Human-Centered Machine Learning: A Statistical and Algorithmic Perspective

Leqi Liu

August 2023
CMU-ML-23-104

Machine Learning Department
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA

Thesis Committee:
Zachary C. Lipton, Chair
Hoda Heidari
Fatma Kılınç-Karzan
Jon Kleinberg (Cornell University)

Submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

This research was sponsored by Air Force Research Laboratory award FA87501720152, contracts from Optum Technology and ZestFinance, and an Open Philanthropy AI Fellowship.

The views and conclusions contained in this document are those of the author and should not be interpreted as representing the official policies, either expressed or implied, of any sponsoring institution, the U.S. government or any other entity.

Keywords: artificial intelligence, machine learning, human-centered machine learning, responsible artificial intelligence, sequential decision-making, causal inference, statistical learning theory, personalization, preference learning, preference dynamics, risk preferences, recommender systems, decision-support systems

To my dearest mother, Guifang Chen,
who taught me that learning is a lifelong endeavor,
and it is never too late to start anything.

To my dearest father, Hongxing Liu,
who taught me to stay curious, cherish the beauty of rigor,
and always be open-minded.

Abstract

Building artificial intelligence systems from a human-centered perspective is increasingly urgent, as large-scale machine learning systems ranging from personalized recommender systems to language and image generative models are deployed to interact with people daily. In this thesis, we propose a guideline for building these systems from a human-centered perspective. Our guideline contains three steps: *(i)* identifying the role of the people of interest and their core characteristics concerned in the learning task; *(ii)* modeling these characteristics in a useful and reliable manner; and *(iii)* incorporating these models into the design of learning algorithms in a principled way.

We ground this guideline in two applications: personalized recommender systems and decision-support systems. For recommender systems, we follow the guideline by *(i)* focusing on users' evolving preferences, *(ii)* modeling them as dynamical systems, and *(iii)* developing efficient online learning algorithms with provable guarantees to interact with users sharing different preference dynamics. For decision-support systems, we *(i)* choose decision-makers' risk preferences to be the core characteristics of concern, *(ii)* model them in the objective function of the system, and *(iii)* provide a general procedure with statistical guarantees for learning models under diverse risk preferences. We conclude by discussing the future of human-centered machine learning and the role of interdisciplinary research in this field.

Acknowledgments

I am extremely lucky to have received enormous support from my advisors, mentors, friends, collaborators, labmates, colleagues and family members during my Ph.D. They gave me inspiration and courage, accompanied me through difficult times, and made sure that I was taken care of in this not-that-short not-that-easy journey. Below is a non-exhaustive list of people who I owe my Ph.D. to.

I owe my Ph.D. to my advisor Zachary C. Lipton. He welcomed me to the ACMI lab with the utmost enthusiasm, gave me the space and freedom to pursue my own research ideas, shared with me sharp, critical and fun perspectives on research, writing and philosophy, connected me with resources that enabled me to pursue a variety of research directions, and helped me keep an open and light mind throughout my Ph.D.

I owe my Ph.D. to my thesis committee members: Hoda Heidari, Fatma Kılınç-Karzan and Jon Kleinberg. They gave me invaluable guidance on research, career and life, empathized with me whenever I sought their advice, and have always served as north stars for me.

I owe my Ph.D. to labmates in the ACMI lab: Nil-Jana Akpinar, Sina Fazelpour, Michael Feffer, Riccardo Fogliato, Kyra Gan, Saurabh Garg, Shantanu Gupta, Jennifer Hsia, Divyansh Kaushik, Kundan Krishna, Tanya Marwah, Danish Pruthi, Jacob Tyo and Tom Yan. Among many things, we bumped project ideas, practiced research talks, and cultivated coding and research habits together.

I owe my Ph.D. to mentors from whom I have sought advice on research, career and beyond: Kamyar Azizzadenesheli, Sarah Dean, Maria De-Arteaga, Nika Haghtalab, Ellen Vitercik and Mariya Toneva. Their openness about their own experience and pursuit of research and excellence have constantly inspired me.

I owe my Ph.D. to colleagues I met during internships at Apple and DeepMind: Jan Balaguer, Tom Gunter, Raphael Koster, Miruna Pislar, Christopher Summerfield, Andrea Tacchetti and many other interns and researchers. They have made the internships both fulfilling and enjoyable, introduced me to new research areas and technical skills, and opened my eyes to industry perspectives.

I owe my Ph.D. to my collaborators and whom I have discussed many research ideas and questions with: Bryon Aragam, Omer Ben-Porat, David Childers, Lee Cohen, Chen Dan, Dylan Hadfield-Menell, Kenneth Holstein, Audrey Huang, Biwei Huang, David Inouye, Yueyi Jiang, Yunfan Jiang, Edward H Kennedy, Joon Sik Kim, Justin Khim, Xinyu Li, Chaochao Lu, Yishay Mansour, Alan Montgomery, Adarsh Prasad, Charvi Rastogi, Pradeep Ravikumar, Meirav Segal, Stephen Tu, William Wong, Yifan Wu, Fan Yang and Kun Zhang. I overly enjoyed brainstorming ideas, learning about new research areas, envisioning projects, formalizing problems, discussing proofs, coding, writing and proofreading together with them.

I owe my Ph.D. to my friends and CMU colleagues. While some of them are my collaborators and labmates, many of them are not. Diane Stidle has always been there for me whenever I needed help navigating through my Ph.D. My officemates Sebastian Caldas, Helen Zhou and Brandon Trabucco supported me through my

job search, and discussed random research and entrepreneurship ideas with me. Going for coffee and lunch together has been a source of refreshing breaks for me. During the pandemic, my friends living around me—Kartik Gupta, Sitoshna Jatty, Kin Gutierrez Olivares, Biswajit Paria, Sai Sandeep, Stefania La Vattiata and Giulio Zhou—have given me great companionship and support. They walked with me in Schenley and Frick Park after a day of work and shared their cooking and bakery with me when they ran into interesting recipes. Shaojie Bai, Yao-Hung Tsai and Chih-Kuan Yeh have explored the Asian food scene in Pittsburgh with me. The many hot pots we had at the early stage of my Ph.D. helped me get used to graduate school much quicker. My Ph.D. would not be complete without the many trips I made with my friends Jin Li, Alena Klindziuk, Feilun Cao, Devendra Chaplot and Kanthashree Mysore. These trips saved me when I was on the verge of burning out. Finally, there are many other friends and colleagues at CMU that I want to express my gratitude to: Maruan Al-Shedivat, Siddharth Ancha, Angela Chen, Valerie Chen, Amanda Coston, Christoph Dann, Priya Donti, Ben Eysenbach, Chirag Gupta, Stefani Karp, Lisa Lee, Liam Li, Bingbin Liu, Yusha Liu, Pratik Patil, Anthony Platanios, Otilia Stretcu, Arun Sai Suggala, Nicholay Topin, Yining Wang, Manzil Zaheer and Han Zhao.

I owe my Ph.D. to my advisers during my undergrad: Lynne Butler, Deepak Kumar, Jia Tao and Lyle Ungar. They introduced me to research and the beauty of mathematics, statistics and computer science, making it possible for me to pursue a Ph.D. in machine learning.

I owe my Ph.D. to my family members. My father Hongxing Liu and mother Guifang Chen have provided me with unconditional love and support throughout my life. It is their firm belief in me that has enabled me to pursue my dreams. My partner Eyan P. Noronha has made my Ph.D. filled with laughter, surprises, adventures, interesting perspectives, exciting road trips, fun puzzles and appreciation of the beauty of nature. His constant love and care have given me the braveness and confidence to face any problem in life.

Words cannot express my gratitude to them all. I hope we continue running into each other throughout our lives. Until next time!

Contents

Chapter 1

Introduction

We are in an exciting era where artificial intelligence systems ranging from personalized recommender systems to large-scale language and image generative models are being deployed to interact with people at scales. To date, one of the most successful paradigms for building these intelligent systems is machine learning. At its core, machine learning deals with data—it uncovers statistical patterns underlying the data and utilizes these patterns to make inferences and decisions on future unseen scenarios. It typically involves specifying a set of models to choose from, a learning objective that emits a score of a model under the given data and settings, and an optimization procedure for finding the model that performs well under the learning objective. In this pipeline, the data is generated by human users or annotators, and the machine learning objectives are specified by human engineers or researchers. Once a model is learned, it may be evaluated by human stakeholders or crowdsourcing workers and later deployed in settings where human users may provide additional feedback to adjust and improve the model. To this extent, though many machine learning models are not explicitly trained to interact with people, the entire machine learning pipeline implicitly (or explicitly) relies on humans in an indispensable manner.

Despite the ubiquity of human presence throughout the machine learning pipeline, traditional academic treatments of machine learning have largely focused on problem settings that either overlook or oversimplify it. This may result in building undesirable systems in terms of *(i)* Usefulness: when the end users of these systems are humans, without knowing their way of using the systems, what we build may not be useful for them; *(ii)* Safety: if we ignore users' preferences and behavior characteristics, the resulting systems may harm the end users; and *(iii)* Reliability: even in contexts where the machine learning systems are not deployed to interact with human users, understanding the influence of other people involved in the pipeline of building these systems will improve the reliability of them. Thus, in this thesis, we focus on the human-centered perspective of building machine learning systems. The central question we aim to answer is the following:

> *Are there some guidelines one could use when building machine learning systems from a human-centered perspective?*

Taking a human-centered perspective on machine learning has become an increasingly

studied subject in recent years—both the empirical and theoretical machine learning communities have started looking into explicitly accounting for human presence in the learning pipeline. For example, given the ambiguity of the learning objective of language generative models, researchers proposed to fine-tune models based on human preferences instead of using a pre-specified reward (hence objective) function (Ziegler et al., 2019; Stiennon et al., 2020). On the other hand, for theoretical machine learning frameworks that aim to account for human presence, humans have mostly been modeled as rational or noisily rational agents (Arora and Doshi, 2021; Haghtalab et al., 2021) and behave adversarially (Podimata, 2022; Haghtalab et al., 2022) or strategically (Hardt et al., 2016a; Kleinberg and Raghavan, 2020). In addition to the machine learning community that develops algorithms for building machine learning models in these settings, researchers from human-computer interaction, public policy, philosophy, and many other fields have proposed new evaluation mechanisms for machine learning pipelines from a human-centered perspective, using human subject studies (Holstein et al., 2019; Heger et al., 2022) and simulations of humans involved in the pipeline (Dai et al., 2021; Martin et al., 2023). As this naturally interdisciplinary field—human-centered machine learning—continues to grow, there is an urgent need to define its scope, understand the role of different disciplines in this field, and outline how researchers across disciplines could collaborate.

1.1 Thesis Statement

This thesis aims to provide principles for developing machine learning systems from a human-centered perspective and illustrate how these principles can be applied to a variety of learning tasks in application domains ranging from decision-support systems to personalized recommender systems. Throughout, we will discuss other disciplines' roles in the process of developing these systems.

Thesis Statement. The following statement is at the center of this thesis: *Building machine learning systems from a human-centered perspective requires us to*

- ***S.1*** *specify the role of the humans of interest and their core characteristics concerned in the learning task;*
- ***S.2*** *develop reliable models that capture these core characteristics;*
- ***S.3*** *incorporate these models into the design of the machine learning systems in a principled way.*

Remark 1.1. *The "model" referred in **S.2** is a mathematical model in the broad sense and may not be a machine learning model. In certain machine learning settings, **S.2** and **S.3** may take place simultaneously. For example, one could learn a preference model of the user and build machine learning systems that interact with the user at the same time. Finally, the overarching goal for **S.1** and **S.2** is to build useful and ideally simple human models for the machine learning setting of interest.*

When illustrating the usage of this guideline, we focus on two application domains: personalized recommender systems and decision-support systems. For recommender systems, we take the human-centered perspective by specifying the humans of interest to be the users and their core characteristics concerned in the learning task to be their evolving preferences (**S.1**). We

use human subject studies to test the existence of such evolving preferences and model them using dynamical systems (S.2). New online learning algorithms with provable guarantees are developed in face of these user preference dynamics (S.3). For decision-support systems, we choose the decision-makers who are facilitated by these systems to be the subject of interest, and their risk preferences to be the core characteristics (S.1). We model these risk preferences in the objective functions of the machine learning models (S.2), and provide algorithms with statistical guarantees for learning such models (S.3).

Due to the interdisciplinary nature of human-centered machine learning, other disciplines play important roles in this field. For example, the human-computer interaction community can help machine learning researchers with S.1 by conducting user studies to understand the core characteristics concerned in a learning task; and with S.2 by testing if the proposed model for capturing those core characteristics is valid. On the other hand, social sciences like behavioral economics and psychology provide machine learning researchers with modeling assumptions on the relevant characteristics of the people of our interest in S.2. These assumptions and structures are especially of need in low-data regimes and may help machine learning researchers build human models that are simple yet useful in S.2.

1.2 Overview

We organize the thesis into three parts. The first part covers S.2 in the thesis statement and focuses on modeling human preferences and decision-making characteristics in recommender systems and decision-support systems, respectively. The second part illustrates S.3 by discussing how one could develop machine learning systems while integrating human preference models. The last part summarizes the takeaways and envisions future directions in human-centered machine learning. We provide a brief overview for the first two parts of the thesis.

1.2.1 Modeling Human Preferences and Decision-Making (Part I)

The first part presents how we could build reliable models for human preferences and decision-making. Using human subject studies, we identify existing oversimplifications of user preferences in the development of recommender system algorithms (Chapter 2). Inspired by models proposed in behavioral economics, we develop new ways for representing different risk preferences in the design of decision-support systems (Chapter 3).

Chapter 2: Understanding the Validity of Assumptions on Human Preferences in Multi-armed Bandits

Theoretical treatments of recommendation systems frequently address the decision-theoretic nature of the problem via the multi-armed bandits (MABs) framework. In this framework, the learner (e.g., the recommender system) interactively chooses actions to take (e.g., recommendations to present) based on the historical rewards it collects. These rewards are users' responses to the recommendations, which are driven by their underlying preferences. Thus, the assumptions on reward distributions, which MAB-based approaches rely heavily on, can be

translated to assumptions on user preferences. Such assumptions are seldom evaluated with real-world human interactants, making it difficult to assess the applicability of the algorithms powered by them. As an example, in the MABs framework, the reward distribution is commonly assumed to be fixed, suggesting that the user preference is static over time. However, it is unclear whether such simplification of user preference is valid: the existing psychology and marketing literatures have shown that people's preferences evolve as a function of their past consumption (e.g., (Tucker, 1964)).

To test the validity of core MABs assumptions, we provide a flexible experimental framework and an open source library for collecting human preference data when interacting with recommendation algorithms in MABs. Using this framework, we conduct human subject studies in a comics recommendation setting and find that the static reward assumption does not hold, and thus bandit algorithms interacting with humans must account for user preference dynamics. We subsequently apply the experimental framework to assess the performance of different MAB algorithms interacting with real human interactants. In general, our toolkits can be used to identify assumptions on reward distributions that better captures the characteristics of user preference dynamics.

Chapter 3: On Human-Aligned Risk Minimization

While the statistical decision-theoretic foundations of modern machine learning have largely focused on the minimization of the expectation of some loss function, seminal results in behavioral economics—Cumulative Prospect Theory (CPT)—have shown that human decision-making is based on different risk measures that overweigh small probabilities and underweigh large probabilities (Tversky and Kahneman, 1992). To model the risk preference of a decision-maker when building machine learning models in decision-support systems, we design a new objective function based on the CPT. We explore how the models learned under this new objective differ from the ones learned under the traditional expectation-based objective and empirically demonstrate that these models have improved performance on desiderata such as fairness.

1.2.2 Integrating Human Characteristics into Machine Learning Algorithm Design (Part II)

The second part illustrates how we could incorporate human models into the design of machine learning algorithms with provable guarantees. For recommender systems, we design online learning algorithms to interact with users with evolving preferences (Chapter 4) who may leave the system (Chapter 5). For decision-support systems, we provide recipes for learning models under different risk preferences in settings including supervised learning (Chapter 6) and optimal treatment regimes (Chapter 7).

Chapter 4: Rebounding Bandits for Modeling Satiation Effects

Psychological research shows that enjoyment of many goods is subject to satiation, with short-term satisfaction declining after repeated exposures to the same item. Nevertheless, proposed algorithms for powering recommender systems seldom model these dynamics, instead proceeding as though user preferences were fixed in time. We introduce rebounding bandits, a

multi-armed bandit setup, where satiation dynamics are modeled as time-invariant linear dynamical systems. Unlike classical bandit settings, methods for tackling rebounding bandits must plan ahead and model-based methods rely on estimating the parameters of the satiation dynamics. We characterize the planning problem and propose efficient learning algorithms that achieve sublinear dynamic regret.

Chapter 5: Modeling Attrition in Recommender Systems with Departing Bandits

We consider settings where the user has evolving preferences over time and may leave the recommender system upon being presented by unsatisfying content. Traditionally, when recommender systems are formalized as multi-armed bandits, the policy of the recommender system influences only the rewards accrued, but not the length of interaction. Our proposed departing bandits setup captures policy-dependent horizons. While naive approaches cannot handle this setting, we provide an efficient learning algorithm that utilizes a simple planning shortcut and achieves sublinear regret.

Chapter 6: Supervised Learning with General Risk Functionals

To date, the vast majority of supervised, unsupervised, and reinforcement learning research has focused on objectives expressible as expectations (over some dataset or distribution) of an underlying loss (or reward) function. However, real-world concerns such as risk aversion, equitable allocations of benefits and harms, or alignment with human preferences, often demand that we address other functionals of the loss distribution. Focusing on supervised learning, consider the common scenario in which a population contains a minority (constituting fraction α of the population) but where group membership was not recorded in the available data. If the pattern relating the features to the label were different for different demographics, a naively trained model might adversely harm members of a minority group. Absent further information, one sensible strategy could be to optimize the worst case performance over all subsets (of size up to α). This would translate to the Conditional Value at Risk (CVaR) objective (i.e., the expectation of the worst 100α percent of the losses incurred by the model).

To account for such risk preferences of decision-makers, we develop a learning framework for obtaining models under general risk functionals. Our results rely on estimating the Cumulative Distribution Function of the loss distribution and identifying that a variety of risk functionals of interest (e.g., CVaR, mean-variance) are Lipschitz. We establish the first uniform convergence results for estimating Lipschitz risk functionals, which licenses us to perform empirical risk minimization.

Chapter 7: Median Optimal Treatment Regimes

We propose a new treatment regime that optimizes a particular risk functional of interest, the Average Conditional Median Effect (ACME), which summarizes across-group median treatment outcomes of a policy, and which the median optimal treatment regime maximizes. This proposed risk functional captures "in-subject robustness" and "across-subject fairness." We give a nonparametric efficiency bound for estimating the ACME of a policy, and propose a new doubly

robust-style estimator that achieves the efficiency bound under weak conditions. To construct the median optimal treatment regime, we introduce a new doubly robust-style estimator for the conditional median treatment effect.

1.3 Bibliographic Notes

Most of the work in this thesis is based on the following papers. The asterisk * in authorship indicates equal contributions to the paper.

Chapter 2 is based on:

- Liu Leqi*, Giulio Zhou*, Fatma Kilinc-Karzan, Zachary Lipton, and Alan Montgomery. "A Field Test of Bandit Algorithms for Recommendations: Understanding the Validity of Assumptions on Human Preferences in Multi-armed Bandits". In: *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI)* (2023)

Chapter 3 is based on:

- Liu Leqi, Adarsh Prasad, and Pradeep Ravikumar. "On Human-Aligned Risk Minimization". In: *Advances in Neural Information Processing Systems (NeurIPS)* (2019)

Chapter 4 is based on:

- Liu Leqi, Fatma Kilinc Karzan, Zachary Lipton, and Alan Montgomery. "Rebounding bandits for modeling satiation effects". In: *Advances in Neural Information Processing Systems (NeurIPS)* (2021)

Chapter 5 is based on:

- Omer Ben-Porat*, Lee Cohen*, Liu Leqi*, Zachary C Lipton, and Yishay Mansour. "Modeling attrition in recommender systems with departing bandits". In: *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)* (2022)

Chapter 6 is based on:

- Liu Leqi, Audrey Huang, Zachary Lipton, and Kamyar Azizzadenesheli. "Supervised Learning with General Risk Functionals". In: *International Conference on Machine Learning (ICML)* (2022)

Chapter 7 is based on:

- Liu Leqi and Edward H Kennedy. "Median optimal treatment regimes". In: *arXiv preprint arXiv:2103.01802* (2021)

Part I

Modeling Human Preferences and Decision-Making

Chapter 2

A Field Test of Bandit Algorithms for Recommendations: Understanding the Validity of Assumptions on Human Preferences in Multi-armed Bandits

2.1 Introduction

As online services have become inescapable fixtures of modern life, recommender systems have become ubiquitous, influencing the music curated into our playlists, the movies pumped into the carousels of streaming services, the news that we read, and the products suggested whenever we visit e-commerce sites. These systems are commonly data-driven and algorithmic, built upon the intuition that historical interactions might be informative of users' preferences, and thus could be leveraged to make better recommendations (Gomez-Uribe and Hunt, 2015). While these systems are prevalent in real-world applications, we often observe misalignment between their behavior and human preferences (Jain et al., 2015). In many cases, such divergence comes from the fact that the underlying assumptions powering the learning algorithms are questionable.

Recommendation algorithms for these systems mostly rely on supervised learning heuristics (Bennett, Lanning, et al., 2007; Gomez-Uribe and Hunt, 2015), including latent factor models such as matrix factorization (Koren et al., 2009) and deep learning approaches that are designed to predict various heuristically chosen targets (Wei et al., 2017) (e.g., purchases, ratings, reviews, watch time, or clicks (Bennett, Lanning, et al., 2007; McAuley and Leskovec, 2013)). Typically they rely on the naive assumption that these behavioral signals straightforwardly indicate the users' preferences. However, this assumption may not hold true in general for a variety of reasons, including exposure bias (users are only able to provide behavioral signals for items they have been recommended) and censoring (e.g., reviews tend to be written by users with strong opinions) (Swaminathan and Joachims, 2015; Joachims et al., 2017).

On the other hand, to study the decision-theoretic nature of recommender systems, the online

decision-making framework—multi-armed bandits (MABs)—has commonly been used (Slivkins, 2019; Lattimore and Szepesvári, 2020). In MABs, at any given time, the decision-maker chooses an arm to pull (a recommendation to make in the recommender system setting) among a set of arms and receives a reward, with the goal of obtaining high expected cumulative reward over time. Theoretical research on MABs centers on algorithms that balance between the exploration and exploitation tradeoff and analyses capturing the performance[1] of these algorithms (Lattimore and Szepesvári, 2020). There is a long line of work on developing MAB-based approaches for recommender systems in settings including traditional K-armed bandits (Barraza-Urbina and Glowacka, 2020, and references therein), contextual bandits (Li et al., 2010; McInerney et al., 2018; Mehrotra et al., 2020, and references therein), and Markov decision processes (Shani et al., 2005a; Chen et al., 2019; Chen et al., 2021a; Chen et al., 2021b; Wang et al., 2022; Chen et al., 2022, and references therein). To guide such research on applying MAB-based algorithms in recommender systems, it is of importance to *test* whether the assumptions that these algorithms are built upon are valid in real-world recommendation settings.

In this work, we focus on the assumption of temporal stability that underlies both practical supervised learning methods and algorithms for classical MAB settings where the reward distribution is assumed to be fixed over time. In a recommendation setting, the reward distribution of an arm corresponds to the user's preference towards that recommendation item. Although the assumption that the reward distribution is fixed may be appropriate to applications driving early MABs research (e.g., sequential experimental design in medical domains) (Robbins, 1952), one may find it to be unreasonable in recommender systems given that the interactants are humans and the reward distributions represent human preferences. For example, consider the task of restaurant recommendations, though a user may be happy with a recommended restaurant for the first time, such enjoyment may decline as the same recommendation is made over and over again. This particular form of evolving preference is known as satiation, and results from repeated consumption (Galak and Redden, 2018). One may also think of cases where a user's enjoyment increases as the same item being recommended multiple times, due to reasons including sensitization (Groves and Thompson, 1970). In both settings, the assumption that reward distributions are fixed is violated and the recommendation algorithms may influence the preferences of their users.

We test the assumption on fixed reward distributions through randomized controlled trials conducted on Amazon Mechanical Turk. In the experiment, we simulate a K-armed bandit setting (a MAB setup where the arm set is the same set of K arms over time) and recommend comics from K comic series to the study participants. After reading a comic, the participants provide an enjoyment score on a 9-point Likert scale (Cox III, 1980), which serves as the reward received by the algorithm for the recommendation (for pulling the corresponding arm). Each comic series belongs to a different genre and represents an arm. Our analyses on the collected dataset reveal that in a bandit recommendation setup, human preferences can evolve, even within a short period of time (less than 30 minutes) (Section 2.4). In particular, between two predefined sequences that result in the same number of pulls of each arm in different order, the mean reward for one arm has a significant difference of 0.57 (95% CI = $[0.30, 0.84]$, p-value < 0.001). This suggests that any MAB algorithms that are applied to recommendation settings should

[1]We provide more details on how performances are defined in MABs in Section 2.3.1.

account for the dynamical aspects of human preferences and the fact that the recommendations made by these algorithms may influence users' preferences.

