{p_i} \Big\{\mu(p_i)(p_i - c_i) + \big(1 - \mu(p_i)\big) \sum_{j\in[n]} \rho_{ij} r_j^t\Big\} \\ &= \arg\max_{p} \Big\{\mu(p)(p - c^0) + \big(1 - \mu(p)(1 - \rho^0)\Big\}\tilde{v}^0,\end{aligned}$$

and

$$\tilde{v}^0 = \max_{p} \Big\{\mu(p)(p - c^0) + \big(1 - \mu(p)(1 - \rho^0)\tilde{v}^0\Big\}.$$

Now, let us consider optimizing the price using Algorithm 6 under $\boldsymbol{\rho}$. We initialize $\mathbf{r}$ as $\tilde{v}^0$. Then, when updating r_i^1,

$$\begin{aligned}r_i^1 &= \max_{p_i} \Big\{\mu(p_i)(p_i - c_i) + \big(1 - \mu(p_i)\big) \sum_{j\in[n]} \rho_{ij}\tilde{v}^0\Big\} \\ &= \max_{p_i} \big\{\mu(p)(p - c^0) + \big(1 - \mu(p)(1 - \rho^0)\big\}\,\tilde{v}^0 \\ &= \tilde{v}^0.\end{aligned}$$

Therefore, by induction, we have that $r_i^t = \tilde{v}^0$, $\forall t$. As a result, $p_i^t = \tilde{p}^0$, $\forall t$, $\forall i \in [n]$. Therefore, Proposition 4.4.2 holds. □

Proof of Lemma 4.4.1. Since p_1 and p_2 are the optimal prices under parameters $(\boldsymbol{\rho}_1, \boldsymbol{\alpha}_1)$ and $(\boldsymbol{\rho}_2, \boldsymbol{\alpha}_2)$ respectively, we know that the iterations in Algorithm 6 stops at p_1 and p_2 under parameters $(\boldsymbol{\rho}_1, \boldsymbol{\alpha}_1)$ and $(\boldsymbol{\rho}_2, \boldsymbol{\alpha}_2)$ respectively. When the iterations in Algorithm 6 stops, we have that

$$r_i(p_k) = \mu(p_{k,i})(p_{k,i} - c_i) + (1 - \mu(p_{k,i})) \sum_{j\in N} \rho_{ij} r_j(p_{k,j}) \text{ for } k = 1, 2.$$

Given $(\boldsymbol{\rho}_1, \boldsymbol{\alpha}_1)$ and $(\boldsymbol{\rho}_2, \boldsymbol{\alpha}_2)$, suppose $i = \arg\max_j\{r_j(p_1; \boldsymbol{\rho}_1, \boldsymbol{\alpha}_1) - r_j(p_2; \boldsymbol{\rho}_1, \boldsymbol{\alpha}_1)\}$. Within

the proof, we denote $r_i(\cdot;\boldsymbol{\rho}_1,\boldsymbol{\alpha}_1)$ by $r_i(\cdot)$ for short. Then, we have that

$$
\begin{aligned}
&r_i(p_1)-r_i(p_2)\\
=&\mu(p_{1,i})(p_{1,i}-c_i)-\mu(p_{2,i})(p_{2,i}-c_i)+(1-\mu(p_{1,i}))\sum_{j\in N}\rho_{ij}r_j(p_{1,j})-(1-\mu(p_{2,i}))\sum_{j\in N}\rho_{ij}r_j(p_{2,j})\\
=&\mu(p_{1,i})(p_{1,i}-c_i)-\mu(p_{2,i})(p_{2,i}-c_i)+(-\mu(p_{1,i})+\mu(p_{2,i}))\sum_{j\in N}\rho_{ij}r_j(p_{1,j})\\
&+(1-\mu(p_{2,i}))\Big(\sum_{j\in N}\rho_{ij}r_j(p_{1,j})-\sum_{j\in N}\rho_{ij}r_j(p_{2,j})\Big)\\
\le&\mu(p_{1,i})(p_{1,i}-c_i)-\mu(p_{2,i})(p_{2,i}-c_i)+(-\mu(p_{1,i})+\mu(p_{2,i}))\sum_{j\in N}\rho_{ij}r_j(p_{1,j})\\
&+(1-\mu(p_{2,i}))\Big(r_i(p_1)-r_i(p_2)\Big).
\end{aligned}
$$

Therefore,

$$
\begin{aligned}
&r_i(p_1)-r_i(p_2) &\text{(C.5)}\\
\le&\frac{1}{\mu(p_{2,i})}\Big(\mu(p_{1,i})(p_{1,i}-c_i)-\mu(p_{2,i})(p_{2,i}-c_i)+(-\mu(p_{1,i})+\mu(p_{2,i}))\sum_{j\in N}\rho_{ij}r_j(p_{1,j})\Big)\\
=&\frac{1}{\mu(p_{2,i})}\Big(\mu(p_{1,i})(p_{1,i}-p_{2,i})+(\mu(p_{1,i})-\mu(p_{2,i}))(p_{2,i}-c_i)+(-\mu(p_{1,i})+\mu(p_{2,i}))\sum_{j\in N}\rho_{ij}r_j(p_{1,j})\Big)\\
=&\frac{1}{\mu(p_{2,i})}\Big(\mu(p_{1,i})(p_{1,i}-p_{2,i})+(\mu(p_{1,i})-\mu(p_{2,i}))\big(p_{2,i}-c_i-\sum_{j\in N}\rho_{ij}r_j(p_{1,j})\big)\Big)\\
\le&\frac{1}{\mu(p_{2,i})}\Big(\mu(p_{1,i})(p_{1,i}-p_{2,i})+(\mu(p_{1,i})-\mu(p_{2,i}))\overline{p}\Big). &\text{(C.6)}
\end{aligned}
$$

We define $\epsilon=p_{1,i}-p_{2,i}$. Then, we have

$$
\begin{aligned}
&\frac{1}{\mu(p_{2,i})}\Big(\mu(p_{1,i})(p_{1,i}-p_{2,i})+(\mu(p_{1,i})-\mu(p_{2,i}))\overline{p}\Big)\\
=&e^{\alpha_{1,i}p_{2,i}}\Big(e^{-\alpha_{1,i}p_1}(p_{1,i}-p_{2,i})+(e^{-\alpha_{1,i}p_{1,i}}-e^{-\alpha_{1,i}p_{2,i}})\overline{p}\Big)\\
=&e^{\alpha_{1,i}(p_{2,i}-p_{1,i})}(p_{1,i}-p_{2,i})+e^{\alpha_{1,i}(p_{2,i}-p_{1,i})}\overline{p}-\overline{p}\\
=&e^{\alpha_{1,i}(p_{2,i}-p_{1,i})}(p_{1,i}-p_{2,i}+\overline{p})-\overline{p}\\
=&e^{-\alpha_{1,i}\epsilon}(\epsilon+\overline{p})-\overline{p}.
\end{aligned}
$$

Define function $f(\epsilon)=e^{-\alpha_{1,i}\epsilon}(\epsilon+\overline{p})-\overline{p}$. Then, we have that $f'(\epsilon)=e^{-\alpha_{1,i}\epsilon}(-\alpha_{1,i}\epsilon-\alpha_{1,i}\overline{p}+1)$. When $\overline{p}>\frac{1}{\alpha_{1,i}}$, if $\epsilon\ge 0$, we have that $-\alpha_{1,i}\epsilon-\alpha_{1,i}\overline{p}+1<0$. Therefore, $f'(\epsilon)\le 0$ when $\epsilon>0$. As a result, when $\epsilon>0$, $f(\epsilon)\le f(0)=0$. Note that although we use the condition $\overline{p}>\frac{1}{\alpha_{1,i}}$,

we can replace $\overline{p}$ with a large enough number in Equation (C.6) and yield the same result. Therefore, no additional condition about $\overline{p}$ is needed.

Then, we have that when $p_{1,i} \geq p_{2,i}$, $r_i(p_1) - r_i(p_2) \leq f(\epsilon) \leq 0$. Therefore, if $i = \arg\max_j\{r_j(p_1) - r_j(p_2)\}$, then, we have $p_{1,i} \leq p_{2,i}$.

As a sequence, to prove Lemma 4.4.1, it suffices to consider the case $p_{1,i} \leq p_{2,i}$.