The line of work that develops contextual bandits and reinforcement learning algorithms for recommender systems (Shani et al., 2005a; McInerney et al., 2018; Chen et al., 2019; Mehrotra et al., 2020; Chen et al., 2021a; Chen et al., 2021b; Wang et al., 2022; Chen et al., 2022) hinges upon the assumption that users' preferences depend on the past recommendations they receive. The proposed algorithms have been deployed to interact with users on large-scale industry platforms (e.g., Youtube). These prior works differ from ours in multiple ways. Among them, the most significant distinction is that our work aims to understand and identify assumptions on reward distributions that better represent user preference characteristics. On the other hand, this line of work asks the question that *once an assumption on the reward distribution has been made*, how to design the recommendation algorithms so that they have better performance. For example, in settings where recommender systems are modeled as contextual bandits (Mehrotra et al., 2020; McInerney et al., 2018), the reward distributions are assumed to take a certain functional form (often linear) in terms of the observable states. When treating recommender systems as reinforcement learning agents in a Markov decision process (Shani et al., 2005a; Chen et al., 2019; Chen et al., 2021a; Chen et al., 2021b; Wang et al., 2022; Chen et al., 2022), one has made the core assumption that the reward is Markovian and depends only on the *observable* user states (e.g., user history) and recommendations. In other words, the reward (and user preference) does not depend on *unobservable* states (e.g., user emotions) as in partially observable Markov decision processes. Under this assumption, the proposed algorithms deal with difficulties (e.g., large action spaces) in the reinforcement learning problem. It is worth noting that in these prior works, the proposed algorithms have been evaluated on industry platforms. For academics who want to evaluate their recommendation algorithms with human interactants, such infrastructure is not easily accessible. We take an initial step to address this need by developing our experimental framework.

The experimental framework we developed is flexible. It allows one to conduct field tests of MAB algorithms, use pre-defined recommendation sequences to analyze human preferences, and ask users to choose an arm to pull on their own. These functionalities can be used to identify assumptions on user preference dynamics and reward distributions that better capture user characteristics. As an illustration of the flexibility of our experimental framework, we have collected data while the participants interact with some traditional MAB algorithms and analyze their experience with these algorithms. Interestingly, we observe that interactants (of a particular algorithm) who have experienced the lowest level of satisfaction are the ones to have the poorest performance in recalling their past ratings for previously seen items.

In summary, we provide a flexible experimental framework that can be used to run field tests with humans for any K-armed bandit algorithms (Section 2.3.3). Using this experimental framework, we have tested the validity of the fixed-reward-distribution (fixed-user-preference) assumption for applying MAB algorithms to recommendation settings (Section 2.4). As an illustration of the flexibility of our experimental framework, we have inspected different bandit algorithms in terms of user enjoyment and attentiveness (Section 2.5). We discuss the limitation of our study in Section 2.6. The code for our experimental framework can be found at https://github.com/HumainLab/human-bandit-evaluation.

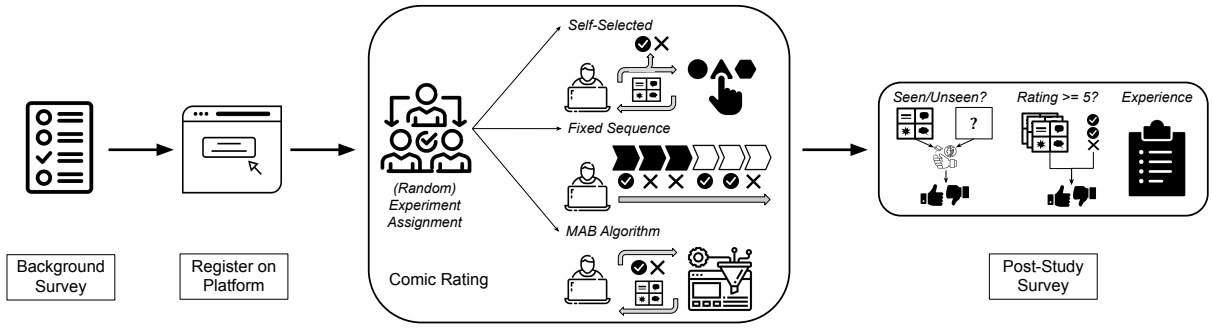


Figure 2.1: Overview of the experimental protocol. Participants first complete a background survey and then register their MTurk ID on our platform to get randomly assigned (without their direct knowledge) one of the following types of experimental setups: Self-selected, one of the fixed sequences, or one of the MAB algorithms. Study participants who are assigned to a fixed sequence and an MAB algorithm only provide ratings (enjoyment scores to the comics). Self-selected study participants provide both a rating as well as the next genre to view. Once the comic rating portion of the study is complete, participants move onto the post-study survey, where they are asked questions related to their experience in the study, e.g., how well they remember the consumed content. Participants must complete all parts of the study and answer attention checks sufficiently correctly to receive full compensation, and therefore be included in the final study data.

2.2 Related Work

The study of evolving preferences has a long history, addressed by such diverse areas as psychology (Christensen and Brooks, 2006; Galak and Redden, 2018), economics (Prelec, 2004; Gul and Pesendorfer, 2005), marketing (Tucker, 1964), operations research (Baucells and Sarin, 2007), philosophy (Liu et al., 2007; Baber, 2007), and recommender systems (Kapoor et al., 2015; Rafailidis and Nanopoulos, 2015; Lee et al., 2014). In the bandits literature, there is a recent line of work, motivated by recommender systems, that aims to incorporate the dynamic nature of human preference into the design of algorithms. These papers have different models on human preferences, expressed as specific forms of how the reward of an arm depends on the arm's past pulls. Levine et al. (2017) and Seznec et al. (2019) model rewards as monotonic functions of the number of pulls. By contrast, Kleinberg and Immorlica (2018), Basu et al. (2019), and Cella and Cesa-Bianchi (2020) consider the reward to be a function of the time elapsed since the last pull of the corresponding arm. In Mintz et al. (2020), rewards are context-dependent, where the contexts are updated based on known deterministic dynamics. Finally, Leqi et al. (2021b) consider the reward dynamics to be unknown stochastic linear dynamical systems. These prior works model the reward (user preferences) in distinct ways, and lack (i) empirical evidence on whether user preferences evolve in the short period of time in a bandit setup; and (ii) datasets and experimental toolkits that can be used to verify the proposed theoretical models and to explore more realistic ways of modeling user preferences. On the other hand, there is a line of work on developing contextual bandits and reinforcement learning algorithms to account for user preference dynamics in recommender systems. The evaluations of these algorithms against human users rely on accessibility to large-scale industry platforms (Shani et al., 2005a;

McInerney et al., 2018; Chen et al., 2019; Mehrotra et al., 2020; Chen et al., 2021a; Chen et al., 2021b; Wang et al., 2022; Chen et al., 2022).

Datasets that are publicly available and can be used to evaluate bandit algorithms in recommendation settings often contain ratings or other types of user feedback of recommendations (Saito et al., 2020; Lefortier et al., 2016). These datasets do not contain trajectories of recommendations and the associated feedback signal for a particular user, making it hard to understand the dynamical aspects of user preferences and to identify effects of past recommendations on the preferences. In addition, existing MAB libraries (e.g., (Holtz et al., 2020)) only consist of implementations of bandit algorithms, but lack the appropriate tools and infrastructure to conduct field tests of these algorithms when interacting with humans. The toolkit we have developed closes this gap and allows one to use these libraries while conducting human experiments on bandit algorithms. Another relevant stream of research that considers human experiments in bandits settings are the ones that ask human participants to make decisions in a bandit task and collect data for modeling their decision-making (Acuna and Schrater, 2008; Reverdy et al., 2014; Lee et al., 2011). In other words, in those experiments, human participants take the role of the algorithm that selects the arm to pull instead of the role of providing reward signals. In one of our experimental setups, the study participants are asked to select comics to read on their own. However, in contrast to reward distributions that are defined by the experiment designers of those experiments, in our setting, the rewards are provided by the human participants, indicating their preferences.

2.3 Experimental Setup

In this section, we first describe a classical MABs setting—stochastic K-armed bandits—and discuss the algorithms we have used in our experiments (Section 2.3.1). We then provide reasoning on why we choose comics as the reommendation domain and selection criteria for the comics used for recommendations (Section 2.3.2). Finally, we discuss our experimental framework (Section 2.3.3).

2.3.1 Stochastic K-armed bandits

In stochastic K-armed bandits, at any time t, the learner (the recommender in our case) chooses an arm $a(t) \in [K]$ to pull (a recommendation to present in our case) and receives a reward $R(t)$. For any horizon T, the goal for the learner is to attain the highest expected cumulative reward $\mathbb{E}[\sum_{t=1}^{T} R(t)]$. In this setup, the reward distribution of each arm $k \in [K]$ is assumed to be fixed over time and the rewards received by pulling the same arm are independently and identically distributed. An oracle (the best policy), in this case, would always play the arm with the highest mean reward. The difference between the expected cumulative reward obtained by the oracle and the one obtained by the learner is known to be the "regret." As is shown in existing literature (Lattimore and Szepesvári, 2020), the regret lower bound for stochastic K-armed bandits is $\Omega(\sqrt{T})$. Many existing algorithms achieve the regret at the optimal rate $O(\sqrt{T})$, including the Upper Confidence Bound (UCB) algorithm and the Thompson Sampling

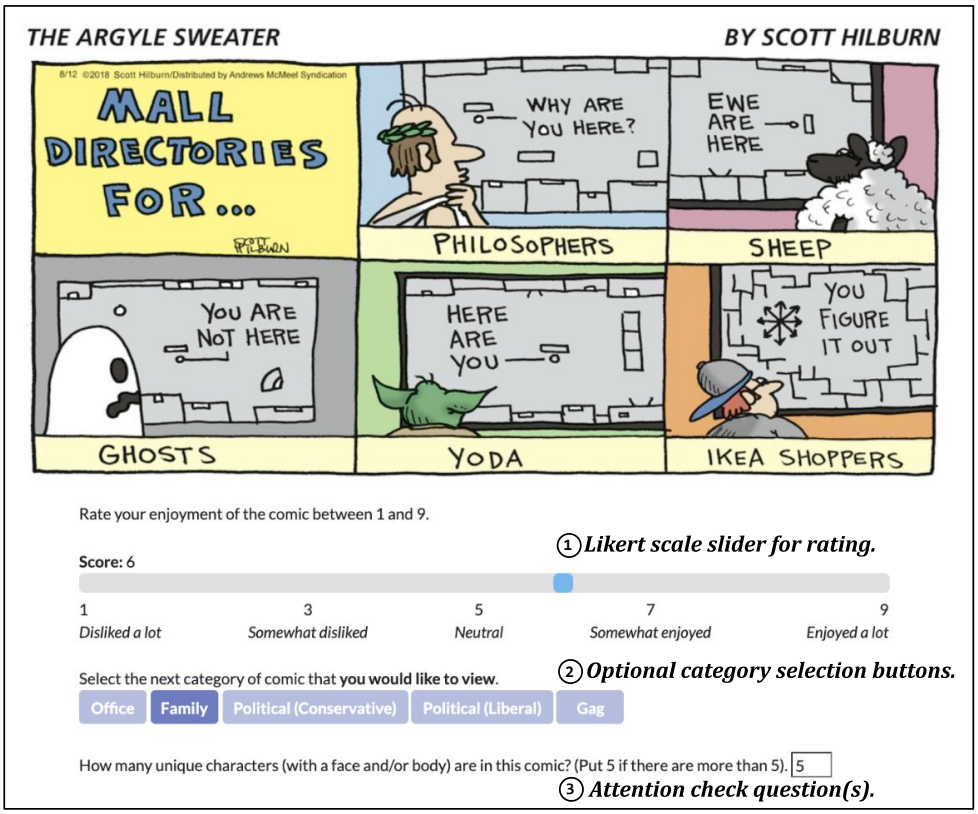


Figure 2.2: The user interface of our experimental platform when the participants read and rate a comic. For each comic, the participants have up to three action items to complete. First, they must provide the comic an enjoyment score (a rating) between 1 and 9 using the Likert scale slider bar, indicating how they like the comic. Second, if the participants are under the Self-selected setting, they must select the genre of comic they would like to view next. For other participants, this step does not exist. Finally, the participants are asked to answer one or more customized attention check questions.

(TS) algorithm. We also include a traditional algorithm Explore-then-Commit (ETC) and a greedy heuristic (ε-Greedy) in our study. These algorithms along with their regret guarantees are built upon the key assumption that the reward distributions are fixed. In Section 2.4, we test the validity of this assumption in a recommendation setting where the interactants are humans and the rewards represent their preferences.

Algorithms We give a brief summary and a high-level intuition for each algorithm. More detailed descriptions of these algorithms can be found in Appendix A.1 in the supplementary material. We use the term "algorithm" in a broad sense and the following items (e.g., Self-selected and Fixed sequence) may not all be traditional MAB algorithms.

- Self-selected: The participants who are assigned to the Self-selected algorithm will choose which arm to interact with by themselves. In other words, at each time t, instead of a prescribed learning policy determining the arm $a(t)$, the participants themselves will choose the arm.
- UCB: At time t, UCB deterministically pulls the arm with the highest upper confidence

bound, a value that for each arm, combines the empirical mean reward with an uncertainty estimate of the mean reward. An arm with high upper confidence bound can have high empirical mean and/or high uncertainty on the mean estimate.

- TS: In Thompson Sampling, a belief distribution is maintained over the possible reward values for each arm. At time t, the algorithm samples a reward from each arm's belief distribution and pulls the arm with the highest sampled reward. When a reward is received after pulling the arm, TS updates the corresponding arm's prior belief distribution to obtain a posterior.
- ETC: Unlike UCB and TS, the Explore-then-Commit algorithm has two separate stages—the exploration stage and the exploitation stage. It starts with an exploration period where the algorithm pulls the arms in a cyclic order and then switches to an exploitation stage where only the arm with the highest empirical mean in the exploration stage will be pulled. Given that the ETC algorithm achieves a regret of $O(T^{2/3})$ when the exploration time is on the order of $T^{2/3}$, we have set the exploration period to be $c \cdot T^{2/3}$ for a positive constant c.
- ε-Greedy: This greedy heuristic pulls the arm with the highest empirical mean with probability $1 - \varepsilon$ where $0 < \varepsilon < 1$, and pulls an arm uniformly at random with probability ε. In a setting with long interaction period, one way of setting ε is to have it decreasing over time, e.g., setting ε to be on the order of $\frac{1}{t}$ (Auer et al., 2002). Given the short interaction period in our setting, we have used a fixed $\varepsilon = 0.1$ (which may result in linear regret).
- Fixed sequence (CYCLE, REPEAT): The fixed sequence algorithms pull arms by following a predefined sequence. That is, the arm pulled at time t only depends on the current time step and does not rely on the received rewards so far. We have used two fixed sequences CYCLE and REPEAT for testing whether the reward distributions (that represent participants' preferences) are fixed over time. We provide more details on these two fixed sequences in Section 2.4.

Next, we present how we have selected the comics used for recommendations.

2.3.2 Comics data

In our experiment, we choose comics as our recommendation domain for the following reasons: (i) Fast consumption time: Given the nature of our experiment where study participants are recommended a sequence of items to consume in a limited amount of time, we require the time for consuming each of the recommendations to be short. For example, recommending movie clips to watch may not be appropriate in our setting given that each clip may take a couple of minutes to finish. (ii) No strong pre-existing preferences: Another important feature of the chosen recommendation domain is that the majority of the study participants should have no strong preference on that subject prior to the experiment. For example, unlike comics, music is a subject that most people tend to already have strong preferences towards (Schäfer and Sedlmeier, 2010). In such cases, the effects of recommendations towards the participants' preferences may be minimal.

We collected comics from 5 comic series on GoComics (GoComics, 2021). Each comic series belongs to a genre and represents an arm for pulling. The 5 comic series along with their genre are Lisa Benson (political, conservative), Nick Anderson (political, liberal), Baldo (family), The Born Loser (office), and The Argyle Sweater (gag). The genres of these comics are assigned by GoComics. For each series, we take the following steps to select the set of comics:

1. We first collect all comics belonging to the comic series from the year 2018. We select this time period to be not too recent so that the results of the study are not heavily influenced by ongoing events. It is also not too distant so that the content is still relevant to all subjects.
2. For each individual comic, we obtain its number of likes on GoComics. Then, we choose the top 60 comics from each comic genre/series in terms of the number of likes. This selection criteria is designed to ensure the quality of the chosen comics.
3. Finally, we randomly assign an ordering to the comics. We keep this ordering fixed throughout the study such that if an arm is pulled (a genre is chosen) at its j-th time, the presented comics will always be the same.

The comics within the same comic series are independent, in the sense that they can generally be read in any order, without requiring much context from previous comics from the same series. For these collected comics, we have labeled the number of unique characters in them, and use it for the attention check questions to ensure that the study participants have read the comics. Although we have adopted the above selection criteria to ensure the quality of comics from the same comic series to be similar, there is heterogeneity among individual comics and thus may influence the interpretation of our results. We provide more discussion on this in Section 2.6. In Section 2.3.3, we discuss our experimental protocol and platform.

2.3.3 Experimental protocol and platform

We first outline our experimental protocol, which consists of the following steps (Figure 2.1):

1. *Background survey (initial filtering)*: We ask the study participants to complete a brief background survey (Appendix A.3 in the supplementary material). The participants will only be given a code to continue to the next step if an arithmetic question is answered correctly. The goal for the first step is to set up an initial filtering for participants.
2. *Registration*: After completing the background survey, participants are then asked to register on our platform using their completion code. Each participant in our study is assigned an algorithm in the following fixed probabilities—0.25 for Self-selected, 0.125 for UCB, 0.125 for TS, 0.125 for ETC, 0.125 for ε-Greedy and 0.125 for each of the two fixed sequences.
3. *Comic rating*: In this step, participants will read a sequence of 50 comics and provide an enjoyment score (a rating) after reading each comic. The sequence of comics can be generated by any of the algorithms discussed in Section 2.3.1. After reading each comic and providing a rating, the participants are also asked to answer an attention check question.
4. *Post-study survey*: Once the participants are finished with the comics rating step of the

study, they are asked to complete a post-study survey about their reading experience. They are asked if they remember reading certain comics and if they have rated them positively, as well as how they perceive the recommendations they are provided. An example of the post-study survey questions can be found in Figure A.2 (Appendix A.4 in the supplementary material).

Our experimental platform is built as a toolkit for running field tests for any MAB algorithms with human users. It consists of (a) the participant-facing web interface, and (b) the server backend that stores and processes incoming data from the web interface. When designing the experimental protocol and platform, we consider the following questions:

1. Given that we are asking users to give subjective feedback, how do we have more user responses that are reflective to the user's true preference?
2. How do we design an experimental interface that impose less bias to the users?
3. Since our study requires users to complete a long (up to 30-minute) sequence of non-independent tasks, how do we have the study to be less interrupted?
4. How do we build the system flexible enough to conduct studies for different MAB algorithms and test different assumptions of MAB setups, including ones that we do not cover in this work?

For (1), we adopt a 9-point Likert scale so that the numbers are sufficiently distinguishable to the participants (Cox III, 1980) and check whether users are paying sufficient attention during the study. In particular, we test the participants on objective properties of the comics they have read, e.g. the number of unique characters (with a face and/or body) in them. We also set a minimum time threshold of 10 seconds before each user response can be submitted so that users spend adequate time on each comic.

For (2), to ensure that the participant's rating is not biased towards (e.g., anchored on) the Likert scale slider's initial value, we set the slider to be transparent before the participant clicks on the scale. In addition, in the Self-selected setting where the participants choose the genre of comics to read next, we minimize the color- and ordering-based biases by setting the category selection buttons to be the same color and in random order.

As stated in design question (3), because we are interested in studying evolving preferences over a sequence of non-independent tasks, we would like to have *continuous and uninterrupted* user attention over a period of time. To do so, we design the system to be stateful so that participants can resume the study where they left off in the event of brief network disconnection and browser issues.

Finally, we discuss the flexibility of our system and address design question (4). Our experimental platform allows the experimenter to specify any recommendation domains and MAB algorithms for interacting with the human interactant. This flexibility allows the experimenter to not only test the performance of different MAB algorithms but also design pull sequences to understand user preference dynamics and test existing assumptions on user preferences. For example, one may design pull sequences to study the correlation between rewards obtained at different time steps and rewards obtained from pulling different but related arms. It is worth noting that the attention checks in our system are also customizable, allowing for diverse atten-

	Self-selected	UCB	TS	ETC	ε-Greedy	CYCLE	REPEAT
# of Participants	74	40	44	39	41	40	38

Table 2.1: Number of participants for each algorithm. The description for Self-selected, UCB, TS, ETC and ε-Greedy are in Section 2.3.1. CYCLE and REPEAT are defined in Section 2.4.

tion checks for each recommendation domain. For more details about the experimental platform and MTurk-related implementation details, we refer the readers to Appendix A.4.

2.3.4 Recruitment and compensation

To ensure the quality of our collected data, we only allow MTurk workers to participate in the study if they are U.S. residents and have completed at least 500 Human Intelligence Tasks (HITs) with an above 97% HIT approval rate. The anticipated (and actual) time to complete the study is less than 30 minutes. For participants who have successfully completed the study and answered the attention check questions correctly 70% of time, we have paid $\$7.5$ (the equivalent hourly salary is above $\$15$/hr). In Appendix A.3, we provide more details on the participants' demographics and backgrounds. Out of the 360 participants who have successfully completed the study, 316 passed the attention check (acceptance rate 87.8%). Our analyses are only conducted on the data collected from these participants. Table 2.1 shows the number of participants for each experimental setup.

2.4 Evolving Preferences in K-armed bandits

As we have previously discussed, in K-armed bandits, the reward distribution of each arm is assumed to be fixed over time (Robbins, 1952; Slivkins, 2019; Lattimore and Szepesvári, 2020), which implies that the mean reward of each arm remains the same. It is unclear whether such an assumption is reasonable in recommender system settings where the reward distributions represent human preferences. Our first aim is to test for the existence of evolving preference in the K-armed bandits setup. In other words, using randomized controlled trials, we want to test the following hypothesis: *In a K-armed bandit recommendation setting, the reward distribution of each arm (i.e., the user preference towards each item) is not fixed over time.*

To answer this, we collect enjoyment scores for two fixed recommendation sequences, where each sequence is of length $T = 50$. The sequence consists of recommendations from $K = 5$ genre of comics. In other words, the total number of arms is 5. The first sequence pulls the arms in a cyclic fashion which we denote by CYCLE, while the second sequence pulls each arm repeatedly for $m = T/K$ times which we denote by REPEAT:

$$\text{CYCLE}: \ (\underbrace{12\ldots K12\ldots K\ldots 12\ldots K}_{12\ldots K \text{ for } m \text{ times}}),$$

$$\text{REPEAT}: \ (\underbrace{22\ldots 2}_{m \text{ times}}\underbrace{1\ldots 1}_{m \text{ times}}3\ldots 3\ldots K\ldots K).$$

We note that for both sequences, the number of arm pulls of each arm is the same (m times). Since the order of the comics is fixed for each arm (e.g., pulling an arm for m times will always result in the same sequence of m comics from that genre), the set of comics recommended by the two sequences are the same. The only difference between the two sequences is the order of the presented comics. Intuitively, if the mean reward of each arm is fixed over time (and does not depend on the past pulls of that arm), then the (empirical) mean reward of each arm should be very similar under the two pull sequences.

In this work, we utilize a modification of the two-sample permutation test (Fisher, 1936) to deal with the different numbers of participants under the two recommendation sequences. We let $\mathcal{P}^{\text{CYCLE}}$ and $\mathcal{P}^{\text{REPEAT}}$ denote the set of participants assigned to the CYCLE and REPEAT recommendation sequence, respectively. For each participant $i \in \mathcal{P}^{\text{CYCLE}}$, we use $a_i(t)$ to denote the pulled arm (the recommended comic genre) at time t and $X_i(t)$ to denote the corresponding enjoyment score (the reward) that the participant has provided. Similarly, for each participant $j \in \mathcal{P}^{\text{REPEAT}}$, we use $a_j(t)$ and $Y_j(t)$ to denote the arm pulled at time t and the enjoyment score collected from participant j at time t. Using these notations, for each arm $k \in [K]$, we define the test statistic as follows:

$$\tau_k = \underbrace{\frac{1}{|\mathcal{P}^{\text{CYCLE}}|} \sum_{i \in \mathcal{P}^{\text{CYCLE}}} \left(\frac{1}{m} \sum_{t \in [T]: a_i(t) = k} X_i(t) \right)}_{M_k^{\text{CYCLE}}} - \underbrace{\frac{1}{|\mathcal{P}^{\text{REPEAT}}|} \sum_{j \in \mathcal{P}^{\text{REPEAT}}} \left(\frac{1}{m} \sum_{t \in [T]: a_j(t) = k} Y_j(t) \right)}_{M_k^{\text{REPEAT}}}.$$

The test statistic τ_k is the difference between the mean reward (enjoyment score) M_k^{CYCLE} under the CYCLE recommendation sequence and the mean reward M_k^{REPEAT} under the REPEAT recommendation sequence for arm k. A non-zero τ_k suggests that the mean reward of arm k is different under CYCLE and REPEATand that the reward distribution is evolving. The higher the absolute value of τ_k is, the bigger the difference between the two mean rewards is. A positive test statistic value indicates that the participants prefer the arm under CYCLE over REPEAT on average.

To quantify the significance of the value of the test statistic, we use a two-sample permutation test to obtain the p-value of the test (Fisher, 1936): First, we permute participants between $\mathcal{P}^{\text{CYCLE}}$ and $\mathcal{P}^{\text{REPEAT}}$ uniformly at random for $10,000$ times and ensure that the size of each group remains the same after each permutation. Then, we recompute the test statistic τ_k for each permutation to obtain a distribution of the test statistic. Finally, we use the original value of our test statistic along with this distribution to determine the p-value.

To report the 95% confidence interval of the test statistic for each arm, we use bootstrap and re-sample the data for $5,000$ times at the level of arm pulls. That is, for each arm pull at time t, we obtain its bootstrapped rewards under CYCLE and REPEAT by resampling from the actual rewards obtained by pulling that arm under CYCLE and REPEAT at time t, respectively.

		family	gag	political (conservative)	office	political (liberal)
Overall	τ_k value	0.290	0.132	0.445	0.047	0.573
	95% CI	$[0.042, 0.549]$	$[-0.096, 0.371]$	$[0.158, 0.744]$	$[-0.196, 0.290]$	$[0.301, 0.837]$
	p-value	0.025^*	0.275	0.004^*	0.694	$< 0.001^*$
Heavy	τ_k value	-0.394	-0.556	-0.694	-0.647	-0.664
	95% CI	$[-0.678, -0.114]$	$[-0.864, -0.243]$	$[-1.056, -0.325]$	$[-0.944, -0.350]$	$[-1.021, -0.302]$
	p-value	0.007^*	$< 0.001^*$	$< 0.001^*$	$< 0.001^*$	$< 0.001^*$
Light	τ_k value	0.784	0.635	1.274	0.552	1.475
	95% CI	$[0.421, 1.150]$	$[0.298, 0.977]$	$[0.860, 1.681]$	$[0.198, 0.905]$	$[1.120, 1.827]$
	p-value	$< 0.001^*$	$< 0.001^*$	$< 0.001^*$	0.001^*	$< 0.001^*$

Table 2.2: The difference between the mean reward under CYCLE and the mean reward under REPEAT for each arm. All results are rounded to 3 digits. The p-values are obtained through permutation tests with $10,000$ permutations. We use asterisk to indicate that the test is significant at the level $\alpha = 0.1$. The 95% confidence intervals are obtained using bootstrap with $5,000$ bootstrapped samples.

Given that we have conducted 5 tests simultaneously (one for each arm), in order to control the family-wise error rate, we need to correct the level α_k $(k \in [K])$ for each test. More formally, to ensure the probability that we falsely reject *any* null hypothesis to be at most α, for each test, the p-value of a test should be at most its corresponding corrected α_k. We adopt the Holm's Sequential Bonferroni Procedure (details presented in Appendix A.2) to perform this correction (Abdi, 2010).

Our results show that for three arms—family, political (conservative) and political (liberal)—the non-zero difference between the mean reward under the two recommendation sequences are significant at level $\alpha = 0.1$ (Table 2.2). These findings confirm our research hypothesis that user preferences are not fixed (even in a short amount of time) in a K-armed bandit recommendation setting. There may be many causes of the evolving reward distributions (preferences). One possibility, among many others, is that the reward of an arm depends on its past pulls. In other words, people's preference towards an item depends on their past consumption of it. For example, existing marketing and psychology literature has suggested that people may experience hedonic decline upon repeated exposures to the same item (Galak and Redden, 2018; Baucells and Sarin, 2007). On the other hand, in music play-listing, one may expect the expected reward of an arm (a genre) to increase due to the taste that the listener has developed for that genre of music (Schäfer and Sedlmeier, 2010). For a more comprehensive discussion on preference formation, we refer the readers to Becker (1996).

Finally, to better understand the nature of our findings, we divide the participants into heavy comic readers who read comics daily and light comic readers who read comics at a lower frequency. Among participants who are assigned the CYCLE sequence, there are 17 heavy readers and 23 light readers. For REPEAT, there are 16 heavy readers and 22 light readers. We perform the same analysis as noted above for each of the two groups. Similar to the overall findings, among both heavy and light readers, evolving preferences (evolving mean reward) have been observed (Table 2.2), confirming our research hypothesis. Interestingly, we find that for each genre, the heavy readers tend to prefer the recommendations from the REPEAT sequence over the CYCLE sequence. On the contrary, for each genre, light readers prefer recommendations

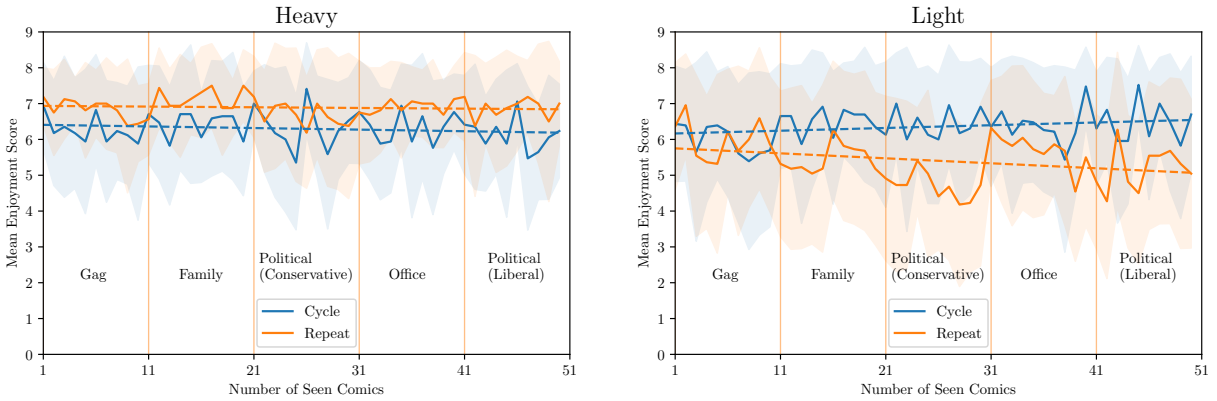


Figure 2.3: The reward at each time step of the CYCLE and REPEAT recommendation sequence averaged across heavy and light readers, respectively. The error bars indicate one standard deviation from the mean. For the REPEAT sequence, we highlight when the arm switches using the vertical lines and add the arm name (comic genre) corresponding to the time period in between the switches using the black texts. The blue and orange dotted lines are fitted through the rewards collected under CYCLE and REPEAT, respectively.

from the CYCLE sequence over the REPEAT sequence. As an initial step towards understanding this phenomenon, we present descriptive data analysis on how rewards (user preferences) evolve for heavy and light readers under the two recommendation sequences. Similar to our results in Table 2.2, for light readers, at most time steps, the reward trajectory of CYCLE has a higher value than the reward trajectory of REPEAT; while for heavy readers, this is the opposite (Figure 2.3). By looking at the reward trajectories (and the lines fitted through the reward trajectories) over the entire recommendation sequence (Figure 2.3), we find additional trends: for light readers, the differences between the rewards collected under CYCLE and REPEAT are increasing over time; while for heavy readers, such differences are relatively stable. This trend is also observed in the reward trajectory (and the line fitted through the reward trajectory) of each arm under CYCLE and REPEAT (Figure 2.4). In particular, for light readers, among all arms (comic genres) except family, we find that the lines fitted through the rewards collected under CYCLE and REPEAT become further apart as the number of arm pulls increases. This distinction between heavy and light users may be due to various reasons. For example, light readers may prefer variety in their recommendations because they are exploring their interests or the light readers and heavy readers share different satiation rates. On a related note, a recent study has shown that the satiation rates of people may depend on their personality traits (Galak et al., 2022). Precisely understanding the causes of this distinction between heavy and light readers is of future interest.

2.5 Usage Example of the Experimental Framework

As an illustration of the usage of our experimental framework, we compare the performance of different algorithms, in terms of both the cumulative rewards, and the users' own reflection on their interactions with these algorithms. Though we are in a simulated and simplified

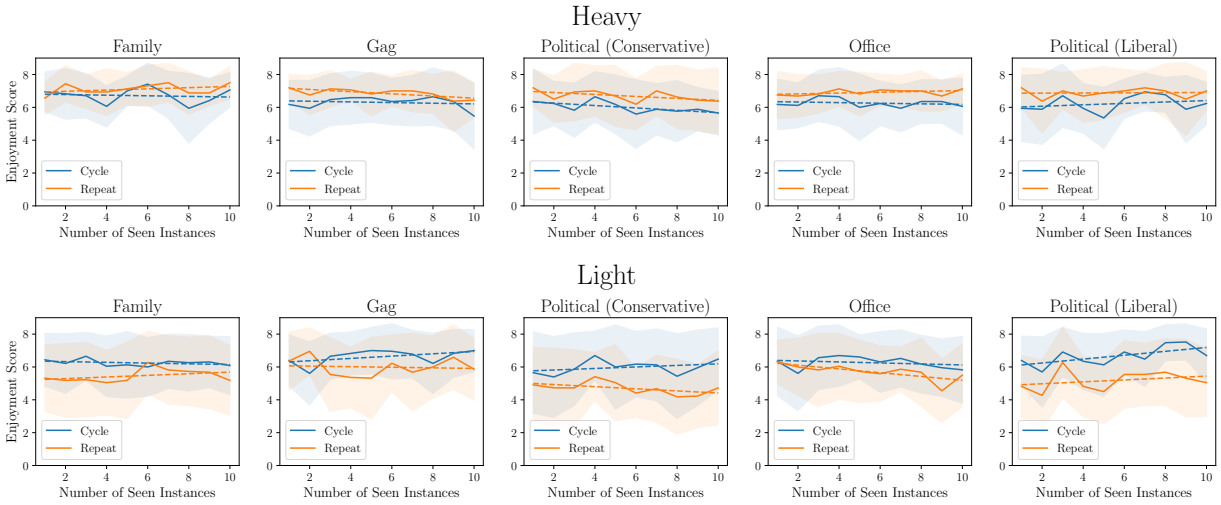


Figure 2.4: Each plot shows the reward collected for a particular arm at each arm pull under the CYCLE and REPEAT recommendation sequences. The error bars indicate one standard deviation from the mean. The reward trajectories are averaged across heavy and light readers, respectively. The blue and orange dotted lines are fitted through the reward trajectories for each arm under CYCLE and REPEAT, respectively.

	Self-selected	UCB	TS	ETC	ε-Greedy	CYCLE	REPEAT
Cumulative reward	319.36 [205, 434]	323.70 [223, 424]	312.37 [204, 420]	324.49 [209, 440]	307.59 [205, 411]	316.5 [216, 417]	301.63 [201, 427]
Hindsight satisfaction	81.08%	75.00%	72.73%	76.92%	80.49%	77.50%	63.16%
Preference towards autonomy	71.62%	50.00%	68.18%	58.97%	58.54%	62.50%	65.79%

Table 2.3: Performances of each algorithm in terms of different enjoyment characterizations. The first row gives the cumulative rewards for each algorithm, averaged over participants who have interacted with it. We also report the 95% confidence intervals for the cumulative rewards. The hindsight-satisfaction row shows the percentage of participants who believe the sequence of comics they have read captures their preference well. The last row provides the percentage of participants who prefer to select comics to read on their own in hindsight.

recommender system setup, we aim to provide some understanding on (i) whether people prefer to be recommended by an algorithm over deciding on their own, and (ii) whether more autonomy (choosing the next comic genre on their own) results in a more attentive and mindful experience. We note that, compared to our findings in Section 2.4 that are obtained through a rigorous hypothesis testing framework, the results in this section are exploratory in nature and should not be interpreted as definitive answers to the above questions.

2.5.1 Enjoyment

We first compare different algorithms (Self-selected, UCB, TS, ETC, ε-Greedy, CYCLE and REPEAT) in terms of the participants' enjoyment. More specifically, we want to compare the participants who are provided recommended comics to read with the ones who choose on their

	Self-selected	UCB	TS	ETC	ε-Greedy	CYCLE	REPEAT
Reading Memory	91.89%	90.00%	92.42%	93.16%	90.24%	90.00%	90.35%
Rating Memory	71.17%	70.00%	70.45%	76.92%	69.92%	69.16%	57.89%

Table 2.4: Average correctness of the two types of memory questions for each algorithm.

own. In general, enjoyment is hard to measure (Payne et al., 1999). To this end, we look at participants' enjoyment and preference towards these algorithms through the following three aspects:

- Cumulative rewards: This metric is closely related to the notion of "regret" that is commonly used to compare the performance of bandit algorithms (Lattimore and Szepesvári, 2020). More formally, for any participant i, the cumulative reward of an algorithm that interacts with participant i is given by $\sum_{t=1}^{T} R_i(t)$ where $R_i(t)$ is the reward provided by participant i at time t. In Table 2.3, for each algorithm, we show their cumulative rewards averaged over participants who have interacted with the algorithm.
- Hindsight satisfaction: After the participants interact with the algorithm, in the post-study survey, we ask them the following question: "Do you feel that the sequence of recommendations[2] captured your preferences well?" Participants' answers to this question provide their hindsight reflections towards how well the algorithm performed and whether they are satisfied with it. The second row of Table 2.3 shows the percentage of participants who believe the sequence of comics they read has captured their preference well.
- Preference towards autonomy: In addition to the previous two metrics, we have explicitly asked the participants to indicate whether they prefer to choose comics to read on their own. In the post-study survey, we ask: "Do you prefer being recommended comics to read or selecting comics to read on your own?" The third row of Table 2.3 provides the percentage of participants who prefer to select comics on their own for each algorithm.

In our collected data, the participants who have given more autonomy (the Self-selected participants) prefer more autonomy in hindsight, compared to other participants. Though Self-selected does not have the highest mean cumulative reward, it has the highest percent of participants who believe that the comics they read have captured their preferences well in hindsight (Table 2.3). This misalignment between mean cumulative reward and hindsight satisfaction also shows in other algorithms (i.e., higher mean cumulative reward may not indicate higher hindsight satisfaction), including ε-Greedy and ETC. On the other hand, UCB performs well in terms of the mean cumulative reward, while having the lowest percentage of participants wanting to have more autonomy.

2.5.2 Attentiveness

We want to understand whether participants who are asked to choose on their own have a more mindful experience, in which the participants are more attentive to what they have read

[2]For Self-selected participants, this should be interpreted as comics they have read/self-selected.

through (Epstein, 1999). There is no consensus on defining and measuring mindfulness of an experience (Grossman, 2008). Some of the existing research uses self-reported mindfulness as a measurement (Greco et al., 2011). In our case, we look at how attentive the participants are to their own experience, though the lens of memory—for each participant, in the post-study survey, we ask them two types of memory questions:

- Reading memory: In the post-study survey, we present the participants three randomly selected comics and ask them to indicate whether they have read the comics before. This question aims to measure the participants' memory of *what they have read*. The first row of Table 2.4 shows the average correctness percentage for the reading memory questions for each algorithm.
- Rating memory: We also look at the participants' memory on *how they have liked* certain comics. To this end, in the post-study survey, we present three randomly selected comics among the ones the participants have read and ask them to indicate whether they have rated the comic with a score of five or above. The second row of Table 2.4 shows the average correctness percentage.

We note that these questions differ from the attention check questions after each comic and are not directly related to the compensation that the participants get. However, the high correctness percentage shown in Table 2.4 on the reading memory questions suggest that the participants have attempted to answer the questions from their memory instead of providing random answers.

In general, the participants have performed much better on the reading memory questions than the rating memory questions, suggesting that they may be more aware of what they have consumed than how they have liked them. In addition, all participants perform similarly in terms of the reading memory correctness percentage with those whose algorithm is ETC or Self-selected performing slightly better. Though it is hard to say which algorithm provides the most attentive experience to the participants and whether Self-selected participants have a more mindful experience, it is relatively clear that participants whose algorithm is REPEAT perform the worst in terms of rating memory. Our data does not provide strong evidence on believing that more autonomy results in more attentive experience, but may suggest that less enjoyable experience (e.g., for participants whose algorithm are REPEAT) correlates to less attentiveness.

2.6 Discussion

Our work provides a general experimental framework and toolkit to understand assumptions on human preferences in a bandit setup. It also allows one to perform field tests of different bandit algorithms that interact with humans. Using these tools, we build a publicly available dataset that contains trajectories of recommendations and ratings of them given by multiple users in a bandit setting. The collected data has been used to check the validity of a core assumption on human preference in MABs literature—that the reward distributions (corresponding to user preferences) are fixed. We show that even in a short time period like our bandit recommendation setup, such dynamical preference exists. Further, our observations on the difference between the light and heavy user in their preference dynamics suggest the need of having a more

granular understanding of human preferences. As an illustrative usage of our experimental framework, we have explored the study participants' preferences towards selecting content to read on their own and being recommended content to read. As we have discussed above, these findings are exploratory in nature with the goal of showcasing different usages of our experimental framework; thus, they should not be interpreted as definitive answers. In our exploratory analysis, we observe that an algorithm achieving the highest cumulative reward does not necessarily imply that it will achieve the highest hindsight satisfaction.

At a higher level, our work fits in the broader picture of understanding the validity of assumptions that machine learning systems (and in our case, recommender systems) rely on. Although all machine learning models are built upon some simplifying assumptions of the world, some of these assumptions oversimplify the world to the extent that they lose the core characterizations of the problem we are interested in. In our case, the assumptions we want to understand are centered around user preferences. We have identified that assumptions on the temporal stability of preferences used in traditional MABs literature are oversimplifications. The balance between identifying simplifications that are helpful for building learning systems and avoiding oversimplifications that discard core characteristics of the problem is difficult. As put by the famous statistician George Box, "all models are wrong, but some are useful" (Box, 1979). Our goal for developing the experimental toolkit and conducting the human subjects study is to provide ways for identifying assumptions on user preferences that are useful for developing MABs algorithms in recommender systems. Below we discuss the limitations and future directions of our work.

2.6.1 Limitations

We discuss several limitations of our study. First, the sizes of the mean reward differences of each arm reported in Section 2.4 are within 1.5 point on the 9-point Likert scale, which may be considered small. This is due to many reasons. For one, the enjoyment score (the reward) provided by each participant is subjective and may have high variance due to this subjectivity. Though a difference may be considered to be strong in a within-subjective study, it may be thought of as small in a between-subject study (the type of study in our case). However, we note that our results obtained using the permutation test show that reward distributions are indeed not fixed over time for multiple arms in the K-armed bandit recommendation setup. Second, each arm represents a comic series. Though we have selected the comics from each series in terms of their quality (the number of likes that the comics receive), there may still be heterogeneity among the selected comics belonging to the same series, and it is up to discussion on whether one should consider these comics to belong to the same arm. Thirdly, many quantities (e.g., enjoyment/satisfaction, mindfulness/attentiveness) we want to measure are less well-defined. Our way of measuring them is from a particular angle, and may not be widely applicable. Fourthly, the study domain is chosen to be comics due to reasons including the short consumption time of a comic and the study requires the participants to read 50 comics. Although all of our experiments are completed within 30 minutes, it is uncommon in real-world settings for people to read 50 comics at a time and thus may introduce uncontrolled boredom effects. Fifthly, in our experiments, we compare the user preferences towards each arm using two fixed sequences.

One may also utilize a purely randomized sequence as a baseline to better understand user preferences. Finally, due to resource constraints (e.g., funding limits for conducting the study and computational limits on the number of simultaneous experiment instances we can host on our server), we recruited 360 participants (with 316 of them passing the attention checks) for our study. Compared to industry-scale experiments, the number of participants in our study is on the lower end. It is also worth mentioning that our implementations of the bandit algorithms ensure that the comic recommendations only depend on the user's own history, which is not a common practice for recommender systems on existing platforms. In practice, one may utilize other users' interaction histories to warm start these bandit algorithms. Though there are these limitations, we would like to emphasize that our work makes a substantive step towards understanding the applicability of traditional assumptions on user preferences in MABs literature.

2.6.2 Future Work

There are multiple future directions for our work. Our findings on the existence of evolving preferences in a K-armed bandits recommendation setting suggest that in order to study the decision-theoretic nature of recommender systems using the MAB framework, one must account for such preference dynamics. The need for learning algorithms (oftentimes reinforcement learning algorithms) that deal with the impact of recommendations on user preferences have also been proposed in recent works (Zheng et al., 2018; Ie et al., 2019; Shani et al., 2005a; McInerney et al., 2018; Chen et al., 2019; Mehrotra et al., 2020; Chen et al., 2021a; Chen et al., 2021b; Wang et al., 2022; Chen et al., 2022). An important building block for this line of research is to have better modeling of human preference dynamics. Our experimental framework and toolkit can provide more grounding and accessibility for research on it. As an example, our observation that heavy and light comic readers have different preference dynamics can be further investigated using our experimental framework, advancing our understanding on evolving preferences in a more granular way. More broadly, our toolkit can be used for: (i) collecting data using fixed or randomized recommendation sequences or bandit algorithms to identify and estimate preference dynamics; and (ii) conducting field tests of bandit algorithms designed to address evolving preferences.

Our exploratory data analysis on the performance of different algorithms suggests that the human interactants of the bandit algorithms may care about other aspects of their experience in addition to cumulative rewards. For example, Self-selected has a higher percentage of satisfied participants in hindsight compared to UCB, though UCB has a higher average cumulative reward. This suggests that besides traditional performance metrics used to analyze and develop these bandit algorithms, we should consider a broader set of objectives and metrics when studying these problems. On a related note, given that we want our algorithm to account for evolving preferences, when regret (the difference between the expected cumulative reward obtained by a proposed policy and an oracle) is used as the performance metric, the oracle should be chosen to be adaptive instead of the best-fixed arm considered in many MABs (including contextual bandits) literature.

Chapter 3

On Human-Aligned Risk Minimization

3.1 Introduction

The decision-theoretic foundations of modern machine learning models have largely focused on estimating model parameters that minimize the expectation of some loss function. This ensures that the resulting model have high average case performance, which loosely is what is meant by good generalization performance. However, as ML models are increasingly deployed in broader societal settings, and in particular, to assist humans in decision-making, it is clear that humans want models to have not just good average performance but also properties like fairness. Due to the importance of these additional desiderata, there have been a burgeoning interest in capturing these properties via appropriate constraints and modifications of the classical objective of expected loss (Kamishima et al., 2011; Garcıa and Fernández, 2015; Sra et al., 2012; Duchi et al., n.d.). In this work, we posit a very natural if simple solution to addressing these varied desiderata that are driven in large part by human considerations. Specifically, we suggest that in contrast to using the standard workhorse of expected loss, we draw from theories of human cognition in psychology and behavioral economics, to consider a "human-aligned" risk instead.

Alternatives to expected loss based risk measures have a long history in decision-making (Howard and Matheson, 1972), with earlier efforts focusing on percentile risk criteria (Filar et al., 1995). In machine learning, instead of minimizing expected loss, various risk measures have been considered in different settings. In risk sensitive reinforcement learning, conditional value-at-risk (CVaR), a percentile risk measure that quantifies the tail performance of a model, has been connected to robustness to modeling errors (Chow et al., 2015; Osogami, 2012). Recently, human-aligned risk measures have also been explored in bandit (Gopalan et al., 2017) and reinforcement learning (Prashanth et al., 2016), where the goal of the agent is to produce long term returns aligned with the preferences of one or more humans.

Contributions. In this work, we introduce a novel notion of human risk minimization (Section 3.3), by bringing ideas from cumulative prospect theory (Section 3.2) into supervised learning. We explore various salient characteristics of our objective such as diminishing sensitivity, decision-making based on higher-order moments and information-content or "surprisal"

view point of human risk (Section 3.4). We also study the implications of minimizing our objective in the context of subpopulation performance (Section 3.5). In particular, our empirical results illustrate that human risk minimization inherently avoids drastic losses across all subgroups.

3.2 Background

3.2.1 Cumulative Prospect Theory

As a seminal work in behavioral economics, cumulative prospect theory (CPT) (Tversky and Kahneman, 1992) provides a framework to emulate human decision-making under uncertainty. In particular, CPT points out that humans overweight extreme events that occur with low probability, rather than treating all the events equally, which is the assumption of expected utility theory (EUT). As an alternative to EUT, CPT has three important components (Rieger and Wang, 2006):

- Outcomes are considered as gains or losses compared to a reference point;
- Value functions are concave for gains, convex for losses and flatter for gains than for losses;
- An inverse S-shaped (first concave then convex) probability weighting function (Figure 3.1 (a)) is used to transform the cumulative distribution function so that small probabilities are inflated and large probabilities are deflated (Wu and Gonzalez, 1996).

Current machine learning follows the EUT framework in the sense that expected losses are minimized. However, as CPT has pointed out, a human evaluates risk differently. For example, given two models $\{\mathcal{M}_1, \mathcal{M}_2\}$ such that $\mathcal{M}_1$ has zero loss with probability $.95$ and loses 100 with probability $.05$ while $\mathcal{M}_2$ loses 5.01 all the time, EUT will choose $\mathcal{M}_1$. The reasoning behind is that the expected loss of $\mathcal{M}_1$ is 5, which is smaller than the expected loss of $\mathcal{M}_2$, which is 5.01. However, from the CPT perspective, $\mathcal{M}_2$ will be chosen because CPT inflates the probability $.05$ and $\mathcal{M}_1$ will end up having a larger risk. In this case, because of the human-innate probability weighting, we end up choosing a model that avoids drastic losses instead of the one with a better average performance.

The inverse S-shaped CPT probability weighting function captures that humans over-weight extreme events with low probability while under-weight "average" events that are more probable but less extreme. Many parametric forms of the probability weighting function have been proposed (Tversky and Kahneman, 1992; Prelec et al., 1998; Wu and Gonzalez, 1996; Rieger and Wang, 2006). To start with, we formally define a class of weighting functions called $\mathcal{W}_{\text{CPT}}$.