In Algorithm 6,

$$\begin{aligned}
p_{k,i} &= \arg\max_{p'_i}\{\mu(p'_i)(p'_i - c_i) + (1-\mu(p'_i))\sum_{j\in N}\rho_{k,ij}r_j(p'_k;\boldsymbol{\rho}_k,\boldsymbol{\alpha}_k)\} \\
&= \arg\max_{p'_i}\{\mu(p'_i)(p'_i - c_i - \sum_{j\in N}\rho_{k,ij}r_j(p'_i;\boldsymbol{\rho}_k,\boldsymbol{\alpha}_k))\} \\
&= \arg\max_{p'_i}\{e^{-\alpha_{k,i}p'_i}(p'_i - c_i - \sum_{j\in N}\rho_{k,ij}r_j(p'_i;\boldsymbol{\rho}_k,\boldsymbol{\alpha}_k))\},\ \text{for } k=1,2.
\end{aligned}$$

By setting the first order derivative as zero, we have that $p_{k,i} = c_i + \sum_{j\in N}\rho_{k,ij}r_j(p_{k,i};\boldsymbol{\rho}_k, \boldsymbol{\alpha}_k) + \frac{1}{\alpha_{k,i}}$. Therefore,

$$\begin{aligned}
&p_{1,i} - p_{2,i} \\
&= \sum_{j\in N}\rho_{1,ij}r_j(p_1;\boldsymbol{\rho}_1,\boldsymbol{\alpha}_1) + \frac{1}{\alpha_{1,i}} - \sum_{j\in N}\rho_{2,ij}r_j(p_2;\boldsymbol{\rho}_2,\boldsymbol{\alpha}_2) - \frac{1}{\alpha_{2,i}} \\
&= \sum_{j\in N}(\rho_{1,ij}-\rho_{2,ij})r_j(p_1;\boldsymbol{\rho}_1,\boldsymbol{\alpha}_1) + \frac{1}{\alpha_{1,i}} - \frac{1}{\alpha_{2,i}} + \sum_{j\in N}\rho_{2,ij}\Big(r_j(p_1;\boldsymbol{\rho}_1,\boldsymbol{\alpha}_1) - r_j(p_2;\boldsymbol{\rho}_2,\boldsymbol{\alpha}_2)\Big).
\end{aligned}$$

We also have that

$$\begin{aligned}
&r_i(p_2;\boldsymbol{\rho}_2,\boldsymbol{\alpha}_2) - r_i(p_1;\boldsymbol{\rho}_1,\boldsymbol{\alpha}_1) \\
=&r_i(p_2;\boldsymbol{\rho}_2,\boldsymbol{\alpha}_2) - r_i(p_2) + r_i(p_2) - r_i(p_1) \\
=&\Big(e^{-\alpha_{2,i}p_{2,i}} - e^{-\alpha_{1,i}p_{2,i}}\Big)(p_2 - c_i) + \Big(1 - e^{-\alpha_{2,i}p_{2,i}}\Big)\sum_{j\in[n]}\rho_{2,ij}r_j(p_2;\boldsymbol{\rho}_2,\boldsymbol{\alpha}_2) - \\
&\qquad\Big(1 - e^{-\alpha_{1,i}p_{2,i}}\Big)\sum_{j\in[n]}\rho_{1,ij}r_j(p_2;\boldsymbol{\rho}_1,\boldsymbol{\alpha}_1) + r_i(p_2) - r_i(p_1) \\
\leq&\alpha_{2,i}|\alpha_{1,i} - \alpha_{2,i}|\overline{p} + \|\boldsymbol{\rho}_1 - \boldsymbol{\rho}_2\|_1\overline{p} + r_i(p_2) - r_i(p_1). \qquad \text{(C.7)}
\end{aligned}$$

If $p_{1,i} \leq p_{2,i}$, then

$$r_i(p_1) - r_i(p_2) \leq \frac{1}{\mu(p_{2,i})}(\mu(p_{1,i}) - \mu(p_{2,i}))\overline{p} \leq \frac{1}{\mu^l}(\mu(p_{1,i}) - \mu(p_{2,i}))\overline{p}.$$