Definition 3.1

Let $w : [0, 1] \to [0, 1]$ be a differentiable function. Then, $w \in \mathcal{W}_{\text{CPT}}$ if and only if

1. $w(0) = 0$ and $w(1) = 1$;
2. there exists $a \in (0, 1)$ such that $w(a) = a$;

3. $w'(x)$ is monotonically decreasing on $x \in [0, a)$ and $w'(x)$ is monotonically increasing on $x \in (a, 1]$.

Traditional CPT probability weighting functions fall into this class, including the original weighting function (for losses) $w(x) = \frac{x^{.69}}{(x^{.69}+(1-x)^{.69})^{1/.69}}$ (Tversky and Kahneman, 1992). For a real-valued continuous random variable X with cumulative distribution function $F(x)$ and a CPT probability weighting function $w \in \mathcal{W}_{\text{CPT}}$, the CPT subjective utility is defined as (Tversky and Kahneman, 1992; Rieger and Wang, 2006):

$$U_{\text{CPT}}(X) = \int_{-\infty}^{+\infty} v(x) dg(F(x)) \tag{3.1}$$

where (1) $v : \mathbb{R} \to \mathbb{R}$ is a value function; (2) $g(F(x)) = w(F(x))$ when $x < 0$ and $g(F(x)) = -w(1 - F(x))$ when $x \geq 0$.

Rank-dependent Utility. As pointed out in (Tversky and Kahneman, 1992), CPT subjective utility is a rank-dependent utility since the decision weight on x depends on the "rank" of x, which is given by $F(x)$. When $F(x)$ is weighted by $w \in \mathcal{W}_{\text{CPT}}$ (Diecidue and Wakker, 2001) and $v(x) = x$, the CPT-weighted rank-dependent utility is:

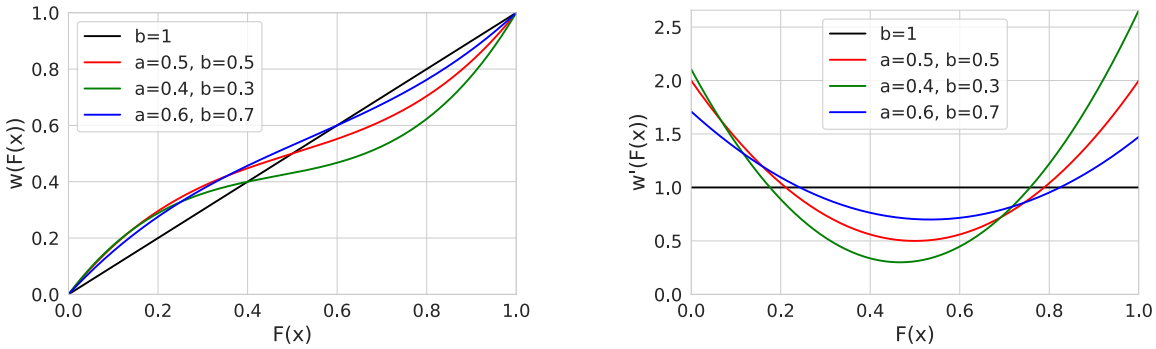
$$U_{\text{CPT-RD}}(X) = \int_{-\infty}^{+\infty} x dw(F(x)). \tag{3.2}$$

We focus on studying the effect of using CPT-weighted cumulative distribution function $w(F(x))$ on training an ML model. Hence, analyzing the effect of using a reference point and a value function v is out of the scope of this paper. As one may have noticed, if $w(F(x)) = F(x)$, then $U_{\text{CPT-RD}}(X) = \mathbb{E}[X]$.

To have a finite CPT subjective utility (Rieger and Wang, 2006) for a real-valued continuous random variable X with $w \in \mathcal{W}_{\text{CPT}}$, it is sufficient to ensure w to be strictly increasing on $[0, 1]$ and continuously differentiable on $[0, 1]$, i.e. $w'(0)$ and $w'(1)$ are finite. As proposed by (Rieger and Wang, 2006), the simplest polynomial that satisfies the above conditions is

$$w_{\text{POLY}}(F(x)) = \frac{3 - 3b}{a^2 - a + 1} \left(F(x)^3 - (a + 1)F(x)^2 + aF(x)\right) + F(x) \tag{3.3}$$

where $a \in (0, 1)$ is the fixed point, i.e. $w_{\text{POLY}}(a) = a$, and $b \in (0, 1)$ controls the curvature of $w_{\text{POLY}}(\cdot) \in \mathcal{W}_{\text{CPT}}$. As b approaches 1, $w_{\text{POLY}}(F(x))$ will converge to the linear function $w(F(x)) = F(x)$. One could interpret b as controlling the sensitivity of the probability weighting function to a unit difference in probability changes (Gonzalez and Wu, 1999). We use $w_{\text{POLY}}(\,\cdot\,; a, b)$ to denote the polynomial form CPT probability weighting function with fixed point a and curvature b.



(a) $w_{\text{POLY}}(F(x))$ **(b)** $w'_{\text{POLY}}(F(x))$

Figure 3.1: (a) Inverse S-shaped probability weighting function w_{POLY} is the steepest near the endpoints $0, 1$. The parametric form of $w_{\text{POLY}}(\,\cdot\,; a, b)$ is shown in Equation 3.3. (b) U-shaped CPT probability weighting function derivative w'_{POLY} up-weights the tails of the original distribution.

Proposition 1. *Given any cumulative distribution function $F(x)$, if a non-decreasing continuous function $w : [0,1] \to [0,1]$ satisfies $w(0) = 0$ and $w(1) = 1$, then $w(F(x))$ is a cumulative distribution function of some random variable.*

For any $w \in \mathcal{W}_{\text{CPT}}$ that is non-decreasing (e.g. w_{POLY}), for a real-valued continuous random variable X with cumulative distribution function $F(x)$ and density $f(x)$, one can think of $f(x)w'(F(x))$ as a CPT-weighted density and $w(F(x))$ is the corresponding CPT-weighted cumulative distribution function. $U_{\text{CPT-RD}}$ is the expectation of the random variable that has the CPT-weighted cumulative distribution function. We denote the set of non-decreasing functions in $\mathcal{W}_{\text{CPT}}$ to be $\overline{\mathcal{W}}_{\text{CPT}}$.

3.2.2 Empirical Risk Minimization

The canonical way of learning an ML model is through empirical risk minimization (ERM). Given n i.i.d. samples $Z_1, \ldots, Z_n \in \mathcal{Z}$, and a loss function $\ell : \Theta \times \mathcal{Z} \to \mathbb{R}$, the *population risk* (expected loss) for model θ is defined to be:

$$R(\theta) = \mathbb{E}[\ell(\theta; Z)].$$

ERM minimizes $\frac{1}{n}\sum_{i=1}^{n} \ell(\theta; Z_i)$ (empirical risk). However, expectation is only one of the many risk measures. For example, value-at-risk and conditional value-at-risk (Rockafellar, Uryasev, et al., 2000) are popular risk measures for evaluating risks of portfolios of financial instruments. CPT defines another way of measuring risk, which aligns with human's preferences. We want to study if minimizing a human-aligned risk will give us ML models that have properties other than a low population risk.

3.3 Human Risk Minimization

Definition 3.2

Given a real-valued random variable $Z \in \mathcal{Z}$, a loss function $\ell : \Theta \times \mathcal{Z} \to \mathbb{R}$ and a CPT probability weighing function $w \in \overline{\mathcal{W}}_{\text{CPT}}$, the human risk is defined to be

$$R_H(\theta; w) \stackrel{\text{def}}{=} \mathbb{E}[\ell(\theta; Z) w'(F(\ell(\theta; Z)))], \tag{3.4}$$

where $F(\ell(\theta; Z))$ is the cumulative distribution function of the loss.

Comparing Equation 3.4 with the CPT-weighted rank-dependent utility in Equation 3.2, we see that $R_H(\theta) = U_{\text{CPT-RD}}(\ell(\theta; Z))$. Given n i.i.d. samples $Z_1, \ldots, Z_n \in \mathcal{Z}$, we define empirical human risk minimization (EHRM) as

$$\theta^* = \arg\min_{\theta} \frac{1}{n} \sum_{i=1}^{n} \ell(\theta; Z_i) w'(F_n(\ell(\theta; Z_i))), \tag{3.5}$$

where F_n is the empirical CDF of the loss.

Optimization. When ℓ is differentiable, we use the following iterative update rule to minimize empirical human risk:

$$\theta^{t+1} = \theta^t - \frac{\eta_t}{n} \sum_{i=1}^{n} w_i^t \nabla_\theta \ell(\theta^t; Z_i) \quad \text{for all} \quad t \in \{0, \cdots, T-1\},$$

where $w_i^t = w'(F_n(\ell(\theta^t; Z_i)))$ and η_t is the learning rate. Note that this heuristic approach relies crucially on the assumption that minor perturbations in θ, don't change $w'(F_n(\ell(\theta; Z))$ drastically. We empirically show that such a heuristic approach performs quite well in practice (See Appendix B.2). Deriving provably optimal optimization algorithms for EHRM is an interesting open problem.

Remarks. In general, there are two levels of decision making in supervised learning: model selection when training a model, and instance prediction when using a model. These two kinds of decisions are very much related. In traditional ERM, a model is selected over others when per-instance predictions are more accurate on average. We explore the consequences of EHRM in both settings: Section 3.2.1 and 3.4.2 discuss the model selection consequences of EHRM; Section 3.5 explores its consequences on per-instance predictions. Different from traditional settings where CPT is considered, supervised learning only evaluates losses. Humans tend to be risk-averse when facing possibilities of large loss. Such a property distinguishes EHRM from ERM. When training machine learning models, surrogate losses are used (e.g., hinge loss is used in replace of $0/1$ loss). Most of the times, such surrogate losses are upper bounds for the original losses. In such cases, the risk-aversion towards possible drastic loss will be carried through when surrogate loss is used instead of the true loss.

To further understand how adding the weights $w'(F(\ell(\theta; Z)))$ to expected loss influences the learned model, we provide the psychological interpretation (Section 3.4.1) and an analytical illustration of how skewness of the loss distribution may influence choices of people with different risk preferences (Section 3.4.2) as well as an information weighting view point (Section 3.4.3) of the probability weighting function.

3.4 Characteristic Properties of Human Risk Minimization

We next review some characteristic properties of human risk minimization, contrasting it with the standard machine learning objective of expected loss minimization. To simplify the notation, we will denote the CDF $F(\ell(\theta; Z))$ of the loss random variable L as $F(\ell)$ in the subsequent analysis.

3.4.1 Diminishing Sensitivity to Probability Changes

Recall from Equation (3.3) that we work with the following polynomial form of CPT probability weighting function

$$w_{\text{POLY}}(F(x)) = \frac{3-3b}{a^2-a+1}\left(F(x)^3 - (a+1)F(x)^2 + aF(x)\right) + F(x),$$

where $a \in (0,1)$ is the fixed point of the function, and $b \in (0,1)$ controls the curvature. For any event E with probability $P(E) \in (0,1)$, given a probability change Δ, we define

$$g(P(E)) = \left(w_{\text{POLY}}(P(E)+\Delta; a, b) - w_{\text{POLY}}(P(E); a, b)\right)/\Delta,$$

which is the ratio between the human perceived probability change and the original probability change. Intuitively, $g(P(E))$ represents human's sensitivity to probability changes.

Lemma 3.1

For any event E with probability $P(E) \in (0,1)$, $\lim_{\Delta \to 0} g(P(E))$ is a monotonically increasing function of $|P(E) - \frac{a+1}{3}|$.

The above stated result can be seen as a quantitative evidence of how CPT probability weighting function captures humans' diminishing sensitivity, which has been long-studied in behavioral economics (Tversky and Kahneman, 1992). Humans are sensitive to probability changes of extreme events. Such sensitivity diminishes as the events become less extreme. When using CPT probability weighting function to weight $F(\ell)$, the event we are considering is $E = \mathbb{I}\{L \leq \ell\}$, i.e. if the loss L is less than or equal to a threshold ℓ. In this case, $P(E) = F(\ell)$. Diminishing sensitivity states that for a given amount of probability change, human's perceived probability change depends on where the probability change happens. The perceived change diminishes as the distance between where it happens and the boundary (impossibility $F(\ell) = 0$ and certainty $F(\ell) = 1$) becomes smaller. The probability changes that happen close to the boundary will be up-weighted while the changes in between will be down-weighted. As shown in Lemma 3.1, for w_{POLY}, the sensitivity of the probability change Δ diminishes as $P(E)$ moves away from 0 (impossibility) and 1 (certainty).

3.4.2 Responsiveness to Skewness of the Loss Distribution

Since the inverse S-shaped probability weighting function exaggerates small probabilities of both good and bad extreme outcomes, intuitively, its overall impact on evaluating a model depends

on higher-order moments of the loss distribution. We highlight this phenomena by considering a family of Bernoulli distributions with same mean and variance. In particular, consider the family of models $\{\mathcal{M}_\theta \mid \theta \in [0,1]\}$ whose losses $\{\ell(\theta) \mid \theta \in [0,1]\}$ are parameterized by θ. For all $\theta \in [0,1]$, suppose that $\ell(\theta)$ follows a Bernoulli distribution (He et al., 2018; Low et al., 2012):

$$P\left(\ell(\theta) = 1 - \left(\frac{1-\theta}{\theta}\right)^{1/2}\right) = \theta, P\left(\ell(\theta) = 1 + \left(\frac{\theta}{1-\theta}\right)^{1/2}\right) = 1-\theta.$$

In the above setup, the mean and variance of the losses are independent of θ. In particular, we have that

$$\mathbb{E}[\ell(\theta)] = 1 \text{ and } \text{Var}(\ell(\theta)) = 1 \text{ for all } \theta \in [0,1].$$

Hence, in this setting empirical risk minimization will treat all the models equally. However, the third central moment (skewness) of $\ell(\theta)$ is given by

$$\text{Skewness}(\ell(\theta)) = \frac{2\theta - 1}{\sqrt{\theta(1-\theta)}}.$$

Observe that $\text{Skewness}(\ell(\theta))$ is a monotonically increasing function of θ, with $\text{Skewness}(\ell(\theta)) = 0$ for $\theta = \frac{1}{2}$. Hence, $\theta < 0.5$ corresponds to models with negatively skewed loss distributions, while $\theta > 0.5$ corresponds to models with positively skewed loss distributions. Then, in this setting, we have the following result:

Lemma 3.2

Consider the human risk objective in Equation (3.4) instantiated with w_{POLY} having fixed point $a = \frac{1}{2}$. Then, we have the following:

1. For $\theta < 0.5$, $R_H(\theta; w_{\text{POLY}}(\,\cdot\,; a, b))$ is a monotonically increasing function of b.
2. For $\theta > 0.5$, $R_H(\theta; w_{\text{POLY}}(\,\cdot\,; a, b))$ is a monotonically decreasing function of b.

Remarks. The above result shows that for models with negatively skewed loss distributions, their expected loss is higher than any human risk, while the opposite is true for positively skewed loss distributions. While empirical risk minimization will treat all the models equally, human risk minimization will distinguish the models through higher-order moments of the loss distribution.

3.4.3 Weighting by Information Content

The information content or "surprisal" of an event is the amount of information gained when the event is observed and is defined as follows.

Definition 3.3

The information content of an event E with probability $P(E)$ is defined as

$$I(E) \stackrel{\text{def}}{=} -\ln[P(E)],$$

where e is used as the base of the logarithm.

Note that the above definition captures the intuition that the observation of a rare event provides more information than a common one. In our setting, the rare event corresponds to the event that the loss L takes extreme values.

Next, we construct a special weighting function $w_{\text{IT}}(\cdot)$ using the information content of the events $E_1 = \mathbb{I}\{L \leq \ell\}$ and $E_2 = \mathbb{I}\{L > \ell\}$. Observe that the information content of the events E_1 and E_2 is given by $-\ln F(\ell)$ and $-\ln(1 - F(\ell))$ respectively. Moreover, it is easy to see that as ℓ gets smaller, the information content of the left tail event E_1 increases; and as ℓ gets larger, the information content of the right tail event E_2 increases.

Then, we can use the information content of E_1 and E_2 to weight the density of L, and define the corresponding weighting function. In particular, the information weighted density is defined to be:

$$\begin{aligned} w'_{\text{IT}}(F(\ell))f(\ell) &= \frac{1}{2}(I(E_1) + I(E_2))f(\ell) \\ &= -\frac{1}{2}f(\ell)\ln\left(F(\ell)\cdot(1 - F(\ell))\right), \end{aligned}$$

and the corresponding **information content weighting function** is given by:

$$\begin{aligned} w_{\text{IT}}(F(\ell)) &= \int_0^1 1 \cdot w'_{\text{IT}}(F(\ell))dF(\ell) \\ &= \frac{1}{2}\big((1 - F(\ell))\cdot\ln(1 - F(\ell)) - F(\ell)\cdot\ln(F(\ell))\big) + F(\ell). \end{aligned}$$

Lemma 3.3

The information content weighing function w_{IT} belongs to $\overline{\mathcal{W}}_{\text{CPT}}$.

Interestingly, using w_{IT} to weight a distribution is of interest in information theory, where it is known as the two-sided information-weighted distribution (Oliveira and Cintra, 2016). Moreover, it is easy to see that $w_{\text{POLY}}(\cdot, a, b)$ with fixed point $a = 1/2$ and curvature $b = \ln 2$ is approximately equal to the third order Taylor approximation of w_{IT}. To the best of our knowledge, this is the first time that information weighting function and CPT probability weighting function have been connected. Uncertainty-aversion in human preferences has also been studied in behavioral economics (Camerer and Weber, 1992; Camerer et al., 2004). In the context of human risk minimization, using w_{IT}, we can define the information-weighted human risk to be

$$R_H(\theta; w_{\text{IT}}) = \mathbb{E}[\ell(\theta; Z)w'_{\text{IT}}(F(\ell(\theta; Z)))].$$

As studied in (Oliveira and Cintra, 2016), the information-weighted distribution $w_{\text{IT}}(F(\ell))$ will be heavy-tailed when the CDF $F(\ell) = e^{-\kappa|\ell|}$ for some $\kappa > 0$. Such heavy-tailedness for the loss distribution may cause human risk to be hard to estimate. We believe that deriving statistically and computationally optimal procedures for minimizing $R_H(\theta; w_{\text{IT}})$ is an interesting direction for future work.

3.5 Implications for Performance over Subgroups

Machine learning models are being increasingly deployed to automate a variety of day-to-day tasks. Employers use such models to select job applicants, provide credit scoring and predicting insurance premiums. With such high stakes, ensuring that learned models are non-discriminatory or *fair* with respect to sensitive features such as gender and race is of utmost importance (Pedreshi et al., 2008; Kamishima et al., 2011; Dwork et al., 2012; Kusner et al., 2017). In this section, we explore the implications of HRM towards subgroup performances. In particular, we know that HRM up-weights possible extreme events, hence, we expect HRM to avoid drastic losses for all subpopulations. We test this hypothesis on both synthetic and real-world datasets and use $w_{\text{POLY}}(\,\cdot\,; a, b)$ specified in Equation 3.3. We have chosen $a = .5$ so that for all $b \in (0, 1)$, $w'_{\text{POLY}}(F(\ell))$ is symmetric about the line $F(\ell) = a$.

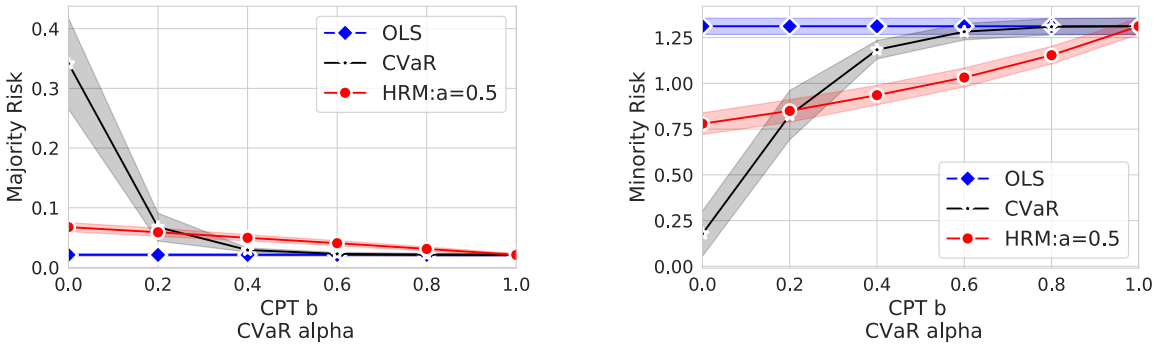
3.5.1 Synthetic Experiment

Setup. In this experiment, we create a synthetic regression task to test the performance of EHRM on the minority subgroup. We follow the setup of (Durrett, 2019) and draw our covariates (features) from an isotropic Gaussian $X \sim \mathcal{N}(0, \mathbf{I}_5)$ in $\mathbb{R}^5$. The noise distribution is fixed as $\epsilon \sim \mathcal{N}(0, .01)$. We draw our response variable Y as,

$$Y = \begin{cases} X^\top \theta^* + \epsilon & \text{if } X^{(1)} \leq 1.645 \\ X^\top \theta^* + X^{(1)} + \epsilon & \text{otherwise} \end{cases}$$

where $\theta^* = [1, 1, 1, 1, 1]$ and $X^{(1)}$ is the first coordinate of X. Observe that since $P\left(X^{(1)} > 1.645\right) = .05$, $\{X \mid X^{(1)} > 1.645\}$ represents our minority subgroup. We fix the squared error $\ell(\theta; (x, y)) = \frac{1}{2}(y - x^T\theta)^2$ as our loss function.

Results. Figure 3.2 plots the risk of minority and majority groups for EHRM and ERM. The empirical risk minimizer is denoted by OLS, the solution of this ordinary least square problem. We see that for different values of $b < 1$, EHRM has a lower minority risk than ERM. Moreover, as b approaches 1, EHRM becomes more similar to ERM. This validates our hypothesis: because the inverse S-shaped probability weighting function inflates small probabilities for extreme losses, drastic losses of the minority group will be exaggerated and human risk minimization trades a low population risk for a better minority performance. Optimization performance of EHRM is shown in Figure B.1 (Appendix B.2). In addition to comparing with ERM, we have also compared EHRM with conditional value-at-risk (CVaR), a risk measure that has been used to measure the worst-case subgroup performance (Duchi et al., n.d.; Williamson and Menon,



(a) Majority Group Performance

(b) Minority Group Performance

Figure 3.2: Majority and minority performance of ERM, EHRM and CVaR on the synthetic dataset. Note that when $b = 1$, the CPT probability weighting function is the identity function. Hence, EHRM is the same as ERM. 2000 training and 20000 testing data points are used in the experiment. Solid lines and shaded area represent the means and one standard derivations of the risks.

2019). $\text{CVaR}_\alpha(\ell(\theta;(x,y)))$ is the expectation of the worst α proportion of the losses. As shown in Figure 3.2, when α is small, CVaR has a lower minority risk than EHRM and ERM, at a cost of a higher majority risk. As α approaches 1, the minority risk increases drastically.

3.5.2 Recidivism Prediction: Similar Subpopulation Performance

Setup. We follow the experimental set up in (Duchi et al., n.d.). Using the fairML toolkit version of the COMPAS recidivism dataset (Adebayo et al., 2016), we want to study the performance of EHRM and ERM on different demographic subgroups. With a 90% and 10% train-test split, ERM and EHRM are used to train a logistic regression model with L_2-regularization. To study the subgroup performances, we report the misclassification rate of different demographic groups on the test set. In particular, out of the 10 binary features in the dataset, we have chosen 7 of them that have more than 10 samples to group the population. For each chosen (binary) attribute, the dataset can be divided into two subgroups. For EHRM, we have chosen b to be .3 so that the EHRM probability weighting function is close to the median estimate of the CPT probability weighting function for high rank losses in (Tversky and Kahneman, 1992). In practice, a and b are application-dependent and user-dependent.

Results. In Figure 3.3, for each attribute, we report the maximum misclassification rate of the two subgroups at test time. Compared to ERM, EHRM has a higher misclassification rate but a more similar worst case performance across different subgroups. Such an observation aligns with our hypothesis that EHRM avoids extremely bad performances for all demographic groups and hence will sacrifice average performance for similar subpopulation performances.

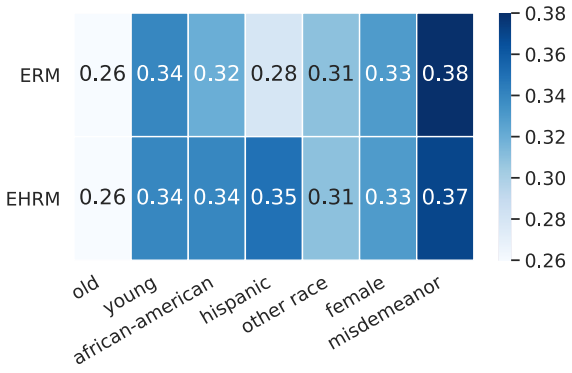


Figure 3.3: Test misclassification rate of the recidivism prediction task. Each box represents the worst performance among the subpopulations grouped by the attribute listed below. EHRM has a more similar performance across subgroups.

3.5.3 Gender Classification based on Facial Image: Fairness Metrics Comparison

Setup. To study EHRM performance on standard fairness metrics, we use the AI Fairness 360 toolkit (Bellamy et al., 2018). In particular, we use the UTKFace dataset (Zhang et al., 2017) to train a neural network[1] for predicting gender based on facial images (male$=0$, female$=1$). As suggested by (Bellamy et al., 2018), we use race as an indicator to divide the population into two groups G_1 (white) and G_2 (other race). The fairness metrics we have used include statistical parity difference, disparate impact, equal opportunity difference, average odds difference, Theil index and false negative rate difference.[2] We train the model with ERM and EHRM over 10 random seeds. For EHRM, b is chosen to be .3 for the same reason mentioned in Section 3.5.2. To minimize empirical human risk, we have used a variant of mini-batch stochastic gradient descent. At each step t, $\theta^{t+1} = \theta^t - \eta_t \sum_{i=1}^{B} w_i^t \nabla_\theta \ell(\theta; Z_i)/B$, where $w_i^t = w'_{\text{POLY}}(F_n(\ell(\theta^t; Z_i)))$, $F_n(\cdot)$ is the empirical CDF of the mini-batch losses, B is the mini-batch size and η_t is the learning rate. As shown in Figure B.2 (Appendix B.2), the empirical human risk of the entire training dataset decreases as training proceeds. We have also compared EHRM with a data pre-processing algorithm named reweighing (Kamiran and Calders, 2012) that re-weights the samples so that statistical dependence between the protected attribute and label are mitigated.

Results. Table 3.1 shows the mean and standard deviation of the test time performance of EHRM, and ERM with and without reweighing pre-processing. ERM performs the best in terms of accuracy at test time. However, reweighing with ERM and EHRM does better in terms of the fairness metrics. The empirical result suggests that the human innate risk-aversion towards possibility of extreme losses has promoted similar performances across different subgroups.

[1] Appendix B.4 consists details of the model configuration.
[2] Appendix B.3 contains definitions of these metrics in terms of G_1 and G_2.

Table 3.1: Mean and standard deviation of accuracy and fairness metrics of models learned by EHRM with $w_{\text{POLY}}(\,\cdot\,; .5, .3)$, and ERM with and without reweighing pre-processing. For each metric, the best performing algorithm is highlighted.

	EHRM(.5, .3)	Reweighing (Kamiran and Calders, 2012)	ERM
Accuracy	.8751 ±.0052	.8767 ±.0067	**.8767 ±.0060**
Stat. Parity Diff.	**-.0825 ±.0220**	-.0875 ±.0212	-.0881 ±.0208
Disparate Impact	**.8475 ±.0411**	.8396 ±.0390	.8368±.0386
Equal Opp. Diff.	**-.0440 ±.0261**	-.0518 ±.0253	-.0502±.0263
Avg. Odds Diff.	**-.0116 ±.0202**	-.0165 ±.0177	-.0173 ±.0188
Theil Index	.0859 ±.0058	**.0824 ±.0038**	.0855 ±.0071
FNR Diff.	**.0440 ±.0261**	.0518 ±.0253	.0502 ±.0263

3.6 Discussion

In this work, we have studied alternatives to empirical risk minimization, and in particular proposed alternate formulations, which are better aligned with human risk measures. We have analyzed several characteristics of human risk minimization such as diminishing sensitivity, model selection based on higher-order moments and information-weighted loss distributions. Further, our empirical analysis has shown that such risk measures have implications for fairness, and in particular trade average performance for similar subgroup performances. Our empirical analysis raises several interesting future directions. Fairness is only one of such desiderata that people start caring about in ML. We would like to study other desiderata that HRM brings. Meanwhile, many risk measures such as conditional value-at-risk (Rockafellar, Uryasev, et al., 2000) can be expressed in a dual form, however, it is not immediately clear if HRM has an equivalent formulation.

Acknowledgement We thank the reviewers for providing thoughtful and constructive feedback for the paper. We thank Hongseok Namkoong for providing the method to optimize conditional value-at-risk. We acknowledge the support of ONR via N000141812861.

Part II

Integrating human characteristics into machine learning algorithm design

Chapter 4

Rebounding Bandits for Modeling Satiation Effects

4.1 Introduction

Recommender systems suggest such diverse items as music, news, restaurants, and even job candidates. Practitioners hope that by leveraging historical interactions, they might provide services better aligned with their users' preferences. However, despite their ubiquity in application, the dominant learning framework suffers several conceptual gaps that can result in misalignment between machine behavior and human preferences. For example, because human preferences are seldom directly observed, these systems are typically trained on the available observational data (e.g., purchases, ratings, or clicks) with the objective of predicting customer behavior (Bennett, Lanning, et al., 2007; McAuley and Leskovec, 2013). Problematically, such observations tend to be confounded (reflecting exposure bias due to the current recommender system) and subject to censoring (e.g., users with strong opinions are more likely to write reviews) (Swaminathan and Joachims, 2015; Joachims et al., 2017).

Even if we could directly observe the utility experienced by each user, we might expect it to depend, in part, on the history of past items consumed. For example, consider the task of automated (music) playlisting. As a user is made to listen to the same song over and over again, we might expect that the utility derived from each consecutive listen would decline (Ratner et al., 1999). However, after listening to other music for some time, we might expect the utility associated with that song to bounce back towards its baseline level. Similarly, a diner served pizza for lunch might feel diminished pleasure upon eating pizza again for dinner.

The psychology literature on *satiation* formalizes the idea that enjoyment depends not only on one's intrinsic preference for a given product but also on the sequence of previous exposures and the time between them (Baucells and Sarin, 2007; Caro and Martıénez-de-Albéniz, 2012). Research on satiation dates to the 1960s (if not earlier) with early studies addressing brand loyalty (Tucker, 1964; McConnell, 1968). Interestingly, even after controlling for marketing variables like price, product design, promotion, etc., researchers still observe brand-switching

behavior in consumers. Such behavior, referred as *variety seeking*, has often been explained as a consequence of utility associated with the change itself (McAlister, 1982; Kahn, 1995). For a comprehensive review on hedonic decline caused by repeated exposure to a stimulus, we refer the readers to (Galak and Redden, 2018).

In this paper, we introduce *rebounding bandits*, a multi-armed bandits (MABs) (Robbins, 1952) framework that models satiation via linear dynamical systems. While traditional MABs draw rewards from *fixed* but unknown distributions, rebounding bandits allow each arm's rewards to evolve as a function of both the per-arm characteristics (susceptibility to satiation and speed of rebounding) and the historical pulls (e.g., past recommendations). In rebounding bandits, even if the dynamics are known and deterministic, selecting the optimal sequence of T arms to play requires planning in a Markov decision process (MDP) whose state space scales exponentially in the horizon T. When the satiation dynamics are known and stochastic, the states are only partially observable, since the satiation of each arm evolves with (unobserved) stochastic noises between pulls. And when the satiation dynamics are unknown, learning requires that we identify a stochastic dynamical system.

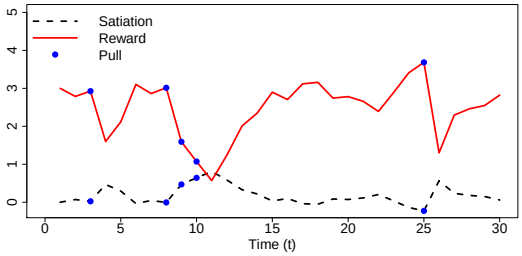
We propose Explore-Estimate-Plan (EEP) an algorithm that (i) collects data by pulling each arm repeatedly, (ii) estimates the dynamics using this dataset; and (iii) plans using the estimated parameters. We provide guarantees for our estimators in Section 4.6.2 and bound EEP's regret in Section 4.6.3.

Our main contributions are: (i) the rebounding bandits problem (Section 4.3), (ii) analysis showing that when arms share rewards and (deterministic) dynamics, the optimal policy pulls arms cyclically, exhibiting variety-seeking behavior (Section 4.4.1); (iii) an estimator (for learning the satiation dynamics) along with a sample complexity bound for identifying an affine dynamical system using a single trajectory of data (Section 4.6.2); (iv) EEP, an algorithm for learning with unknown stochastic dynamics that achieves sublinear w-step lookahead regret (Pike-Burke and Grunewalder, 2019) (Section 4.6); and (v) experiments demonstrating EEP's efficacy (Section 4.7).

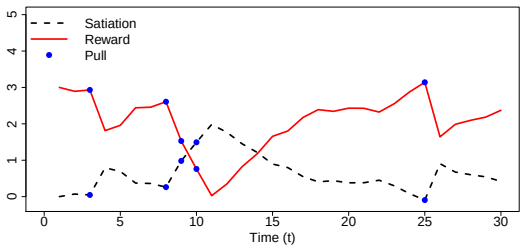
4.2 Related Work

Satiation effects have been addressed by such diverse disciplines as psychology, marketing, operations research, and recommendation systems. In the psychology and marketing literatures, satiation has been proposed as an explanation for variety-seeking consumer behavior (Galak and Redden, 2018; McAlister, 1982; McAlister and Pessemier, 1982). In operations research, addressing continuous consumption decisions, (Baucells and Sarin, 2007) propose a deterministic linear dynamical system to model satiation effects. In the recommendation systems community, researchers have used semi-Markov models to explicitly model two states: (i) *sensitization*—where the user is highly interested in the product; and (ii) *boredom*—where the user is not engaged (Kapoor et al., 2015).

The bandits literature has proposed a variety of extensions where rewards depend on past exposures, both to address satiation and other phenomena. (Heidari et al., 2016; Levine et al., 2017; Seznec et al., 2019) tackle settings where each arm's expected reward grows (or shrinks) monotonically in the number of pulls. By contrast, (Kleinberg and Immorlica, 2018; Basu et al.,



(a) $\gamma_k = .5, \lambda_k = 3, b_k = 3$



(b) $\gamma_k = .8, \lambda_k = 1.5, b_k = 3$

Figure 4.1: These plots illustrate the satiation level and reward of an arm from time 1 to 30. The two plots are generated with the same pull sequence, base rewards $b_k = 3$ and realized noises with variance $\sigma_z = .1$. In Figure 4.1a, $\gamma_k = .5$ and $\lambda_k = 3$. In Figure 4.1b, $\gamma_k = .8$ and $\lambda_k = 1.5$. In both cases, the arm has started with 0 as its base satiation level. **Black dashed line:** the satiation level. **Red solid line:** the reward. **Blue dots:** time steps where the arm is pulled.

2019; Cella and Cesa-Bianchi, 2020) propose models where rewards increase as a function of the time elapsed since the last pull. (Pike-Burke and Grunewalder, 2019) model the expected reward as a function of the time since the last pull drawn from a Gaussian Process with known kernel. (Warlop et al., 2018) propose a model where rewards are linear functions of the recent history of actions and (Mintz et al., 2020) model the reward as a function of a context that evolves according to known deterministic dynamics. In rested bandits (Gittins, 1979), an arm's expected rewards change only when it is played, and in restless bandits (Whittle, 1988) rewards evolve independently from the play of each arm.

Key Differences This may be the first bandits paper to model evolving rewards through continuous-state linear stochastic dynamical systems with unknown parameters. Our framework captures several important aspects of satiation: rewards decline by diminishing amounts with consecutive pulls and rebound towards the baseline with disuse. Unlike models that depend only on fixed windows or the time since the last pull, our model expresses satiation more organically as a quantity that evolves according to stochastic dynamics. To estimate the reward dynamics, we leverage recent advances in the identification of linear dynamical systems (Simchowitz et al., 2018; Sarkar and Rakhlin, 2019) that rely on the theory of self-normalized processes (Peña et al., 2009; Abbasi-Yadkori et al., 2011) and block martingale conditions (Simchowitz et al., 2018).

4.3 Rebounding Bandits Problem Setup

Consider the set of K arms $[K] := \{1, \ldots, K\}$ with bounded base rewards $b_1, \ldots, b_K$. Given a horizon T, a policy $\pi_{1:T} := (\pi_1, \ldots, \pi_T)$ is a sequence of actions, where $\pi_t \in [K]$ depends on past actions and observed rewards. For any arm $k \in [K]$, we denote its pull history from time 0 to T as the binary sequence $u_{k,0:T} := (u_{k,0}, \ldots, u_{k,T})$, where $u_{k,0} = 0$ and for $t \in [T]$, $u_{k,t} = 1$ if $\pi_t = k$ and $u_{k,t} = 0$ otherwise. The subsequence of $u_{k,0:T}$ from t_1 to t_2 (including both endpoints) is denoted by $u_{k,t_1:t_2}$.

At time t, each arm k has a satiation level $s_{k,t}$ that depends on a *satiation retention* factor $\gamma_k \in [0, 1)$, as follows

$$s_{k,t} := \gamma_k(s_{k,t-1} + u_{k,t-1}) + z_{k,t-1},\ \forall t > t_0^k, \tag{4.1}$$

where $t_0^k := \min_t\{t :\ u_{k,t} = 1\}$ is the first time arm k is pulled and $z_{k,t-1}$ is independent and identically distributed noise drawn from $\mathcal{N}(0, \sigma_z^2)$, accounting for incidental (uncorrelated) factors in the satiation dynamics. Because satiation requires exposure, arms only begin to have nonzero satiation levels after their first pull, i.e., $s_{k,0} = \ldots = s_{k,t_0^k} = 0$.

At time $t \in [T]$, if arm k is played with a current satiation level $s_{k,t}$, the agent receives reward $\mu_{k,t} := b_k - \lambda_k s_{k,t}$, where b_k is the base reward for arm k and $\lambda_k \geq 0$ is a bounded *exposure influence* factor. We use *satiation influence* to denote the product of the exposure influence factor λ_k and the satiation level $s_{k,t}$. In Figure 4.1, we show how rewards evolve in response to both pulls and the stochastic dynamics under two sets of parameters. The expected reward of arm k (where the expectation is taken over all noises associated with the arm) monotonically decreases by diminishing amounts with consecutive pulls and increases with disuse by diminishing amounts.

Remark 4.1. *We note that there exist choices of b_k, γ_k, λ_k for which the expected reward of arm k can be negative. In the traditional bandits setup, one must pull an arm at every time step. Thus, what matters are the relative rewards and the problem is mathematically identical, regardless of whether the expected rewards range from -10 to 0 or 0 to 10. In addition, one might construct settings where negative expected rewards are reasonable. For example, when one of the arms corresponds to no recommendation with 0 being its expected reward (e.g., $b_k = 0$, $\lambda_k = 0$), then the interpretation of negative expected reward would be that pulling (recommending) the corresponding arm (item) is less preferred relative to not pulling (no recommendation).*

Given horizon $T \geq 1$, we seek a pull sequence $\pi_{1:T}$, where π_t depends on past rewards and actions $(\pi_1, \mu_{\pi_1,1}, \ldots, \pi_{t-1}, \mu_{\pi_{t-1},t-1})$, that maximizes the expected cumulative reward:

$$G_T(\pi_{1:T}) := \mathbb{E}\left[\sum_{t=1}^{T} \mu_{\pi_t,t}\right]. \tag{4.2}$$

Additional Notation Let $\overline{\gamma} := \max_{k\in[K]} \gamma_k$ and $\overline{\lambda} := \max_{k\in[K]} \lambda_k$. We use $a \lesssim b$ when $a \leq Cb$ for some positive constant C.

4.4 Planning with Known Dynamics

Before we can hope to learn an optimal policy with unknown stochastic dynamics, we need to establish a procedure for planning when the satiation retention factors, exposure influence factors, and base rewards are known. We begin by presenting several planning strategies and analyzing them under deterministic dynamics, where the past pulls exactly determine each arm's satiation level, i.e., $s_{k,t} = \gamma_k(s_{k,t-1} + u_{k,t-1})$, $\forall t > t_0^k$. With some abuse of notation, at time $t \geq 2$, given a pull sequence $u_{k,0:t-1}$, we can express the satiation and the expected[1] reward

[1]We use "expected reward" to emphasize that all results in this section apply to settings where the satiation dynamics are deterministic but the rewards are stochastic, i.e., $\mu_{k,t} = b_k - \lambda_k s_{k,t} + e_{k,t}$ for independent mean-zero noises $e_{k,t}$.

of each arm as

$$s_{k,t}(u_{k,0:t-1}) = \gamma_k \left(s_{k,t-1} + u_{k,t-1}\right) = \gamma_k \left(\gamma_k \left(s_{k,t-2} + u_{k,t-2}\right)\right) + \gamma_k u_{k,t-1} = \textstyle\sum_{i=1}^{t-1} \gamma_k^{t-i} u_{k,i},$$
$$\mu_{k,t}(u_{k,0:t-1}) = b_k - \lambda_k \left(\textstyle\sum_{i=1}^{t-1} \gamma_k^{t-i} u_{k,i}\right). \tag{4.3}$$

At time $t = 1$, we have that $s_{k,1}(u_{k,0:0}) = 0$ and $\mu_{k,1}(u_{k,0:0}) = b_k$ for all $k \in [K]$. Since the arm parameters $\{\lambda_k, \gamma_k, b_k\}_{k=1}^K$ are known, our goal (4.2) simplifies to finding a pull sequence that solves the following bilinear integer program:

$$\max_{u_{k,t}} \left\{ \sum_{k=1}^{K} \sum_{t=1}^{T} u_{k,t} \left(b_k - \lambda_k \sum_{i=0}^{t-1} \gamma_k^{t-i} u_{k,i} \right) : \begin{array}{l} \sum_{k=1}^{K} u_{k,t} = 1, \quad \forall t \in [T], \\ u_{k,t} \in \{0,1\}, \;\; u_{k,0} = 0, \; \forall k \in [K], \forall t \in [T] \end{array} \right\} \tag{4.4}$$

where the objective maximizes the expected cumulative reward associated with the pull sequence and the constraints ensure that at each time period we pull exactly one arm. Note that (4.4) includes products of decision variables $u_{k,t}$ leading to bilinear terms in the objective. In Appendix C.1, we provide an equivalent integer linear program.

4.4.1 The Greedy Policy

At each step, the greedy policy π^g picks the arm with the highest instantaneous expected reward. Formally, at time t, given the pull history $\{u_{k,0:t-1}\}_{k=1}^K$, the greedy policy picks

$$\pi_t^g \in \arg\max_{k \in [K]} \mu_{k,t}(u_{k,0:t-1}).$$

In order to break ties, when all arms have the same expected reward, the greedy policy chooses the arm with the lowest index.

Note that the greedy policy is not, in general, optimal. Sometimes, we are better off allowing the current best arm to rebound even further, before pulling it again.

Example 4.1. *Consider the case with two arms. Suppose that arm 1 has base reward b_1, satiation retention factor $\gamma_1 \in (0,1)$, and exposure influence factor $\lambda_1 = 1$. For any fixed time horizon $T > 2$, suppose that arm 2 has $b_2 = b_1 + \frac{\gamma_2 - \gamma_2^T}{1-\gamma_2}$ where $\gamma_2 \in (0,1)$ and $\lambda_2 = 1$. The greedy policy $\pi_{1:T}^g$ will keep pulling arm 2 until time $T-1$ and then play arm 1 (or arm 2) at time T. This is true because if we keep pulling arm 2 until $T-1$, at time T, we have $\mu_{2,T}(u_{2,0:T-1}) = b_1 = \mu_{1,T}(u_{1,0:T-1})$. However, the policy $\pi_{1:T}^n$, where $\pi_t^n = 2$ if $t \leq T-2$, $\pi_{T-1}^n = 1$, and $\pi_T^n = 2$, obtains a higher expected cumulative reward. In particular, the difference $G_T(\pi_{1:T}^n) - G_T(\pi_{1:T}^g)$ will be $\gamma_2 - \gamma_2^{T-1}$.*

4.4.2 When is Greedy Optimal?

When the satiation effect is always 0, e.g., when the satiation retention factors $\gamma_k = 0$ for all $k \in [K]$, we know that the greedy policy (which always plays the arm with the highest instantaneous expected reward) is optimal. However, when satiation can be nonzero, it is less

clear under what conditions the greedy policy performs optimally. This question is of special interest when we consider human decision-making, since we cannot expect people to solve large-scale bilinear integer programs every time they pick music to listen to.

In this section, we show that when all arms share the same properties (γ_k, λ_k, b_k are identical for $k \in [K]$), the greedy policy is optimal. In this case, the greedy policy exhibits variety-seeking behavior as it plays the arms cyclically. Interestingly, this condition aligns with early research that has motivated studies on satiation (Tucker, 1964; McConnell, 1968): when controlling for marketing variables (e.g., the arm parameters γ_k, λ_k, b_k), researchers still observe variety-seeking behaviors of consumers (e.g., playing arms in a cyclic order).

Assumption 4.1. $\gamma_1 = \ldots = \gamma_K = \gamma,\ \lambda_1 = \ldots = \lambda_K = \lambda$, *and* $b_1 = \ldots = b_K = b$.

We start with characterizing the greedy policy when Assumption 4.1 holds.

Lemma 4.1: Greedy Policy Characterization

Under Assumption 4.1 and the tie-breaking rule that when all arms have the same expected reward, the greedy policy chooses the one with the lowest arm index, the sequence of arms pulled by the greedy policy forms a periodic sequence: $\pi_1 = 1, \pi_2 = 2, \cdots, \pi_K = K$, and $\pi_{t+K} = \pi_t,\ \forall t \in \mathbb{N}_+$.

In this case, the greedy policy is equivalent to playing the arms in a cyclic order. All proofs for the paper are deferred to the Appendices.

Theorem 4.1

Under Assumption 4.1, given any horizon T, the greedy policy $\pi^g_{1:T}$ is optimal.

Proof Sketch. First, when $T \leq K$, greedy policy is optimal since its cumulative expected reward is Tb. So, we consider the case of $T > K$. Assume for contradiction that there exists another policy $\pi^o_{1:T}$ that is optimal and is not greedy, i.e., $\exists t \in [T], \pi^o_t \notin \arg\max_{k\in[K]} b - \lambda s^o_{k,t}$ where $s^o_{k,t}$ denotes the satiation level of arm k at time t under the policy $\pi^o_{1:T}$. We will construct a new policy $\pi^n_{1:T}$ that obtains a higher cumulative expected reward than $\pi^o_{1:T}$. Throughout the proof, we use $s^n_{k,t}$ to denote the satiation levels for the new policy.

We first note two illustrative facts to give the intuition of the proof.

Fact 1: Any policy $\pi^o_{1:T}$ that does not pick the arm with the lowest satiation level (i.e., highest expected reward) at the last time step T is not optimal.
Proof of Fact 1: In this case, the policy $\pi^n_{1:T} = (\pi^o_1, \ldots, \pi^o_{T-1}, \pi_T)$ where $\pi_T \in \arg\max_{k\in[K]} b - \lambda s^o_{k,T}$ will obtain a higher cumulative expected reward.

Fact 2: If a policy $\pi^o_{1:T}$ picks the lowest satiation level for the final pull π^o_T but does not pick the arm with the lowest satiation level at time $T-1$, we claim that $\pi^n_{1:T} = (\pi^o_1, \ldots, \pi^o_{T-2}, \pi^o_T, \pi^o_{T-1}) \neq \pi^o_{1:T}$ obtains a higher cumulative expected reward.
Proof of Fact 2: First, note that $\pi^o_{T-1} \neq \pi^o_T$ because otherwise π^o_{T-1} is the arm with the lowest satiation level at $T-1$. Moreover, at time $T-1$, $\pi^o_T \in \arg\min_k s^o_{k,T-1}$ has the smallest satiation, since if not, then there exists another arm $k \neq \pi^o_T$ and $k \neq \pi^o_{T-1}$ that has a smaller satiation

level than π^o_T at time $T-1$. In that case, π^o_T will not be the arm with the lowest satiation at time T, which is a contradiction. Then, we deduce $s^o_{\pi^o_{T-1},T-1} > s^o_{\pi^o_T,T-1}$. Combining this with $\pi^o_{T-1} \neq \pi^o_T$, we arrive at

$$G_T(\pi^n_{1:T}) - G_T(\pi^o_{1:T}) = \lambda(1-\gamma)\left(s^o_{\pi^o_{T-1},T-1} - s^o_{\pi^o_T,T-1}\right) > 0.$$

For the general case, given any policy $\pi^o_{1:T}$ that is not a greedy policy, we construct the new policy $\pi^n_{1:T}$ that has a higher cumulative expected reward through the following procedure:

1. Find $t^* \in [T]$ such that for all $t > t^*$, $\pi^o_t \in \arg\max_{k\in[K]} b - \lambda s^o_{k,t}$ and $\pi^o_{t^*} \notin \arg\max_{k\in[K]} b - \lambda s^o_{k,t^*}$. Further, we know that $\pi^o_{t^*+1} \in \arg\max_{k\in[K]} b - \lambda s^o_{k,t^*}$, using the same reasoning as the above example, i.e., otherwise $\pi^o_{t^*+1} \notin \arg\max_{k\in[K]} b - \lambda s^o_{k,t^*+1}$. To ease the notation, we use k_1 to denote $\pi^o_{t^*}$ and k_2 to denote $\pi^o_{t^*+1}$.
2. For the new policy, we choose $\pi^n_{1:t^*+1} = (\pi^o_1, \ldots, \pi^o_{t^*-1}, k_2, k_1)$. Let $A^o_{t_1,t_2}$ denote the set $\{t' : t^*+2 \leq t' \leq t_2, \pi^o_{t'} = \pi^o_{t_1}\}$. $A^o_{t_1,t_2}$ contains a set of time indices in between t^*+2 and t_2 when arm $\pi^o_{t_1}$ is played under policy $\pi^o_{1:T}$. We construct the following three sets $T_A := \{t : t^*+2 \leq t \leq T, |A^o_{t^*,t}| < |A^o_{t^*+1,t}|\}$, $T_B := \{t : t^*+2 \leq t \leq T, |A^o_{t^*,t}| > |A^o_{t^*+1,t}|\}$ and $T_C := \{t : t^*+2 \leq t \leq T, |A^o_{t^*,t}| = |A^o_{t^*+1,t}|\}$. For time $t \geq t^*+2$, we consider the following three cases:

Case I. $T_B = \varnothing$, which means that at any time t in between t^*+2 and T, arm k_1 is played more than arm k_2 from t^*+2 to t. In this case, the new policy follows $\pi^n_{t^*+2:T} = \pi^o_{t^*+2:T}$.