By the definition of $\mu(\cdot)$, we have that

$$\begin{aligned}
&\mu(p_{1,i}) - \mu(p_{2,i}) = e^{-\alpha_{1,i}p_{1,i}} - e^{-\alpha_{1,i}p_{2,i}} \leq \alpha_{1,i}(p_{2,i} - p_{1,i}) \\
&= \alpha_{1,i}\Big(-\sum_{j\in N}(\rho_{1,ij}-\rho_{2,ij})r_j(p_1) - \frac{1}{\alpha_{1,i}} + \frac{1}{\alpha_{2,i}} - \sum_{j\in N}\rho_{2,ij}\Big(r_j(p_1) - r_j(p_2;\boldsymbol{\rho}_2,\boldsymbol{\alpha}_2)\Big)\Big).
\end{aligned}$$

Combining the result in (C.6) and (C.7), we have that

$$
\begin{aligned}
&r_i(p_1) - r_i(p_2) \le \frac{\overline{p}}{\underline{\mu}}(\mu(p_{1,i}) - \mu(p_{2,i})) \\
&\le \frac{\overline{p}}{\underline{\mu}}\alpha_{1,i}\Big(-\sum_{j\in N}(\rho_{1,ij} - \rho_{2,ij})r_j(p_1) - \frac{1}{\alpha_{1,i}} + \frac{1}{\alpha_{2,i}} - \sum_{j\in N}\rho_{2,ij}\Big(r_j(p_1) - r_j(p_2;\boldsymbol{\rho}_2,\boldsymbol{\alpha}_2)\Big)\Big) \\
&\le \frac{\overline{p}\alpha_{1,i}}{\underline{\mu}}\Big(\|\boldsymbol{\rho}_1 - \boldsymbol{\rho}_2\|_1\overline{p} + \frac{1}{\alpha_{2,i}} - \frac{1}{\alpha_{1,i}} + \max_{j\in[n]} : \Big(r_j(p_2;\boldsymbol{\rho}_2,\boldsymbol{\alpha}_2) - r_j(p_1)\Big)\Big) \\
&\le \frac{\overline{p}\alpha_{1,i}}{\underline{\mu}}\Big(\|\boldsymbol{\rho}_1 - \boldsymbol{\rho}_2\|_1\overline{p} + \frac{1}{\alpha_{2,i}} - \frac{1}{\alpha_{1,i}} + \max_{j\in[n]} : \alpha_{2,j}|\alpha_{1,j} - \alpha_{2,j}|\overline{p} + \|\boldsymbol{\rho}_1 - \boldsymbol{\rho}_2\|_1\overline{p} + r_i(p_2) - r_i(p_1)\Big) \\
&= \frac{\overline{p}\alpha_{1,i}}{\underline{\mu}}\Big(2\overline{p}\|\boldsymbol{\rho}_1 - \boldsymbol{\rho}_2\|_1 + \frac{1}{\alpha_{2,i}} - \frac{1}{\alpha_{1,i}} + \max_{j\in[n]} : \alpha_{2,j}|\alpha_{1,j} - \alpha_{2,j}|\overline{p} + r_i(p_2) - r_i(p_1)\Big).
\end{aligned}
$$

Therefore,

$$
(1 + \frac{\overline{p}\alpha_{1,i}}{\underline{\mu}})\big(r_i(p_1) - r_i(p_2)\big) \le \frac{\overline{p}\alpha_{1,i}}{\underline{\mu}}\Big(2\overline{p}\|\boldsymbol{\rho}_1 - \boldsymbol{\rho}_2\|_1 + \frac{1}{\alpha_{2,i}} - \frac{1}{\alpha_{1,i}} + \max_{j\in[n]} : \alpha_{2,j}|\alpha_{1,j} - \alpha_{2,j}|\overline{p}\Big).
$$

By the boundedness of p_i and α_i, we have that

$$
\begin{aligned}
r_i(p_1) - r_i(p_2) &\le \frac{\alpha_{i,i}\overline{p}}{\underline{\mu} + \alpha_{1,i}\overline{p}}\Big(2\overline{p}\|\boldsymbol{\rho}_1 - \boldsymbol{\rho}_2\|_1 + \frac{1}{\alpha_{2,i}\alpha_{1,i}}\|\boldsymbol{\alpha}_1 - \boldsymbol{\alpha}_2\|_\infty + \overline{\alpha p}\|\boldsymbol{\alpha}_1 - \boldsymbol{\alpha}_2\|_\infty\Big) \\
&\le \frac{\alpha_{1,i}\overline{p}}{\underline{\mu} + \alpha_{1,i}\overline{p}}\Big(2\overline{p}\|\boldsymbol{\rho}_1 - \boldsymbol{\rho}_2\|_1 + \Big(\frac{1}{\underline{\alpha}^2} + \overline{\alpha p}\Big)\|\alpha_1 - \alpha_2\|_\infty\Big).
\end{aligned}
$$