Case II. $T_A = \varnothing$, which means that at any time t in between t^*+2 and T, arm k_2 is played more than arm k_1 from t^*+2 to t. In this case, the new policy satisfies: for all $t \geq t^*+2$, 1) $\pi^n_t = \pi^o_t$ if $\pi^o_t \neq k_1$ and $\pi^o_t \neq k_2$; 2) $\pi^n_t = k_2$ if $\pi^o_t = k_1$; and 3) $\pi^n_t = k_1$ if $\pi^o_t = k_2$.

Case III. $T_A \neq \varnothing$ and $T_B \neq \varnothing$. Then, starting from t^*+2, if $t \in T_A$, π^n_t follows the new policy construction in Case I, i.e., $\pi^n_t = \pi^o_t$. If $t \in T_B$, π^n_t follows the new policy construction in Case II. Finally, for all $t \in T_C$, define $t'_{A,t} = \max_{\substack{t'\in T_A:\\ t'<t}} t'$ and $t'_{B,t} = \max_{\substack{t'\in T_B:\\ t'<t}} t'$. If $t'_{A,t} > t'_{B,t}$, then π^n_t follows the new policy construction as Case I. If $t'_{A,t} < t'_{B,t}$, π^n_t follows the new policy construction as Case II. We note that $t'_{A,t} \neq t'_{B,t}$ since $T_A \cap T_B = \varnothing$.

When $T_A = \varnothing$ and $T_B = \varnothing$, we know that k_1 and k_2 are not played in $\pi^o_{t^*+2:T}$. In this case, the new policy construction can follow either Case I or Case II. The proof proceeds by analyzing the difference between $G_T(\pi^n_{1:T})$ and $G_T(\pi^o_{1:T})$. We provide the complete proof in Appendix C.2.2. □

Remark 4.2. *Theorem 4.1 suggests that when the (deterministic) satiation dynamics and base rewards are identical across arms, planning does not require knowledge of those parameters, since playing the arms in a cyclic order is optimal.*

Lemma 4.1 and Theorem 4.1 lead us to conclude the following result: when recommending items that share the same properties, the best strategy is to show the users a variety of recommendations by following the greedy policy.

On a related note, Theorem 4.1 also gives an exact Max K-Cut of a complete graph $\mathcal{K}_T$ on T vertices, where the edge weight connecting vertices i and j is given by $e(i,j) = \lambda\gamma^{|j-i|}$ for $i \neq j$. The Max K-Cut problem partitions the vertices of a graph into K subsets $P_1, \dots P_K$, such that the sum of the edge weights connecting the subsets are maximized (Frieze and Jerrum, 1997). Mapping the Max K-Cut problem back to our original setup, each vertex represents a time step. If vertex i is assigned to subset P_k, it suggests that arm k should be played at time i. The edge weights $e(i,j) = \lambda\gamma^{|j-i|}$ for $i \neq j$ can be seen as the reduction in satiation influence achieved by not playing the same arm at both time i and time j. The goal (4.4) is to maximize the total satiation influence reduction.

Proposition 4.1: Connection to Max K-Cut

Under Assumption 4.1, an optimal solution to (4.4) is given by a Max K-Cut on $\mathcal{K}_T$, where $\mathcal{K}_T$ is a complete graph on T vertices with edge weights $e(i,j) = \lambda\gamma^{|j-i|}$ for all $i \neq j$.

Using Lemma 4.1 and Theorem 4.1, we obtain an exact Max K-Cut of $\mathcal{K}_T$: $\forall k \in [K], P_k = \{t \in [T] : t \equiv k \pmod K\}$.

4.4.3 The w-lookahead Policy

To model settings where the arms correspond to items with different characteristics (e.g., we can enjoy tacos on consecutive days but require time to recover from a trip to the steakhouse) we must allow the satiation parameters to vary across arms. Here, the greedy policy may not be optimal. Thus, we consider more general lookahead policies (the greedy policy is a special case). Given a window of size w and the current satiation levels, the w-lookahead policy picks actions to maximize the total reward over the next w time steps. Let l denote $\lceil T/w \rceil$. Define $t_i = \min\{iw, T\}$ for $i \in [l]$ and $t_0 = 0$. More formally, the w-lookahead policy $\pi^w_{1:T}$ is defined as follows: for any $i \in [l]$, given the previously chosen arms' corresponding pull histories $\{u^w_{k,0:t_{i-1}}\}_{k=1}^K$ where $u^w_{k,0} = 0$ and $u^w_{k,t} = 1$ if (and only if) $\pi^w_t = k$, the next w (or T mod w) actions $\pi^w_{t_{i-1}+1:t_i}$ are given by

$$\max_{\pi^w_{t_{i-1}+1:t_i}} \left\{ \sum_{t=t_{i-1}+1}^{t_i} \mu_{\pi^w_t,t}(u_{\pi_t,0:t-1}) : \begin{array}{ll} u_{k,0:t_{i-1}} = u^w_{k,0:t_{i-1}}, & \forall k \in [K], \\ \sum_{k=1}^K u_{k,t} = 1, & \forall t \in [t_i], \\ u_{k,t} \in \{0,1\}, & \forall k \in [K], t \in [t_i] \end{array} \right\} \tag{4.5}$$

In the case of a tie, one can pick any of the sequences that maximize (4.5). We recover the greedy policy when the window size $w = 1$, and finding the w-lookahead policy for the window size $w = T$ is equivalent to solving (4.4).

Remark 4.3. *Another reasonable lookahead policy, which requires planning ahead at every time step, would be the following: at every time t, plan for the next w actions and follow them for a single time step. To lighten the computational load, we adopt the current w-lookahead policy which only requires planning every w time steps.*

For the rest of the paper, we use $\texttt{Lookahead}(\{\lambda_k, \gamma_k, b_k\}_{k=1}^K, \{u^w_{k,0:t_{i-1}}\}_{k=1}^K, t_{i-1}, t_i)$ to refer to the solution of (4.5), when the arm parameters are $\{\lambda_k, \gamma_k, b_k\}_{k=1}^K$, the historical pull sequences of all arms till time t_{i-1} are given by $\{u^w_{k,0:t_{i-1}}\}_{k=1}^K$. The solution corresponds to the actions that should be taken by the w-lookahead policy for the next $t_i - t_{i-1}$ time steps.

Theorem 4.2

Given any horizon T, let $\pi^*_{1:T}$ be a solution to (4.4). For a fixed window size $w \leq T$, we have that

$$G_T(\pi^*_{1:T}) - G_T(\pi^w_{1:T}) \leq \frac{\overline{\lambda}\overline{\gamma}(1-\overline{\gamma}^{T-w})}{(1-\overline{\gamma})^2}\lceil T/w \rceil.$$

Remark 4.4. *Note that when $w = T$, the w-lookahead policy by definition is the optimal policy and in such a case, the upper bound for the optimality gap of w-lookahead established in Theorem 4.2 is also 0. In contrast to the optimal policy, the computational benefit of the w-lookahead policy becomes apparent when the horizon T is large since it requires solving for a much smaller program (4.5). In general, the w-lookahead policy is expected to perform much better than the greedy policy (which corresponds to the case of $w = 1$) at the expense of a higher computational cost. Finally, we note that for the window size of $w = \sqrt{T}$, we obtain $G_T(\pi^*_{1:T}) - G_T(\pi^w_{1:T}) \leq O(\sqrt{T})$.*

4.5 Learning with Unknown Dynamics: Preliminaries

When the satiation dynamics are unknown and stochastic ($\sigma_z > 0$), the learner faces a continuous-state partially observable MDP because the satiation levels are not observable. To set the stage, we first introduce our state representation (Section 4.5.1) and a regret-based performance measure (Section 4.5.2). In the next section, we will introduce EEP, our algorithm for rebounding bandits.

4.5.1 State Representation

Following (Ortner et al., 2012), at any time $t \in [T]$, we define a state vector x_t in the state space $\mathcal{X}$ to be $x_t = (x_{1,t}, n_{1,t}, x_{2,t}, n_{2,t}, \ldots, x_{K,t}, n_{K,t})$, where $n_{k,t} \in \mathbb{N}$ is the number of steps at time t since arm k was last selected and $x_{k,t}$ is the satiation influence (product of λ_k and the satiation level) as of the most recent pull of arm k. Since the most recent pull happens at $t - n_{k,t}$, we have $x_{k,t} = b_k - \mu_{k,t-n_{k,t}} = \lambda_k s_{k,t-n_{k,t}}$. Recall that $\mu_{k,t-n_{k,t}}$ is the reward collected by pulling arm k at time $t - n_{k,t}$. Note that b_k is directly observed when arm k is pulled for the first time because there is no satiation effect. The state at the first time step is $x_1 = (0, \ldots, 0)$. At time t, if arm k is chosen at state x_t, and reward $\mu_{k,t}$ is obtained, then the next state x_{t+1} will satisfy (i) for the pulled arm k, $n_{k,t+1} = 1$ and $x_{k,t+1} = b_k - \mu_{k,t}$; (ii) for other arms $k' \neq k$, $n_{k',t+1} = n_{k',t} + 1$ if $n_{k',t} \neq 0$, $n_{k',t+1} = 0$ if $n_{k',t} = 0$, and the satiation influence remains the same $x_{k',t+1} = x_{k',t}$.

Given $\{\gamma_k, \lambda_k, b_k\}_{k=1}^K$, the reward function $r : \mathcal{X} \times [K] \to \mathbb{R}$ represents the *expected* reward of pulling arm k under state x_t:

If $n_{k,t} = 0$, then $r(x_t, k) = b_k$. If $n_{k,t} \geq 1$, $r(x_t, k) = \mathbb{E}[\mu_{k,t}|x_t] = b_k - \gamma_k^{n_{k,t}} x_{k,t} - \lambda_k \gamma_k^{n_{k,t}}$, where the expectation is taken over the noises in between the current pull and the last pull of

arm k. See Appendix C.3.1 for the full description of the MDP setup (including the transition kernel and value function definition) of rebounding bandits.

4.5.2 Evaluation Criteria: w-step Lookahead Regret

In reinforcement learning (RL), the performance of a learner is often measured through a regret that compares the expected cumulative reward obtained by the learner against that of an optimal policy in a competitor class (Lattimore and Szepesvári, 2020). In most episodic (e.g., finite horizon) RL literature (Ortner and Ryabko, 2012; Jaksch et al., 2010), regrets are defined in terms of episodes. In such cases, the initial state is reset (e.g., to a fixed state) after each episode ends, independent of previous actions taken by the leaner. Unlike these episodic RL setups, in rebounding bandits, we cannot restart from the initial state because the satiation level cannot be reset and user's memory depends on past received recommendations. Instead, (Pike-Burke and Grunewalder, 2019) proposed a version of *w-step lookahead regret* that divides the T time steps into $\lceil T/w \rceil$ episodes where each episode (besides the last) consists of w time steps. At the beginning of each episode, the initial state is reset but depends on how the learner has interacted with the user previously. In particular, at the beginning of episode $i+1$ (at time $t = iw+1$), given that the learner has played $\pi_{1:iw}$ with corresponding pull sequence $u_{k,0:iw}$ for $k \in [K]$, we reset the initial state to be $x^i = (\mu_{1,iw+1}(u_{1,0:iw}), n_{1,iw+1}, \ldots, \mu_{K,iw+1}(u_{K,0:iw}), n_{K,iw+1})$ where $\mu_{k,t}(\cdot)$ is defined in (4.3) and $n_{k,iw+1}$ is the number of steps since arm k is last pulled by the learner as of time $iw+1$. Then, given the learner's policy $\pi_{1:T}$, where $\pi_t : \mathcal{X} \to [K]$, the w-step lookahead regret, against a competitor class $\mathcal{C}^w$ (which we define later), is defined as follows:

$$\begin{aligned}\mathrm{Reg}^w(T) = \sum_{i=0}^{\lceil T/w \rceil - 1} \max_{\tilde{\pi}_{1:w} \in \mathcal{C}^w} \mathbb{E}\left[\sum_{j=1}^{\min\{w, T-iw\}} r(x_{iw+j}, \tilde{\pi}_j(x_{iw+j})) \middle| x_{iw+1} = x^i\right] \\ - \mathbb{E}\left[\sum_{j=1}^{\min\{w, T-iw\}} r(x_{iw+j}, \pi_{iw+j}(x_{iw+j})) \middle| x_{iw+1} = x^i\right], \end{aligned} \tag{4.6}$$

where the expectation is taken over $x_{iw+2}, \ldots, x_{\min\{(i+1)w, T\}}$.

The competitor class $\mathcal{C}^w$ that we have chosen consists of policies that depend on time steps, i.e., $\mathcal{C}^w = \{\tilde{\pi}_{1:w} : \tilde{\pi}_t = \tilde{\pi}_t(x_t) = \tilde{\pi}_t(x'_t), \tilde{\pi}_t \in [K], \forall t \in [w], x_t, x'_t \in \mathcal{X}\}$. We note that $\mathcal{C}^w$ subsumes many traditional competitor classes in bandits literature, including the class of fixed-action policies considered in adversarial bandits (Lattimore and Szepesvári, 2020) and the class of periodic ranking policies (Cella and Cesa-Bianchi, 2020). In our paper, the w-lookahead policy (including the T-lookahead policy given by (4.4)) is a time-dependent policy that belongs to $\mathcal{C}^w$, since at time t, it will play a fixed action by solving (4.5) using the true reward parameters $\{\lambda_k, \gamma_k, b_k\}_{k=1}^K$. The time-dependent competitor class $\mathcal{C}^w$ differs from a state-dependent competitor class which includes all measurable functions $\tilde{\pi}_t$ that map from $\mathcal{X}$ to $[K]$. The state-dependent competitor class contains the optimal policy π^* where $\pi_t^*(x_t)$ depends on not just the time step but also the exact state x_t. Finding the optimal state-dependent policy requires optimal planning for a continuous-state MDP, which relies on state space discretizion (Ortner and Ryabko, 2012) or function approximation (e.g., approximate dynamic programming algorithms (Munos, 2007; Ernst et al., 2005; Riedmiller, 2005)). In Appendix C.3, we provide discussion and analysis on an algorithm compared against the optimal state-dependent policy. We proceed the rest of the main paper with $\mathcal{C}^w$ defined above.

When $w = 1$, the 1-step lookahead regret is also known as the instantaneous regret, which is commonly used in restless bandits literature and some nonstationary bandits papers including (Mintz et al., 2020). Note that low instantaneous regret does not imply high expected cumulative reward in the long-term, i.e., one may benefit more by waiting for certain arms to rebound. When $w = T$, we recover the full horizon regret. As we have noted earlier, finding the optimal competitor policy in this case is computationally intractable because the number of states, even when the satiation dynamics are deterministic, grows exponentially with the horizon T. Finally, we note that the w-step lookahead regret can be obtained for not just policies designed to look w steps ahead but any given policy. For a more comprehensive discussion on these notions of regret, see (Pike-Burke and Grunewalder, 2019, Section 4).

4.6 Explore-Estimate-Plan

We now present *Explore-Estimate-Plan (EEP)*, an algorithm for learning in rebounding bandits with stochastic dynamics and unknown parameters, that (i) collects data by pulling each arm a fixed number of times; (ii) estimates the model's parameters based on the logged data; and then (iii) plans according to the estimated model. Finally, we analyze EEP's regret.

Because each arm's base reward is known from the first pull, whenever arm k is pulled at time t and $n_{k,t} \neq 0$, we measure the satiation influence $\lambda_k s_{k,t}$, which becomes the next state $x_{k,t+1}$:

$$\begin{aligned} x_{k,t+1} = \lambda_k s_{k,t} &= \lambda_k \gamma_k^{n_{k,t}} s_{k,t-n_{k,t}} + \lambda_k \gamma_k^{n_{k,t}} + \lambda_k \textstyle\sum_{i=0}^{n_{k,t}-1} \gamma_k^i z_{k,t-1-i} \\ &= \gamma_k^{n_{k,t}} x_{k,t+1-n_{k,t}} + \lambda_k \gamma_k^{n_{k,t}} + \lambda_k \textstyle\sum_{i=0}^{n_{k,t}-1} \gamma_k^i z_{k,t-1-i}. \end{aligned} \tag{4.7}$$

We note that the current state $x_{k,t}$ equals $x_{k,t+1-n_{k,t}}$, since $x_{k,t+1-n_{k,t}}$ is the last observed satiation influence for arm k and $n_{k,t}$ is the number of steps since arm k was last pulled.

4.6.1 The Exploration Phase: Repeated Pulls

We collect a dataset $\mathcal{P}_k^n$ by consecutively pulling each arm $n+1$ times, in turn, where $n \geq \lfloor T^{2/3}/K \rfloor$ (Line 4-7 of Algorithm 4.1). Specifically, for each arm $k \in [K]$, the dataset $\mathcal{P}_k^n$ contains a single trajectory of $n+1$ observed satiation influences $\tilde{x}_{k,1}, \ldots, \tilde{x}_{k,n+1}$, where $\tilde{x}_{k,1} = 0$ and $\tilde{x}_{k,j}$ $(j > 1)$ is the difference between the first reward and the j-th reward from arm k. Thus, for $\tilde{x}_{k,j}, \tilde{x}_{k,j+1} \in \mathcal{P}_k^n$, using (4.7) with $n_{k,t} = 1$ (because pulls are consecutive), it follows that

$$\tilde{x}_{k,j+1} = \gamma_k \tilde{x}_{k,j} + d_k + \tilde{z}_{k,j}, \tag{4.8}$$

where $d_k = \lambda_k \gamma_k$ and $\tilde{z}_{k,j}$ are independent samples from $\mathcal{N}(0, \sigma_{z,k}^2)$ with $\sigma_{z,k}^2 = \lambda_k^2 \sigma_z^2$. In Appendix C.5.2, we discuss other exploration strategies (e.g., playing the arms in a cyclic order) for EEP and their regret guarantees.

4.6.2 Estimating the Reward Model and Satiation Dynamics

For all $k \in [K]$, given the dataset $\mathcal{P}_k^n$, we estimate $A_k = (\gamma_k, d_k)^\top$ using the *ordinary least squares estimator*:

$$\widehat{A}_k \in \arg\min_{A \in \mathbb{R}^2} \|\mathbf{Y_k} - \overline{\mathbf{X}}_\mathbf{k} A\|_2^2, \tag{4.9}$$

where $\mathbf{Y_k} \in \mathbb{R}^n$ is an n-dimensional vector whose j-th entry is $\tilde{x}_{k,j+1}$ and $\overline{\mathbf{X}}_\mathbf{k} \in \mathbb{R}^{n \times 2}$ takes as its j-th row the vector $\overline{x}_{k,j} = (\tilde{x}_{k,j}, 1)^\top$, i.e., $\tilde{x}_{k,j+1}$ is treated to be the response to the covariates $\overline{x}_{k,j}$. For $n \geq 2$, we have that

$$\widehat{A}_k = \begin{pmatrix} \widehat{\gamma}_k \\ \widehat{d}_k \end{pmatrix} = \left(\overline{\mathbf{X}}_\mathbf{k}{}^\top \overline{\mathbf{X}}_\mathbf{k}\right)^{-1} \overline{\mathbf{X}}_\mathbf{k}{}^\top \mathbf{Y_k}, \tag{4.10}$$

and we take $\widehat{\lambda}_k = |\widehat{d}_k / \widehat{\gamma}_k|$.

The difficulty in analyzing the ordinary least squares estimator (4.10) for identifying an affine dynamical system (4.8) using a single trajectory of data comes from the fact that the samples are not independent. Asymptotic guarantees of the ordinary least squares estimators in this case have been studied previously in the control theory and time series communities (Hamilton, 1994; Ljung, 1999). Recent work on system identifications for linear dynamical systems focuses on the sample complexity (Simchowitz et al., 2018; Sarkar and Rakhlin, 2019). Adapting the proof of (Simchowitz et al., 2018, Theorem 2.4), we derive the following theorem for identifying our affine dynamical system (4.8).

Theorem 4.3

Fix $\delta \in (0, 1)$. For all $k \in [K]$, there exists a constant $n_0(\delta, k)$ such that if the dataset $\mathcal{P}_k^n$ satisfies $n \geq n_0(\delta, k)$, then

$$\mathbb{P}\left(\|\widehat{A}_k - A_k\|_2 \gtrsim \sqrt{1/(\psi n)}\right) \leq \delta,$$

where $\psi = \sqrt{\min\left\{\frac{\sigma_{z,k}^2 (1-\gamma_k)^2}{16 d_k^2 (1-\gamma_k^2) + (1-\gamma_k)^2 \sigma_{z,k}^2}, \frac{\sigma_{z,k}^2}{4(1-\gamma_k^2)}\right\}}$.

As shown in Theorem 4.3, when $d_k = \lambda_k \gamma_k$ gets larger, the convergence rate for $\widehat{A}_k$ gets slower. Given a single trajectory of sufficient length, we obtain $|\widehat{\gamma}_k - \gamma_k| \leq O(1/\sqrt{n})$ and $|\widehat{d}_k - d_k| \leq O(1/\sqrt{n})$. In Corollary 4.1, we show that the estimator of λ_k also achieves $O(1/\sqrt{n})$ estimation error.

Corollary 4.1. *Fix $\delta \in (0, 1)$. Suppose that for all $k \in [K]$, we have $\mathbb{P}(\|\widehat{A}_k - A_k\|_2 \gtrsim 1/\sqrt{n}) \leq \delta$ and $\widehat{\gamma}_k > 0$. Then, with probability $1 - \delta$, we have that for all $k \in [K]$,*

$$|\widehat{\gamma}_k - \gamma_k| \leq O\left(\frac{1}{\sqrt{n}}\right), \quad |\widehat{\lambda}_k - \lambda_k| \leq O\left(\frac{1}{\sqrt{n}}\right).$$

Algorithm 4.1 w-lookahead Explore-Estimate-Plan

Require: Lookahead window size w, Number of arms K, Horizon T **Initialize** $t = 1$, $\pi_{1:T}$ to be an empty array of length T and $\widetilde{T} = T^{2/3} + w - (T^{2/3} \bmod w)$.
for $k = 1, \dots, K$ **do**
Set $t' = t$ and initialize an empty array $\mathcal{P}_k^n$.

for $c = 0, \dots, \lfloor \widetilde{T}/K \rfloor$ **do**
Play arm k to obtain reward $\mu_{k,t'+c}$ and add $\mu_{k,t'} - \mu_{k,t'+c}$ to $\mathcal{P}_k^n$.
Set $\pi_t = k$ and increase t by 1.
end for
Obtain $\widehat{\gamma}_k, \widehat{d}_k$ using the estimator (4.9),set $\widehat{\lambda}_k = |\widehat{d}_k/\widehat{\gamma}_k|$ and $\widehat{b}_k = \mu_{k,t'}$.
end for Let $t_0 = \widetilde{T}$, set $\pi_{t:t_0} = (1, \dots, \widetilde{T} - t + 1)$, and play $\pi_{t:t_0}$.
for $i = 1, \dots, \lceil \frac{T-t_0}{w} \rceil$ **do**
Set $t_i = \min\{t_{i-1} + w, T\}$.
Obtain $\pi_{t_{i-1}+1:t_i} = \texttt{Lookahead}(\{\widehat{\lambda}_k, \widehat{\gamma}_k, \widehat{b}_k\}_{k=1}^K, \{u_{k,0:t_{i-1}}\}_{k=1}^K, t_{i-1}, t_i)$ where $\{u_{k,0:t_{i-1}}\}_{k=1}^K$ are the arm pull histories correspond to $\pi_{1:t_{i-1}}$.
Play $\pi_{t_{i-1}+1,t_i}$.
end for

4.6.3 Planning and Regret Bound

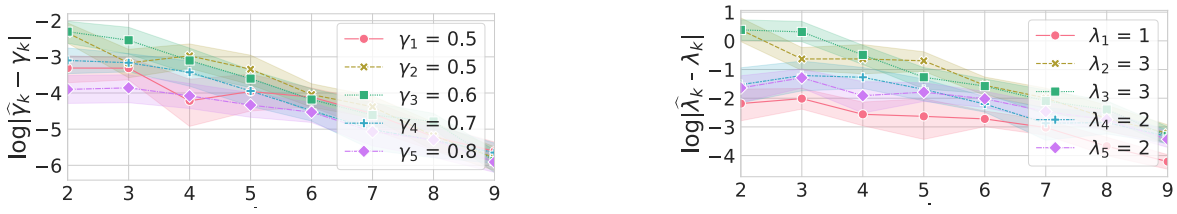
In the planning stage of Algorithm 4.1 (Line 11-15), at time $t_{i-1} + 1$, the next w arms to play are obtained through the Lookahead function defined in (4.5) based on the estimated parameters from the estimation stage (Line 8). Using the results in Corollary 4.1, we obtain the following sublinear regret bound for w-lookahead EEP.

Theorem 4.4

There exists a constant T_0 such that for all $T > T_0$ and $w \leq T^{2/3}$, the w-step lookahead regret of w-lookahead Explore-Estimate-Plan satisfies

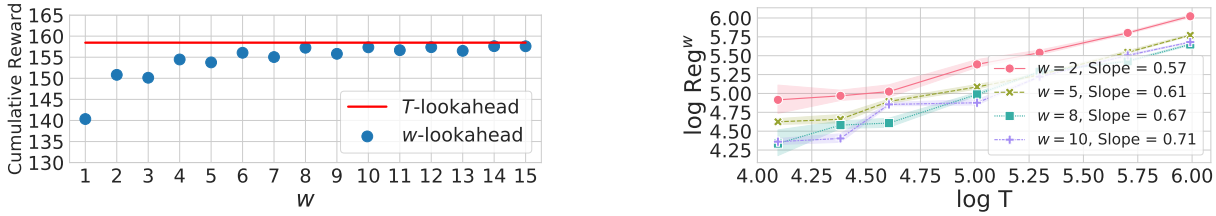
$$\text{Reg}^w(T) \leq O(K^{1/2} T^{2/3} \log T).$$

Remark 4.5. *The fact that EEP incurs a regret of order $O(T^{2/3})$ is expected for two reasons: First, EEP can be viewed as an explore-then-commit (ETC) algorithm that first explores then exploits. The regret of EEP resembles the $O(T^{2/3})$ regret of the ETC algorithm in the classical K-armed bandits setting (Lattimore and Szepesvári, 2020). In rebounding bandits, the fundamental obstacle to mixing the exploration and exploitation stages is the need to estimate the satiation dynamics. When the rewards of each arm are not observed periodically, the obtained satiation influences can no longer be viewed as samples from the same time-invariant affine dynamical system, since the parameters of the system depend on the duration between pulls. In practice, one may utilize the maximum likelihood estimator to obtain estimates of the reward parameters but obtaining the sample complexity of such an estimator with dependent data is difficult. Second, it has been shown in (Besbes et al., 2019) that when the rewards of the arms have temporal variation that depends on*



(a) $\log|\widehat{\gamma}_k - \gamma_k|$ v.s. $\log n$ **(b)** $\log|\widehat{\lambda}_k - \lambda_k|$ v.s. $\log n$

Figure 4.2: Figure 4.2a and 4.2b are the log-log plots of absolute errors of $\widehat{\gamma}_k$ and $\widehat{\lambda}_k$ with respect to the number of samples n in a single trajectory. The results are averaged over 30 random runs, where the shaded area represents one standard deviation.



(a) $T = 30$ **(b)** $\log \mathrm{Reg}^w$ v.s. $\log T$

Figure 4.3: Figure 4.3a shows the expected cumulative reward collected by the T-lookahead policy (red line) and w-lookahead policy (blue dots) when $T = 30$. Figure 4.3b shows the log-log plot of the w-step lookahead regret of w-lookahead EEP under different T averaged over 20 random runs.

the horizon T, the worst case instantaneous regret has a lower bound $\Omega(T^{2/3})$. On the other hand, in the traditional K-armed bandits setup, the regret (following the classical definition (Lattimore and Szepesvári, 2020)) is lower bounded by $\Omega(T^{1/2})$, and can be attained by methods like the upper confidence bound algorithm (Lattimore and Szepesvári, 2020). Precisely characterizing the regret lower bound for rebounding bandits is of future interest.

4.7 Experiments

We now evaluate the performance of EEP experimentally, separately investigating the sample efficiency of our proposed estimators (4.10) for learning the satiation and reward models (Figure 4.2) and the computational performance of the w-lookahead policies (4.5) (Figure 4.3a). For the experimental setup, we have 5 arms with satiation retention factors $\gamma_1 = \gamma_2 = .5, \gamma_3 = .6, \gamma_4 = .7, \gamma_5 = .8$, exposure influence factors $\lambda_1 = 1, \lambda_2 = \lambda_3 = 3, \lambda_4 = \lambda_5 = 2$, base rewards $b_1 = 2, b_2 = 3, b_3 = 4, b_4 = 2, b_5 = 10$, and noise with variance $\sigma_z = 0.1$.

Parameter Estimation We first evaluate our proposed estimator for using a single trajectory per arm to estimate the arm parameters γ_k, λ_k. In Figure 4.2, we show the absolute error (averaged over 30 random runs) between the estimated parameters and the true parameters for each arm. Aligning with our theoretical guarantees (Corollary 4.1), the log-log plots show that the convergence rate of the absolute error is on the scale of $O(n^{-1/2})$.

w-lookahead Performance To evaluate w-lookahead policies, we solve (4.5) using the true reward parameters and report expected cumulative rewards of the obtained w-lookahead policies (Figure 4.3a). Recall that the greedy policy is precisely the 1-lookahead policy. In order to solve the resulting integer programs, we use Gurobi 9.1 (LLC, 2021) and set the number of threads for solving the problem to 10. When $T = 30$, the T-lookahead policy (expected cumulative rewards given by the red line in Figure 4.3a) solved through (4.4) is obtained in 1610s. On the other hand, all w-lookahead policies (expected cumulative rewards given by the blue dots in Figure 4.3a) for w in between 1 and 15 are solved within 2s. We provide the results when $T = 100$ in Appendix C.7. Despite using significantly lower computational time, w-lookahead policies achieve a similar expected cumulative reward to the T-lookahead policy.

EEP Performance We evaluate the performance of EEP when T ranges from 60 to 400. For each horizon T, we examine the w-step lookahead regret of w-lookahead EEP where $w = 2, 5, 8, 10$. All results are averaged over 20 random runs. As T increases, the exploration stage of EEP becomes longer, which results in collecting more data for estimating the reward parameters and lower variance of the parameter estimators. We fit a line for the regrets with the same lookahead size w to examine the order of the regret with respect to the horizon T. The slopes of the lines (see Figure 4.3b's legend) are close to $2/3$, which aligns with our theoretical guarantees (Theorem 4.4), i.e., the regrets are on the order of $O(T^{2/3})$. In Appendix C.7, we present additional experimental setups and results.

4.8 Discussion

While our work has taken strides towards modeling the exposure-dependent evolution of preferences through dynamical systems, there are many avenues for future work. First, while our satiation dynamics are independent across arms, a natural extension might allow interactions among the arms. For example, a diner sick of pizza after too many trips to Di Fara's, likely would also avoid Grimaldi's until the satiation effect wore off. On the system identification side, we might overcome our reliance on evenly spaced pulls, producing more adaptive algorithms (e.g., optimism-based algorithms) that can refine their estimates, improving the agent's policy even past the pure exploration period. Finally, our satiation model captures just one plausible dynamic according to which preferences might evolve in response to past recommendations. Characterizing other such dynamics (e.g., the formation of brand loyalty where the rewards of an arm increase with more pulls) in bandits setups is of future interest.

Chapter 5

Modeling Attrition in Recommender Systems with Departing Bandits

5.1 Introduction

At the heart of online services spanning such diverse industries as media consumption, dating, financial products, and more, recommendation systems (RSs) drive personalized experiences by making curation decisions informed by each user's past history of interactions. While in practice, these systems employ diverse statistical heuristics, much of our theoretical understanding of them comes via stylized formulations within the multi-armed bandits (MABs) framework. While MABs abstract away from many aspects of real-world systems they allow us to extract crisp insights by formalizing fundamental tradeoffs, such as that between exploration and exploitation that all RSs must face (Joseph et al., 2016; Liu and Ho, 2018; Patil et al., 2020; Ron et al., 2021). As applies to RSs, exploitation consists of continuing to recommend items (or categories of items) that have been observed to yield high rewards in the past, while exploration consists of recommending items (or categories of items) about which the RS is uncertain but that could *potentially* yield even higher rewards.

In traditional formalizations of RSs as MABs, the recommender's decisions affect only the rewards obtained. However, real-life recommenders face a dynamic that potentially alters the exploration-exploitation tradeoff: Dissatisfied users have the option to depart the system, never to return. Thus, recommendations in the service of exploration not only impact instantaneous rewards but also risk driving away users and therefore can influence long-term cumulative rewards by shortening trajectories of interactions.

In this work, we propose *departing bandits* which augment conventional MABs by incorporating these policy-dependent horizons. To motivate our setup, we consider the following example: An RS for recommending blog articles must choose at each time among two categories of articles, e.g., economics and sports. Upon a user's arrival, the RS recommends articles sequentially. After each recommendation, the user decides whether to "click" the article and continue to the next recommendation, or to "not click" and may leave the system. Crucially, the user interacts with

the system for a random number of rounds. The user's departure probability depends on their satisfaction from the recommended item, which in turn depends on the user's unknown *type*. A user's type encodes their preferences (hence the probability of clicking) on the two topics (economics and sports).

When model parameters are given, in contrast to traditional MABs where the optimal policy is to play the best fixed arm, departing bandits require more careful analysis to derive an optimal planning strategy. Such planning is a local problem, in the sense that it is solved for each user. Since the user type is never known explicitly (the recommender must update its beliefs over the user types after each interaction), finding an optimal recommendation policy requires solving a specific partially observable MDP (POMDP) where the user type constitutes the (unobserved) state (more details in Section 5.5.1). When the model parameters are unknown, we deal with a learning problem that is global, in the sense that the recommender (learner) is learning for a stream of users instead of a particular user.

We begin with a formal definition of departing bandits in Section 5.2, and demonstrate that any fixed-arm policy is prone to suffer linear regret. In Section 5.3, we establish the UCB-based learning framework used in later sections. We instantiate this framework with a single user type in Section 5.4, where we show that it achieves $\tilde{O}(\sqrt{T})$ regret for T being the number of users. We then move to the more challenging case with two user types and two recommendation categories in Section 5.5. To analyze the planning problem, we effectively reduce the search space for the optimal policy by using a closed-form of the *expected return* of any recommender policy. These results suggest an algorithm that achieves $\tilde{O}(\sqrt{T})$ regret in this setting. In the full version of this paper (**Ben-Porat22**), we also show an efficient optimal planning algorithm for multiple user types and two recommendation categories, and describe a scheme to construct semi-synthetic problem instances for this setting using real-world datasets.

5.1.1 Related Work

MABs have been studied extensively by the online learning community (Cesa-Bianchi and Lugosi, 2006; Bubeck, Cesa-Bianchi, et al., 2012). The contextual bandit literature augments the MAB setup with context-dependent rewards (Abbasi-Yadkori et al., 2011; Slivkins, 2019; Mahadik et al., 2020; Korda et al., 2016; Lattimore and Szepesvári, 2020). In contextual bandits, the learner observes a *context* before they make a decision, and the reward depends on the context. Another line of related work considers the dynamics that emerge when users act strategically (Kremer et al., 2014; Mansour et al., 2015; Cohen and Mansour, 2019; Bahar et al., 2016; Bahar et al., 2020). In that line of work, users arriving at the system receive a recommendation but act strategically: They can follow the recommendation or choose a different action. This modeling motivates the development of incentive-compatible mechanisms as solutions. In our work, however, the users are modeled in a stochastic (but not strategic) manner. Users may leave the system if they are dissatisfied with recommendations, and this departure follows a fixed (but possibly unknown) stochastic model.

The departing bandits problem has two important features: Policy-dependent horizons, and multiple user types that can be interpreted as unknown states. Existing MAB works (Azar et al.,

2013; Cao et al., 2020) have addresses these phenomena separately but we know of no work that integrates the two in a single framework. In particular, while Azar et al. (2013) study the setting with multiple user types, they focus on a fixed horizon setting. Additionally, while Cao et al. (2020) deal with departure probabilities and policy-dependent interaction times for a single user type, they do not consider the possibility of multiple underlying user types.

The planning part of our problem falls under the framework of using Markov Decision Processes for modeling recommender-user dynamics (Shani et al., 2005b). Specifically, our problem works with partially observable user states which have also been seen in many recent bandits variants (Pike-Burke and Grunewalder, 2019; Leqi et al., 2021b). Unlike these prior works that focus on interactions with a single user, departing bandits consider a stream of users each of which has an (unknown) type selected among a finite set of user types.

More broadly, our RS learning problem falls under the domain of reinforcement learning (RL). Existing RL literature that considers departing users in RSs include Zhao et al. (2020b), Lu and Yang (2016), and Zhao et al. (2020a). While Zhao et al. (2020b) handle users of a single type that depart the RS within a bounded number of interactions, our work deals with multiple user types. In contrast to Zhao et al. (2020a), we consider an online setting and provide regret guarantees that do not require bounded horizon. Finally, Lu and Yang (2016) use POMDPs to model user departure and focus on approximating the value function. They conduct an experimental analysis on historical data, while we devise an online learning algorithm with theoretical guarantees.

5.2 Departing Bandits

We propose a new online problem, called *departing bandits*, where the goal is to find the optimal recommendation algorithm for users of (unknown) types, and where the length of the interactions depends on the algorithm itself. Formally, the departing bandits problem is defined by a tuple $\langle [M], [K], \mathbf{q}, \mathbf{P}, \mathcal{L} \rangle$, where M is the number of user *types*, K is the number of *categories*, $\mathbf{q} \in [0,1]^M$ specifies a prior distribution over types, and $\mathbf{P} \in (0,1)^{K \times M}$ and $\mathcal{L} \in (0,1)^{K \times M}$ are the *click-probability* and the *departure-probability* matrices, respectively.[1]

There are T users who arrive sequentially at the RS. At every episode, a new user $t \in [T]$ arrives with a type $type(t)$. We let $\mathbf{q}$ denote the prior distribution over the user types, i.e., $type(t) \sim \mathbf{q}$. Each user of type x *clicks* on a recommended category a with probability $\mathbf{P}_{a,x}$. In other words, each click follows a Bernoulli distribution with parameter $\mathbf{P}_{a,x}$. Whenever the user clicks, she stays for another iteration, and when the user does not click (no-click), she *departs* with probability $\mathcal{L}_{a,x}$ (and stays with probability $1 - \mathcal{L}_{a,x}$). Each user t interacts with the RS (the learner) until she departs.

We proceed to describe the user-RS interaction protocol. In every iteration j of user t, the learner recommends a category $a \in [K]$ to user t. The user clicks on it with probability $\mathbf{P}_{a,type(t)}$. If the user clicks, the learner receives a reward of $r_{t,j}(a) = 1$.[2] If the user does not click, the

[1] We denote by $[n]$ the set $\{1, \ldots, n\}$.

[2] We formalize the reward as is standard in the online learning literature, from the perspective of the learner. However, defining the reward from the user perspective by, e.g., considering her utility as the number of clicks she gives or the number of articles she reads induces the same model.

Algorithm 5.1 The Departing Bandits Protocol

Input: number of types M, number of categories K, and number of users (episodes) T

Hidden Parameters: types prior $\mathbf{q}$, click-probability $\mathbf{P}$, and departure-probability $\mathcal{L}$

for episode $t \leftarrow 1, \ldots, T$ **do**
  a new user with type $type(t) \sim \mathbf{q}$ arrives
  $j \leftarrow 1$, $depart \leftarrow false$
  while $depart$ is $false$ **do**
    the learner picks a category $a \in [K]$
    with probability $\mathbf{P}_{a,x}$, user t clicks on a and $r_{t,j}(a) \leftarrow 1$; otherwise, $r_{t,j}(a) \leftarrow 0$
    if $r_{t,j}(a) = 0$ **then**
      with probability $\mathcal{L}_{a,x}$: $depart \leftarrow true$ and user t departs
    end if
    the learner observes $r_{t,j}(a)$ and $depart$
    if $depart$ is $false$ **then**
      $j \leftarrow j + 1$
    end if
  end while
end for

learner receives no reward (i.e., $r_{t,j}(a) = 0$), and user t departs with probability $\mathcal{L}_{a,type(t)}$. We assume that the learner knows the value of a constant $\epsilon > 0$ such that $\max_{a,x} \mathbf{P}_{a,x} \leq 1 - \epsilon$ (i.e., ϵ does not depend on T). When user t departs, she does not interact with the learner anymore (and the learner moves on to the next user $t+1$). For convenience, the departing bandits problem protocol is summarized in Algorithm 5.1.

Having described the protocol, we move on to the goals and performance of the learner. Without loss of generality, we assume that the online learner's recommendations are made based on a *policy* π, which is a mapping from the history of previous interactions (with that user) to recommendation categories. For each user (episode) $t \in [T]$, the learner selects a policy π_t that recommends category $\pi_{t,j} \in [K]$ at every iteration $j \in [N^{\pi_t}(t)]$, where $N^{\pi_t}(t)$ denotes the episode length (i.e., total number of iterations policy π_t interacts with user t until she departs).[3] The *return* of a policy π, denoted by V^π is the cumulative reward the learner obtains when executing the policy π until the user departs. Put differently, the return of π from user t is the random variable $V^\pi = \sum_{j=1}^{N^\pi(t)} r_{t,j}(\pi_{t,j})$.

We denote by π^* an optimal policy, namely a policy that maximizes the expected return, $\pi^* = \arg\max_\pi \mathbb{E}[V^\pi]$. Similarly, we denote by V^* the optimal return, i.e., $V^* = V^{\pi^*}$.

We highlight two algorithmic tasks. The first is the planning task, in which the goal is to find an optimal policy π^*, given $\mathbf{P}, \mathcal{L}, \mathbf{q}$. The second is the online learning task. We consider settings where the learner knows the number of categories, K, the number of types, M, and the number of users, T, but has no prior knowledge regarding $\mathbf{P}$, $\mathcal{L}$ or $\mathbf{q}$. In the online learning task,

[3] We limit the discussion to deterministic policies solely; this is w.l.o.g. (see Subsection 5.5.1 for further details).

	Type x	Type y
Category 1	$\mathbf{P}_{1,x} = 0.5$	$\mathbf{P}_{1,y} = 0.28$
Category 2	$\mathbf{P}_{2,x} = 0.4$	$\mathbf{P}_{2,y} = 0.39$
Prior	$\mathbf{q}_x = 0.4$	$\mathbf{q}_y = 0.6$

Table 5.1: The departing bandits instance in Section 5.2.1.

the *value* of the learner's algorithm is the sum of the returns obtained from all the users, namely

$$\sum_{t=1}^{T} V^{\pi_t} = \sum_{t=1}^{T} \sum_{j=1}^{N^{\pi}(t)} r_{t,j}(\pi_{t,j}).$$

The performance of the leaner is compared to that of the best policy, formally defined by the *regret* for T episodes,

$$R_T = T \cdot \mathbb{E}[V^{\pi^*}] - \sum_{t=1}^{T} V^{\pi_t}. \tag{5.1}$$

The learner's goal is to minimize the expected regret $\mathbb{E}[R_T]$.

5.2.1 Example

The motivation for the following example is two-fold. First, to get the reader acquainted with our notations; and second, to show why fixed-arm policies are inferior in our setting.

Consider a problem instance with two user types ($M = 2$), which we call x and y for convenience. There are two categories ($K = 2$), and given no-click the departure is deterministic, i.e., $\mathcal{L}_{a,\tau} = 1$ for every category $a \in [K]$ and type $\tau \in [M]$. That is, every user leaves immediately if she does not click. Furthermore, let the click-probability $\mathbf{P}$ matrix and the user type prior distribution $\mathbf{q}$ be as in Table 5.1.

Looking at $\mathbf{P}$ and $\mathbf{q}$, we see that Category 1 is better for Type x, while Category 2 is better for type y. Notice that without any additional information, a user is more likely to be type y. Given the prior distribution, recommending Category 1 in the first round yields an expected reward of $\mathbf{q}_x\mathbf{P}_{1,x} + \mathbf{q}_y\mathbf{P}_{1,y} = 0.368$. Similarly, recommending Category 2 in the first round results in an expected reward of 0.394. Consequently, if we recommend *myopically*, i.e., without considering the user type, always recommending Category 2 is better than always recommending Category 1.

Let π^a denote the fixed-arm policy that always selects a single category a. Using the tools we derive in Section 5.5 and in particular Theorem 5.3, we can compute the expected returns of π^1 and π^2, $\mathbb{E}[V^{\pi^1}]$ and $\mathbb{E}[V^{\pi^2}]$. Additionally, using results from Section 5.5.2, we can show that the optimal policy for the planning task, π^*, recommends Category 2 until iteration 7, and then recommends Category 1 for the rest of the iterations until the user departs.

Using simple calculations, we see that $\mathbb{E}[V^{\pi^*}] - \mathbb{E}[V^{\pi^1}] > 0.0169$ and $\mathbb{E}[V^{\pi^*}] - \mathbb{E}[V^{\pi^2}] > 1.22 \times 10^{-5}$; hence, the expected return of the optimal policy is greater than the returns of both

Algorithm 5.2 UCB-based algorithm with hybrid radii: UCB-Hybrid (Jia et al., 2021)

Input: set of policies Π, number of users T, $\tilde{\tau}, \eta$
Initialize: $\forall \pi \in \Pi : U_0(\pi) \leftarrow \infty, n(\pi) = 0$
for user $t \leftarrow 1, \ldots, T$ **do**
  Execute π_t such that $\pi_t \in \arg\max_{\pi \in \Pi} U_{t-1}(\pi)$ and receive return $\hat{V}^{\pi_t}[n(\pi_t)] \leftarrow \sum_{j=1}^{N^{\pi_t}(t)} r_{t,j}(\pi_{t,j})$
  $n(\pi_t) \leftarrow n(\pi_t) + 1$
  if $n(\pi_t) < 8\eta \ln T$ **then**
    Update $U_t(\pi_t) = \frac{\sum_{i=1}^{n(\pi_t)} \hat{V}^{\pi_t}[i]}{n(\pi_t)} + \frac{8\sqrt{\eta} \cdot \tilde{\tau} \ln T}{n(\pi_t)}$
  else
    Update $U_t(\pi_t) = \frac{\sum_{i=1}^{n(\pi_t)} \hat{V}^{\pi_t}[i]}{n(\pi_t)} + \sqrt{\frac{8\tilde{\tau}^2 \ln T}{n(\pi_t)}}$
  end if
end for

fixed-arm policies by a constant. As a result, if the learner only uses fixed-arm policies (π^a for every $a \in [K]$), she suffers linear expected regret, i.e., $\mathbb{E}[R_T] = T \cdot \mathbb{E}[V^{\pi^*}] - \sum_{t=1}^{T} \mathbb{E}[V^{\pi^a}] = \Omega(T)$.

5.3 UCB Policy for Sub-exponential Returns

In this section, we introduce the learning framework used in the paper and provide a general regret guarantee for it.

In standard MAB problems, at each $t \in [T]$ the learner picks a single arm and receives a single sub-Gaussian reward. In contrast, in departing bandits, at each $t \in [T]$ the learner receives a return V^{π}, which is the cumulative reward of that policy. The return V^{π} depends on the policy π not only through the obtained rewards at each iteration but also through the total number of iterations (trajectory length). Such returns are not necessarily sub-Gaussian. Consequently, we cannot use standard MAB algorithms as they usually rely on concentration bounds for sub-Gaussian rewards. Furthermore, as we have shown in Section 5.2.1, in departing bandits fixed-arm policies can suffer linear regret (in terms of the number of users), which suggests considering a more expressive set of policies. This in turn yields another disadvantage for using MAB algorithms for departing bandits, as their regret is linear in the number of arms (categories) K.

As we show later in Sections 5.4 and 5.5, for some natural instances of the departing bandits problem, the return from each user is sub-exponential (Definition 5.1). Algorithm 5.2, which we propose below, receives a set of policies Π as input, along with other parameters that we describe shortly. The algorithm is a restatement of the *UCB-Hybrid* Algorithm from Jia et al. (2021), with two modifications: (1) The input includes a set of policies rather than a set of actions/categories, and accordingly, the confidence bound updates are based on return samples (denoted by $\hat{V}^{\pi}$) rather than reward samples. (2) There are two global parameters ($\tilde{\tau}$ and η) instead of two local parameters per action. If the return from each policy in Π is sub-exponential, Algorithm 5.2 not only handles sub-exponential returns, but also comes with the following guarantee: Its expected

value is close to the value of the best policy in Π.

5.3.1 Sub-exponential Returns

For convenience, we state here the definition of sub-exponential random variables (Eldar and Kutyniok, 2012).

Definition 5.1

We say that a random variable X is *sub-exponential* with parameters (τ^2, b) if for every γ such that $|\gamma| < 1/b$,

$$\mathbb{E}[\exp(\gamma(X - \mathbb{E}[X]))] \leq \exp(\frac{\gamma^2\tau^2}{2}).$$

In addition, for every (τ^2, b)-sub-exponential random variables, there exist constants $C_1, C_2 > 0$ such that the above is equivalent to each of the following properties:

1. Tails: $\forall v \geq 0 : \mathbb{P}[|X| > v] \leq \exp(1 - \frac{v}{C_1})$.
2. Moments: $\forall p \geq 1 : (\mathbb{E}[|X|^p])^{1/p} \leq C_2 p$.

Let Π be a set of policies with the following property: There exist $\tilde{\tau}, \eta$ such that the return of every policy $\pi \in \Pi$ is (τ^2, b)-sub-exponential with $\tilde{\tau} \geq \tau$ and $\eta \geq \frac{b^2}{\tau^2}$. The following Algorithm 5.2 receives as input a set of policies Π with the associated parameters, $\tilde{\tau}$ and η. Similarly to the UCB algorithm, it maintains an upper confidence bound U for each policy, and balances between exploration and exploitation. Theorem 5.1 below shows that Algorithm 5.2 always gets a value similar to that of the best policy in Π up to an additive factor of $\tilde{O}\left(\sqrt{|\Pi|\,T} + |\Pi|\right)$. The theorem follows directly from Theorem 3 from Jia et al. (2021) by having policies as arms and returns as rewards.