Thus there exist two positive numbers L_1 and L_2, where $L_1 = \frac{2\overline{\alpha p}^2}{\underline{\mu}+\underline{\alpha}\overline{p}}$ and $L_2 = \frac{(1+\overline{\alpha}^3\overline{p})\overline{p}}{\underline{\alpha}(\underline{\mu}+\underline{\alpha}\overline{p})}$, such that

$$
r_i(p_1) - r_i(p_2) \le L_1\|\boldsymbol{\rho}_1 - \boldsymbol{\rho}_2\|_1 + L_2\|\boldsymbol{\alpha}_1 - \boldsymbol{\alpha}_2\|_\infty, \text{ for any product } i \in [n].
$$

□

Proof of Lemma 4.4.2. We first bound the one-step regret. Since there are three types of parameters, the analysis of regret has three parts, i.e. low-rank estimation $\hat{\boldsymbol{\rho}}$, the price elasticity $\hat{\alpha}$, and arrival rate λ.

We first claim that we can ignore the error of λ in the analysis, because of the following two reasons. First, in algorithms 6, the optimal pricing does not depend on the arrival rate λ, therefore, if the error of $\boldsymbol{\rho}$ and α tends to zero, the price tends to be optimal. Secondly, because $\sum_i \lambda_i = 1$ and the stationary revenue from one customer is $\sum_i \lambda_i r_i$, therefore the error of lambda has limited influence on the revenue, given the price. Then, it suffices to analyze the error $\hat{r}_i^t - r_i^*$.

Since r can be written as a function of price p, it suffices to analyze $r(p_t) - r(p^*)$. By Lemma 4.4.1, we have that $r(p_t) - r(p^*) \le L_1\|\boldsymbol{\rho}^* - \boldsymbol{\rho}^t\|_1 + L_2\|\boldsymbol{\alpha}_1 - \boldsymbol{\alpha}_2\|_\infty$.

To analyze the error of $\hat{\boldsymbol{\alpha}}$, we first show that for each product i, the number of collected transition pairs starting from i grows at least in the order of $\mathcal{O}(T)$ with high probability. To show that, we observe that the transition probability $\rho_{ij} \geq \beta_1$, $\forall i, j$, and the purchase probability for any product is at most $\overline{\mu}$. Hence, the probability to have at least one transition pair starting from product i for any customer is at least $(1-\overline{\mu})\beta_1$. Let U_i^t denote the collected number of transition pairs starting from product i by customer t, then, by Chernoff's inequality, we have, for any $\delta \in (0,1)$,

$$\mathbb{P}\Big(U_i^t \geq t(1-\overline{\mu})\beta_1 - \sqrt{t\ln(1/\delta)}\Big) \geq 1-\delta. \tag{C.8}$$

Then, to obtain the error of $\hat{\alpha}$, we use the result in Theorem 1 in L. Li, Yu Lu, and Zhou, 2017, which characterizes the confidence bound for MLE of generalized linear models (GLM). Since prices and price elasticity are assumed to be within some intervals in Assumption 4.4.1, the conditions in Theorem 1 in L. Li, Yu Lu, and Zhou, 2017 is satisfied when $t \geq C(1+\ln(1/\delta))$, for some constant C. By Theorem 1 in L. Li, Yu Lu, and Zhou, 2017, we have that there exists a constant C, such that when $t \geq C(1+\ln(1/\delta))$, for any price p', for any product i,

$$p'(\hat{\alpha}_i - \alpha_i^*) \leq \frac{1}{\sqrt{U_i^t}} \frac{\overline{p}}{\underline{p}} \sqrt{\ln(1/\delta)}.$$

Setting $p' = \overline{p}$, we have that

$$|\hat{\alpha}_i - \alpha_i^*| \leq \frac{1}{\sqrt{U_i^t}} \frac{1}{\underline{p}} \sqrt{\ln(1/\delta)}.$$

Setting $\delta \leftarrow \frac{\delta}{T}$, we have that at each time t, the error from α_i is less than

$$L_2 \frac{1}{\sqrt{U_i^t}} \frac{1}{\underline{p}} \sqrt{\ln(T/\delta)}.$$

Combining the result in (C.8), by the union bound for all products, we have with probability at least $1-n\delta$,

$$\|\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^*\|_\infty \leq \frac{1}{\sqrt{t(1-\overline{\mu})\beta_1 - \sqrt{t\ln(1/\delta)}}} \frac{1}{\underline{p}} \sqrt{\ln(T/\delta)}. \tag{C.9}$$

When $t \geq \frac{2\ln(1/\delta)}{(1-\overline{\mu}\beta_1)}$, we have $t(1-\overline{\mu})\beta_1 - \sqrt{t\ln(1/\delta)} \geq \frac{1}{2}t(1-\overline{\mu})\beta_1$. Thus, when $t \geq \frac{2\ln(1/\delta)}{(1-\overline{\mu}\beta_1)}$, we have $\|\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^*\|_\infty \leq L_2 \frac{1}{\sqrt{t(1-\overline{\mu})\beta_1}} \frac{1}{\underline{p}} \sqrt{\ln(T/\delta)}$.

Thus, summing over $t = 1...T$, by the fact that $\sum_{t=1}^T \frac{1}{\sqrt{t}} \leq 2\sqrt{T}$, when $t \geq \frac{2\ln(1/\delta)}{(1-\overline{\mu}\beta_1)}$, we have that, there exists a constant $c_1 > 0$, the cumulative revenue loss from the second term is at most $2\frac{L_2 c_1}{\underline{p}(1-\overline{\mu})}\sqrt{T\ln(T/\delta)}$.

We denote the upper bound of $\|\hat{\Theta}^S - \Theta^{*S}\|_F$ in Lemma 4.3.3 by $\mathcal{B}(n, N)$, where n is the number of products and N is the number of click transitions. Then, the l_2-norm of each row in $\hat{\Theta}^S - \Theta^{*S}$ is no more than $\mathcal{B}(n, N)$.

For the error from $\boldsymbol{\rho}$, suppose N is the number of click transitions. Then, by the inequality that $\|X\|_1 \leq \sqrt{n}\|X\|_2$, we have that the error of first term $\|\boldsymbol{\rho}^* - \boldsymbol{\rho}^t\|_1$ is less than $\sqrt{n}\mathcal{B}(n, N)$. By Chernoff's bound, we have that with probability at least $1 - \delta$,

$$N \geq \frac{t}{\overline{\mu}} - \sqrt{\frac{1}{t}\ln(1/\delta)}. \tag{C.10}$$

In $\mathcal{B}(n, N)$, setting $4(2n)^{-\tau/c_1} = \frac{\delta}{T}$, we have that $\tau = c_1 \ln(\frac{4T}{\delta})\frac{1}{\ln(2n)}$, then we have

$$\mathcal{B}(n, N) = \frac{128}{\beta_1^2}\sqrt{\frac{2c_1 \ln(\frac{4T}{\delta})r}{N_S \beta_1}}. \tag{C.11}$$

Combining the lower bound in (C.10), we have that with probability at least $1 - 2\frac{\delta}{T}$, the error, $\|\boldsymbol{\rho}^* - \boldsymbol{\rho}^t\|_1$, is less than

$$\frac{128}{\beta_1^{2.5}}\sqrt{\frac{2c_1 \ln(\frac{4T}{\delta})nr}{\frac{t}{\overline{\mu}} - \sqrt{\frac{1}{t}\ln(\frac{t}{\delta})}}}. \tag{C.12}$$

Since $\sum_{t=1}^{T} \frac{1}{\sqrt{t}} \leq 2\sqrt{T} \in \mathcal{O}(\sqrt{T})$, we have that when $T \geq C(1 + \ln(\frac{1}{\delta}))$, with probability at least $1 - 2\frac{\delta}{T}$, the total regret of revenue in the first term is at most $\frac{c_2 L_1}{\beta_1^{2.5}}\sqrt{nrT\ln(\frac{T}{\delta})}$ for some constant c_2.

Then, adding up the regret of the first term and second term and sum over all the products, with probability at least $1 - 3n\delta$, the total regret is at most:

$$\frac{2L_2 c_1}{\underline{p}(1 - \overline{\mu}\beta_1)}\sqrt{T\ln\Big(\frac{T}{\delta}\Big)} + \frac{c_2 L_1}{\beta_1^{2.5}}\sqrt{nrT\ln\Big(\frac{T}{\delta}\Big)} \leq \Big(\frac{2L_2 c_1}{\underline{p}(1 - \overline{\mu}\beta_1)} + \frac{c_2 L_1}{\beta_1^{2.5}}\sqrt{nr}\Big)\sqrt{T\ln\Big(\frac{T}{\delta}\Big)}.$$

Thus, substituting $\delta \leftarrow \frac{\delta}{n}$, we conclude the regret bound in Lemma 4.4.2. □

Proof of Theorem 4.4.1. In the availability-focused pricing policy, the training set of each available product set is independent, so the estimation of the transition submatrix for each available product set is independent as well. Suppose that after observing T customers, the set of available product sets, which are shown at least one time, is denoted by $\mathbb{S}$. For each available product set $S_i \in \mathbb{S}$, we use $N_{S,i}$ to denote the number of click transition pairs that are collected under the S_i. Then, by (C.11), we have that for each $S_i \in \mathbb{S}$, with probability at least $1 - \delta/T$,

$$\|\hat{\Theta}^{S,i} - \Theta^{*S,i}\|_F \leq \frac{128}{\beta_1^2}\sqrt{\frac{2c_1 \ln(\frac{4T}{\delta})r}{N_{S,i}\beta_1}}. \tag{C.13}$$

By the fact that $\sum_{t=1}^{T} \frac{1}{\sqrt{t}} \leq 2\sqrt{T} \in \mathcal{O}(\sqrt{T})$, we have that when $T \geq C(1 + \ln(\frac{1}{\delta}))$, with probability at least $1 - \frac{\delta}{T}$, the regret of the revenue in the first term in Lemma 4.4.1 is at most $\frac{c_2 L_1}{\beta_1^{2.5}} \sqrt{nrN_{S,i} \ln(\frac{T}{\delta})}$ for some constant c_2. By the proof of Lemma 4.4.2, we have that for any available product set, the regret of the revenue from the second part in Lemma 4.4.1 is at most $2\frac{L_2 c_1}{\underline{p}(1-\overline{\mu})} \sqrt{T \ln(T/\delta)}$.

Thus, for any given available product set, we have that with probability at least $1 - \delta$, the cumulative regret is at most

$$\frac{c_2 L_1}{\beta_1^{2.5}} \sqrt{nrN_{S,i} \ln(\frac{T}{\delta})} + 2\frac{L_2 c_1}{\underline{p}(1-\overline{\mu})} \sqrt{T \ln(T/\delta)}.$$

Utilizing the Cauchy–Schwarz inequality and the fact that $\sum_{i=1}^{|\mathbb{S}|} N_{S,i} = T$, we have that

$$\sum_{i=1}^{|\mathbb{S}|} \sqrt{N_{S,i}} \leq \sqrt{T|\mathbb{S}|}. \tag{C.14}$$

Since $|\mathbb{S}| \leq T$, we have that with probability at least $1 - 2T\delta$, the total cumulative regret is at most

$$\frac{c_2 L_1}{\beta_1^{2.5}} \sqrt{nr|\mathbb{S}|T \ln(\frac{T}{\delta})} + 2\frac{L_2 c_1}{\underline{p}(1-\overline{\mu})} \sqrt{T \ln(T/\delta)}.$$

In the worst case, $|\mathbb{S}| \leq 2^n$, so we obtain that with probability at least $1 - 2T\delta$, for any sequence of available product sets, the cumulative regret is at most

$$\frac{c_2 L_1}{\beta_1^{2.5}} \sqrt{nr2^n T \ln(\frac{T}{\delta})} + 2\frac{L_2 c_1}{\underline{p}(1-\overline{\mu})} \sqrt{T \ln(T/\delta)}.$$

By replacing T with T/δ, we obtain the results in Theorem 4.4.1. □

Proof of Theorem 4.4.2. We use $\tilde{T}^S$ to denote the number of times that available product set S is shown to customers. Because at each iteration, the available product set is uniformly drawn from the cover $\mathbb{S}$, the probability of selecting one specific available product set S is $1/|\mathbb{S}|$. Thus, by Chernoff's bound, we can obtain the following lower bound for $\tilde{T}^S$ with high probability for any $\delta \in (0, 1)$:

$$\mathbb{P}\Big(\tilde{T}^S \geq \frac{T}{|\mathbb{S}|} - \sqrt{\frac{2T}{|\mathbb{S}|} \ln(\frac{1}{\delta})}\Big) \geq 1 - \delta.$$

Thus, for any available product set $S \in \mathbb{S}$, by the union bound, the probability that $\tilde{T}^S \geq \frac{T}{|\mathbb{S}|} - \sqrt{\frac{2T}{|\mathbb{S}|} \ln(\frac{1}{\delta})}$ is at least $1 - |\mathbb{S}|\delta$. Redefining $\delta \leftarrow |\mathbb{S}|\delta$, we have that for any available

product set $S \in \mathbb{S}$,

$$\mathbb{P}\Big(\tilde{T}^S \geq \frac{T}{|\mathbb{S}|} - \sqrt{\frac{2T}{|\mathbb{S}|}\ln(\frac{|\mathbb{S}|}{\delta})}\Big) \geq 1-\delta.$$

When $T \geq 8|\mathbb{S}|\ln(|\mathbb{S}|/\delta)$, we have that $\tilde{T}^S \geq \frac{T}{|\mathbb{S}|} - \sqrt{\frac{2T}{|\mathbb{S}|}\ln(\frac{|\mathbb{S}|}{\delta})} \geq \frac{T}{2|\mathbb{S}|}$. Thus, when $T \geq 8|\mathbb{S}|\ln(|\mathbb{S}|/\delta)$, we have for any available product set $S \in \mathbb{S}$,

$$\mathbb{P}\Big(\tilde{T}^S \geq \frac{T}{2|\mathbb{S}|}\Big) \geq 1-\delta. \tag{C.15}$$

According to Algorithm 5, the final estimation of the entire cover matrix is the combination of the rescaled submatrix under each availability. Thus, for each available product set $S \in \mathbb{S}$, the submatrix is the same as optimal solution Θ^S of problem 4.2. Thus, by Equ. (C.12) in the proof of Lemma 4.4.2, when Equ. (C.15) holds, we have that for any available product set $S \in \mathbb{S}$, with probability at least $1-2\delta/T$,

$$\|\Theta_T^S - \Theta^{*S}\|_1 \leq \frac{128}{\beta_1^{2.5}}\sqrt{\frac{2c_1\ln(\frac{4T}{\delta})nr}{\frac{t}{|\mathbb{S}|\overline{\mu}} - \sqrt{\frac{|\mathbb{S}|}{t}\ln(\frac{t}{|\mathbb{S}|\delta})}}}.$$

Next, regarding the estimation error for α_i, by (C.9) in the proof of Lemma 4.4.2, when Equ. (C.15) holds, with probability at least $1-\delta/T$, the error from α_i is less than

$$\|\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^*\|_\infty \leq \frac{1}{\sqrt{t(1-\overline{\mu})\beta_1 - \sqrt{t\ln(1/\delta)}}}\frac{1}{\underline{p}}\sqrt{|\mathbb{S}|\ln(T/\delta)}.$$

Using Lemma 4.4.1, the one-step regret bound for any available product set S can be upper bounded by $L_1\|\Theta^S - \Theta^{*S}\|_1 + L_2\|\boldsymbol{\alpha}_1 - \boldsymbol{\alpha}_2\|_\infty$. Finally, to achieve the regret bound for the cumulative revenue loss, we sum over t from 1 to T and utilize the fact that $\sum_{t=1}^T \frac{1}{\sqrt{t}} \leq 2\sqrt{T}$. This yields that when $T \geq C|\mathbb{S}|(1+\ln(\frac{|\mathbb{S}|}{\delta}))$, with probability $1-4\delta$, the regret of Algorithm 7 is at most

$$\text{Regret}(\{\mathbf{p}_t\}_{t=1}^T) \leq \Big(\frac{2L_2}{\underline{p}} + \frac{c_2L_1}{\beta_1^{2.5}}\sqrt{\tilde{N}r}\Big)\sqrt{|\mathbb{S}|T\ln\Big(\frac{\tilde{N}T}{\delta}\Big)},$$

for some constant $c_2, C > 0$.

□