Theorem 5.1

Let Π be a set of policies with the associated parameters $\tilde{\tau}, \eta$. Let $\pi_1, \ldots, \pi_T$ be the policies Algorithm 5.2 selects. It holds that

$$\mathbb{E}\left[\max_{\pi \in \Pi} T \cdot V^{\pi} - \sum_{t=1}^{T} V^{\pi_t}\right] = O(\sqrt{|\Pi|\,T \log T} + |\Pi| \log T).$$

There are two challenges in leveraging Theorem 5.1. The first challenge is crucial: Notice that Theorem 5.1 does not imply that Algorithm 5.2 has a low regret; its only guarantee is w.r.t. the policies in Π received as an input. As the number of policies is infinite, our success will depend on our ability to characterize a "good" set of policies Π. The second challenge is technical: Even if we find such Π, we still need to characterize the associated $\tilde{\tau}$ and η. This is precisely what we do in Section 5.4 and 5.5.

5.4 Single User Type

In this section, we focus on the special case of a single user type, i.e., $M = 1$. For notational convenience, since we only discuss single-type users, we associate each category $a \in [K]$ with its two unique parameters $\mathbf{P}_a := \mathbf{P}_{a,1}, \mathcal{L}_a := \mathcal{L}_{a,1}$ and refer to them as scalars rather than vectors. In addition, We use the notation N_a for the random variable representing the number of iterations until a random user departs after being recommended by π^a, the fixed-arm policy that recommends category a in each iteration.

To derive a regret bound for single-type users, we use two main lemmas: Lemma 5.1, which shows the optimal policy is fixed, and Lemma 5.2, which shows that returns of fixed-arm policies are sub-exponential and calculate their corresponding parameters. These lemmas allow us to use Algorithm 5.2 with a policy set Π that contains all the fixed-arm policies, and derive a $\tilde{O}(\sqrt{T})$ regret bound. All omitted proofs can be found in the full version of this paper (**Ben-Porat22**).

To show that there exists a category $a^* \in [K]$ for which π^{a^*} is optimal, we rely on the assumption that all the users have the same type (hence we drop the type subscripts t), and as a result the rewards of each category $a \in [K]$ have an expectation that depends on a single parameter, namely $\mathbb{E}[r(a)] = \mathbf{P}_a$. Such a category $a^* \in [K]$ does not necessarily have the maximal click-probability nor the minimal departure-probability, but rather an optimal combination of the two (in a way, this is similar to the knapsack problem, where we want to maximize the reward while having as little weight as possible). We formalize it in the following lemma.

Lemma 5.1

A policy π^{a^*} is optimal if

$$a^* \in \arg\max_{a \in [K]} \frac{\mathbf{P}_a}{\mathcal{L}_a(1 - \mathbf{P}_a)}.$$

As a consequence of this lemma, the planning problem for single-type users is trivial—the solution is a fixed-arm policy π^{a^*} given in the lemma. However, without access to the model parameters, identifying π^{a^*} requires learning. We proceed with a simple observation regarding the random number of iterations obtained by executing a fixed-arm policy. The observation would later help us show that the return of any fixed-arm policy is sub-exponential.

Observation 5.1. *For every $a \in [K]$ and every $\mathcal{L}_a > 0$, the random variable N_a follows a geometric distribution with success probability parameter $\mathcal{L}_a[1 - \mathbf{P}_a] \in (0, 1 - \epsilon]$.*

Using Observation 5.1 and previously known results (stated as Lemma D.3 in the appendix), we show that N_a is sub-exponential for all $a \in [K]$. Notice that return realizations are always upper bounded by the trajectory length; this implies that returns are also sub-exponential. However, to use the regret bound of Algorithm 5.2, we need information regarding the parameters (τ_a^2, b_a) for every policy π^a. We provide this information in the following Lemma 5.2.

Lemma 5.2

For each category $a \in [K]$, the centred random variable $V^{\pi^a} - \mathbb{E}[V^{\pi^a}]$ is sub-exponential with parameters (τ_a^2, b_a), such that

$$\tau_a = b_a = -\frac{8e}{\ln(1 - \mathcal{L}_a(1 - \mathbf{P}_a))}.$$

Proof sketch. We rely on the equivalence between the subexponentiality of a random variable and the bounds on its moments (Property 2 in Definition 5.1). We bound the expectation of the return V^{π^a}, and use Minkowski's and Jensen's inequalities to show in Lemma D.2 that $\mathbb{E}[|V^{\pi^a} - \mathbb{E}[V^{\pi^a}]|^p])^{1/p}$ is upper bounded by $-4/\ln(1-\mathcal{L}_a(1-\mathbf{P}_a))$ for every $a \in [K]$ and $p \geq 1$. Finally, we apply a normalization trick and bound the Taylor series of $\mathbb{E}[\exp(\gamma(V^{\pi^a} - \mathbb{E}[V^{\pi^a}]))]$ to obtain the result. □

An immediate consequence of Lemma 5.2 is that the parameters $\tilde{\tau} = 8e/\ln(\frac{1}{1-\epsilon})$ and $\eta = 1$ are valid upper bounds for τ_a and b_a/τ_a^2 for each $a \in [K]$ (I.e., $\forall a \in [K] : \ \tilde{\tau} \geq \tau_a$ and $\eta \geq b_a^2/\tau_a^2$). We can now derive a regret bound using Algorithm 5.2 and Theorem 5.1.

Theorem 5.2

For single-type users ($M = 1$), running Algorithm 5.2 with $\Pi = \{\pi^a : a \in [K]\}$ and $\tilde{\tau} = \frac{8e}{\ln(\frac{1}{1-\epsilon})}$, $\eta = 1$ achieves an expected regret of at most

$$\mathbb{E}[R_T] = O(\sqrt{KT \log T} + K \log T).$$

5.5 Two User Types and Two Categories

In this section, we consider cases with two user types ($M = 2$), two categories ($K = 2$) and departure-probability $\mathcal{L}_{a,\tau} = 1$ for every category $a \in [K]$ and type $\tau \in [M]$. Even in this relatively simplified setting, where users leave after the first "no-click", planning is essential. To see this, notice that the event of a user clicking on a certain category provides additional information about the user, which can be used to tailor better recommendations; hence, algorithms that do not take this into account may suffer a linear regret. In fact, this is not just a matter of the learning algorithm at hand, but rather a failure of all fixed-arm policies; there are instances where all fixed-arm policies yield high regret w.r.t. the baseline defined in Equation (5.1). Indeed, this is what the example in Section 5.2.1 showcases. Such an observation suggests that studying the optimal planning problem is vital.

In Section 5.5.1, we introduce the partially observable MDP formulation of departing bandits along with notion of *belief-category walk*. We use this notion to provide a closed-form formula for policies' expected return, which we use extensively later on. Next, in Section 5.5.2 we characterize the optimal policy, and show that we can compute it in constant time relying on the closed-form

formula. This is striking, as generally computing optimal POMDP policies is computationally intractable since, e.g., the space of policies grows exponentially with the horizon. Conceptually, we show that there exists an optimal policy that depends on a belief threshold: It recommends one category until the posterior belief of one type, which is monotonically increasing, crosses the threshold, and then it recommends the other category. Finally, in Section 5.5.3 we leverage all the previously obtained results to derive a small set of threshold policies of size $O(\ln T)$ with corresponding sub-exponential parameters. Due to Theorem 5.1, this result implies a $\tilde{O}(\sqrt{T})$ regret.

5.5.1 Efficient Planning

To recap, we aim to find the optimal policy when the click-probability matrix and the prior over user types are known. Namely, given an instance in the form of $\langle \mathbf{P}, \mathbf{q} \rangle$, our goal is to efficiently find the optimal policy.

For planning purposes, the problem can be modeled by an episodic POMDP, $\langle S, [K], O, \text{Tr}, \mathbf{P}, \Omega, \mathbf{q}, O \rangle$. A set of states, $S = [M] \cup \{\perp\}$ that comprises all types $[M]$, along with a designated absorbing state $\perp$ suggesting that the user departed (and the episode terminated). $[K]$ is the set of the actions (categories). $O = \{stay, depart\}$ is the set of possible observations. The transition and observation functions, $\text{Tr} : S \times [K] \to S$ and $\Omega : S \times [K] \to O$ (respectively) satisfy $\text{Tr}(\perp | i, a) = \Omega(depart | i, a) = 1 - \mathbf{P}_{i,a}$ and $\text{Tr}(i | i, a) = \Omega(stay | i, a) = \mathbf{P}_{i,a}$ for every type $i \in [M]$ and action $a \in [K]$. Finally, $\mathbf{P}$ is the expected reward matrix, and $\mathbf{q}$ is the initial state distribution over the M types.

When there are two user types and two categories, the click-probability matrix is given by Table 5.2 where we note that the prior on the types holds $\mathbf{q}_y = 1 - \mathbf{q}_x$, thus can be represented by a single parameter $\mathbf{q}_x$.

Remark 5.1. *Without loss of generality, we assume that $\mathbf{P}_{1,x} \geq \mathbf{P}_{2,x}, \mathbf{P}_{1,y}, \mathbf{P}_{2,y}$ since one could always permute the matrix to obtain such a structure.*

Since the return and number of iterations for the same policy is independent of the user index, we drop the subscript t in the rest of this subsection and use .

	Type x	Type y
Category 1	$\mathbf{P}_{1,x}$	$\mathbf{P}_{1,y}$
Category 2	$\mathbf{P}_{2,x}$	$\mathbf{P}_{2,y}$
Prior	$\mathbf{q}_x$	$\mathbf{q}_y = 1 - \mathbf{q}_x$

Table 5.2: Click probabilities for two user types and two categories.

As is well-known in the POMDP literature (Kaelbling et al., 1998), the optimal policy π^* and its expected return are functions of belief states that represent the probability of the state at each time. In our setting, the states are the user types. We denote by b_j the belief that the state is (type) x at iteration j. Similarly, $1 - b_j$ is the belief that the state is (type) y at iteration j. Needless to say, once the state $\perp$ is reached, the belief over the type states $[M]$ is irrelevant, as users do not come back. Nevertheless, we neglect this case as our analysis does not make use it.

We now describe how to compute the belief. At iteration $j = 1$, the belief state is set to be $b_1 = \mathbb{P}(state = x) = \mathbf{q}_x$. At iteration $j > 1$, upon receiving a positive reward $r_j = 1$, the belief is updated from $b_{j-1} \in [0, 1]$ to

$$b_j(b_{j-1}, a, 1) = \frac{b_{j-1} \cdot \mathbf{P}_{a,x}}{b_{j-1} \cdot \mathbf{P}_{a,x} + \mathbf{P}_{a,y}(1 - b_{j-1})}, \tag{5.2}$$

where we note that in the event of no-click, the current user departs the system, i.e., we move to the absorbing state $\perp$. For any policy $\pi : [0, 1] \to \{1, 2\}$ that maps a belief to a category, its expected return satisfies the Bellman equation,

$$\mathbb{E}[V^\pi(b)] = \left(b\mathbf{P}_{\pi(b),x} + (1 - b)\mathbf{P}_{\pi(b),y}\right) \cdot (1 + \mathbb{E}[V^\pi(b'(b, \pi(b), 1))]).$$

To better characterize the expected return, we introduce the following notion of belief-category walk.

Definition 5.2: Belief-category walk

Let $\pi : [0, 1] \to \{1, 2\}$ be any policy. The sequence

$$b_1, a_1 = \pi(b_1), b_2, a_2 = \pi(b_2), \ldots$$

is called the *belief-category walk*. Namely, it is the induced walk of belief updates and categories chosen by π, given all the rewards are positive ($r_j = 1$ for every $j \in \mathbb{N}$).

Notice that every policy induces a single, well-defined and deterministic belief-category walk (recall that we assume departure-probabilities satisfy $\mathcal{L}_{a,\tau} = 1$ for every $a \in [K], \tau \in [M]$). Moreover, given any policy π, the trajectory of every user recommended by π is fully characterized by belief-category walk clipped at $b_{N^\pi(t)}, a_{N^\pi(t)}$.

In what follows, we derive a closed-form expression for the expected return as a function of b, the categories chosen by the policy, and the click-probability matrix.

Theorem 5.3

For every policy π and an initial belief $b \in [0, 1]$, the expected return is given by

$$\mathbb{E}[V^\pi(b)] = \sum_{i=1}^{\infty} b \cdot \mathbf{P}_{1,x}^{m_{1,i}} \cdot \mathbf{P}_{2,x}^{m_{2,i}} + (1 - b)\mathbf{P}_{1,y}^{m_{1,i}} \cdot \mathbf{P}_{2,y}^{m_{2,i}},$$

where $m_{1,i} := |\{a_j = 1, j \leq i\}|$ and $m_{2,i} := |\{a_j = 2, j \leq i\}|$ are calculated based on the belief-category walk $b_1, a_1, b_2, a_2, \ldots$ induced by π.

5.5.2 Characterizing the Optimal Policy

Using Theorem 5.3, we show that the planning problem can be solved in $O(1)$. To arrive at this conclusion, we perform a case analysis over the following three structures of the click-probability

matrix $\mathbf{P}$:

- *Dominant Row*, where $\mathbf{P}_{1,y} \geq \mathbf{P}_{2,y}$;
- *Dominant Column*, where $\mathbf{P}_{2,x} \geq \mathbf{P}_{2,y} > \mathbf{P}_{1,y}$;
- *Dominant Diagonal*, where $\mathbf{P}_{1,x} \geq \mathbf{P}_{2,y} > \mathbf{P}_{1,y}, \mathbf{P}_{2,x}$.

Crucially, any matrix $\mathbf{P}$ takes exactly one of the three structures. Further, since $\mathbf{P}$ is known in the planning problem, identifying the structure at hand takes $O(1)$ time. Using this structure partition, we characterize the optimal policy.

Dominant Row We start by considering the simplest structure, in which the Category 1 is preferred by both types of users: Since $\mathbf{P}_{1,y} \geq \mathbf{P}_{2,y}$ and $\mathbf{P}_{1,x} \geq \mathbf{P}_{2,x}, \mathbf{P}_{1,y}, \mathbf{P}_{2,y}$ (Remark 5.1), there exists a dominant row, i.e., Category 1.

Lemma 5.3

For any instance such that $\mathbf{P}$ has a dominant row a, the fixed policy π^a is an optimal policy.

As expected, if Category 1 is dominant then the policy that always recommends Category 1 is optimal.

Dominant Column In the second structure we consider the case where there is no dominant row, and that the column of type x is dominant, i.e., $\mathbf{P}_{1,x} \geq \mathbf{P}_{2,x} \geq \mathbf{P}_{2,y} > \mathbf{P}_{1,y}$. In such a case, which is also the one described in the example in Section 5.2.1, it is unclear what the optimal policy would be since none of the categories dominates the other.

Surprisingly, we show that the optimal policy can be of only one form: Recommend Category 2 for some time steps (possibly zero) and then always recommend Category 1. To identify when to switch from Category 2 to Category 1, one only needs to compare four expected returns.

Theorem 5.4

For any instance such that $\mathbf{P}$ has a dominant column, one of the following four policies is optimal:

$$\pi^1, \pi^2, \pi^{2:\lfloor N^* \rfloor}, \pi^{2:\lceil N^* \rceil},$$

where $N^* = N^*(\mathbf{P}, \mathbf{q})$ is a constant, and $\pi^{2:\lfloor N^* \rfloor}$ ($\pi^{2:\lceil N^* \rceil}$) stands for recommending Category 2 until iteration $\lfloor N^* \rfloor$ ($\lceil N^* \rceil$) and then switching to Category 1.

The intuition behind the theorem is as follows. If the prior tends towards type y, we might start with recommending Category 2 (which users of type y are more likely to click on). But after several iterations, and as long as the user stays, the posterior belief b increases since $\mathbf{P}_{2,x} > \mathbf{P}_{2,y}$ (recall Equation (5.2)). Consequently, since type x becomes more probable, and since $\mathbf{P}_{1,x} \geq \mathbf{P}_{2,x}$, the optimal policy recommends the best category for this type, i.e., Category 1. For the exact expression of N^*, we refer the reader to Appendix D.5.3.

Using Theorem 5.3, we can compute the expected return for each of the four policies in $O(1)$, showing that we can find the optimal policy when $\mathbf{P}$ has a column in $O(1)$.

Dominant Diagonal In the last structure, we consider the case where there is no dominant row (i.e., $\mathbf{P}_{2,y} > \mathbf{P}_{1,y}$) nor a dominant column (i.e., $\mathbf{P}_{2,y} > \mathbf{P}_{2,x}$). At first glance, this case is more complex than the previous two, since none of the categories and none of the types dominates the other one. However, we uncover that the optimal policy can be either always recommending Category 1 or always recommending Category 2. Theorem 5.5 summarizes this result.

Theorem 5.5

For any instance such that $\mathbf{P}$ has a dominant diagonal, either π^1 or π^2 is optimal.

With the full characterization of the optimal policy derived in this section (for all the three structures), we have shown that the optimal policy can be computed in $O(1)$.

5.5.3 Learning: UCB-based Regret Bound

In this section, we move from the planning task to the learning one. Building on the results of previous sections, we know that there must exist a threshold policy—a policy whose belief-category walk has a finite prefix of one category, and an infinite suffix with the other category—which is optimal. However, there can still be infinitely many such policies. To address this problem, we first show how to reduce the search space for approximately optimal policies with negligible additive factor to a set of $|\Pi| = O(\ln(T))$ policies. Then, we derive the parameters $\tilde{\tau}$ and η required for Algorithm 5.2. As an immediate consequence, we get a sublinear regret algorithm for this setting. We begin with defining threshold policies.

Definition 5.3: Threshold Policy

A policy π is called an (a, h)-threshold policy if there exists an number $h \in \mathbb{N} \cup \{0\}$ in π's belief-category walk such that

- π recommends category a in iterations $j \leq h$, and
- π recommends category a' in iterations $j > h$,

for $a, a' \in \{1, 2\}$ and $a \neq a'$.

For instance, the policy π^1 that always recommends Category 1 is the $(2, 0)$-threshold policy, as it recommends Category 2 until the zero'th iteration (i.e., never recommends Category 2) and then Category 1 eternally. Furthermore, the policy $\pi^{2:\lfloor N^* \rfloor}$ introduced in Theorem 5.4 is the $(2, \lfloor N^* \rfloor)$-threshold policy.

Next, recall that the chance of departure in every iteration is greater or equal to ϵ, since we assume $\max_{a,\tau} \mathbf{P}_{a,\tau} \leq 1 - \epsilon$. Consequently, the probability that a user will stay beyond H iterations is exponentially decreasing with H. We could use high-probability arguments to claim that it suffices to focus on the first H iterations, but without further insights this would yield $\Omega(2^H)$ candidates for the optimal policy. Instead, we exploit our insights about threshold policies.

Let Π_H be the set of all (a, h)-threshold policies for $a \in \{1, 2\}$ and $h \in [H] \cup \{0\}$. Clearly,

$|\Pi_H| = 2H + 2$. Lemma 5.4 shows that the return obtained by the best policy in Π_H is not worse than that of the optimal policy π^* by a negligible factor.

Lemma 5.4

For every $H \in \mathbb{N}$, it holds that

$$\mathbb{E}\left[V^{\pi^*} - \max_{\pi \in \Pi_H} V^{\pi}\right] \leq \frac{1}{2^{O(H)}}.$$

Before we describe how to apply Algorithm 5.2, we need to show that returns of all the policies in Π_H are sub-exponential. In Lemma 5.5, we show that V^{π} is (τ^2, b)-sub-exponential for every threshold policy $\pi \in \Pi_H$, and provide bounds for both τ and b^2/τ^2.

Lemma 5.5

Let $\tilde{\tau} = \frac{8e}{\ln(\frac{1}{1-\epsilon})}$ and $\eta = 1$. For every threshold policy $\pi \in \Pi_H$, the centred random variable $V^{\pi} - \mathbb{E}[V^{\pi}]$ is (τ^2, b)-sub-exponential with (τ^2, b) satisfying $\tilde{\tau} \geq \tau$ and $\eta \geq b^2/\tau^2$.

We are ready to wrap up our solution for the learning task proposed in this section. Let $H = \Theta(\ln T)$, Π_H be the set of threshold policies characterized before, and let $\tilde{\tau}$ and η be constants as defined in Lemma 5.5.

Theorem 5.6

Applying Algorithm 5.2 with $\Pi_H, T, \tilde{\tau}, \eta$ on the class of two-types two-categories instances considered in this section always yields an expected regret of

$$\mathbb{E}[R_T] \leq O(\sqrt{T} \ln T).$$

Proof. It holds that

$$\begin{aligned}
\mathbb{E}[R_T] &= \mathbb{E}\left[TV^{\pi^*} - \sum_{t=1}^{T} V^{\pi_t}\right] \\
&= \mathbb{E}\left[TV^{\pi^*} - \max_{\pi \in \Pi_H} TV^{\pi}\right] + \mathbb{E}\left[\max_{\pi \in \Pi_H} TV^{\pi} - \sum_{t=1}^{T} V^{\pi_t}\right] \\
&\leq \frac{T}{2^{O(H)}} + O(\sqrt{HT \log T} + H \log T) = O(\sqrt{T} \ln T),
\end{aligned}$$

where the inequality follows from Theorem 5.1 and Lemma 5.4. Finally, setting $H = \Theta(\ln T)$ yields the desired result. □

5.6 Discussion

This paper introduces a MAB model in which the recommender system influences both the rewards accrued *and* the length of interaction. We dealt with two classes of problems: A single user type with general departure probabilities (Section 5.4) and the two user types, two categories where each user departs after her first no-click (Section 5.5). For each problem class, we started with analyzing the planning task, then characterized a small set of candidates for the optimal policy, and then applied Algorithm 5.2 to achieve sublinear regret.

In the full version (**Ben-Porat22**), we also consider a third class of problems: Two categories, multiple user types ($M \geq 2$) where user departs with their first no-click. We use the closed-form expected return derived in Theorem 5.3 to show how to use dynamic programming to find approximately optimal planning policies. We formulate the problem of finding an optimal policy for a finite horizon H in a recursive manner. Particularly, we show how to find a $1/2^{O(H)}$ additive approximation in run-time of $O(H^2)$. Unfortunately, this approach cannot assist us in the learning task. Dynamic programming relies on skipping sub-optimal solutions to sub-problems (shorter horizons in our case), but this happens on the fly; thus, we cannot a-priori define a small set of candidates like what Algorithm 5.2 requires. More broadly, we could use this dynamic programming approach for more than two categories, namely for $K \geq 2$, but then the run-time becomes $O(H^K)$.

There are several interesting future directions. First, achieving low regret for the setup in Section 5.5 with $K \geq 2$. We suspect that this class of problems could enjoy a solution similar to ours, where candidates for optimal policies are mixing two categories solely. Second, achieving low regret for the setup in Section 5.5 with uncertain departure (i.e., $\mathcal{L} \neq 1$). Our approach fails in such a case since we cannot use belief-category walks; these are no longer deterministic. Consequently, the closed-form formula is much more complex and optimal planning becomes more intricate. These two challenges are left open for future work.

Chapter 6

Supervised Learning with General Risk Functionals

6.1 Introduction

To date, the vast majority of supervised, unsupervised, and reinforcement learning research has focused on objectives expressible as expectations (over some dataset or distribution) of an underlying loss (or reward) function. This focus is understandable. The expected loss is mathematically convenient and a reasonable default, and a special case of nearly every proposed family of risks. To be sure, this focus has paid off: we now possess a rich body of theory and methods for evaluating, optimizing, and providing theoretical guarantees on the expected loss.

However, real-world concerns such as risk aversion, equitable allocations of benefits and harms, or alignment with human preferences, often demand that we address other functionals of the loss distribution. For example, in finance, the expectation of returns must be weighed against their variance to determine an ideal portfolio allocation, as codified, e.g., in the mean-variance objective (Björk et al., 2014). Focusing on supervised learning, consider the common scenario in which a population contains a minority (constituting fraction α of the population) but where group membership was not recorded in the available data. If the pattern relating the features to the label were different for different demographics, a naively trained model might adversely harm members of a minority group. Absent further information, one sensible strategy could be to optimize the worst case performance over all subsets (of size up to α). This would translate to the familiar Conditional Value at Risk (CVaR) objective (Rockafellar, Uryasev, et al., 2000). In addition, even in settings where a model is evaluated in terms of the expected loss at test time, the training objective may be chosen as other functionals due to reasons including distribution shifts (Duchi and Namkoong, 2018), noisy labels (Lee et al., 2020), imbalanced dataset (Li et al., 2020), etc.

Risk-sensitive learning research addresses the problem of learning models under many families of (risk) functionals, including (among others) distortion risks (Wirch and Hardy, 2001), coherent risks (Artzner et al., 1999), spectral risks (Acerbi, 2002), and cumulative prospect theory

risks (Prashanth et al., 2016). Subsuming these risks under a common framework addressing bounded losses/rewards, Huang et al. (2021) recently introduced Lipschitz risk functionals, for which differences in the risk are bounded by (sup norm) differences in the Cumulative Distribution Function (CDF) of losses. Thus, because a single CDF estimate can be used to estimate all Lipschitz risks, sup norm concentration of the CDF estimate entails corresponding (simultaneous) concentration of all Lipschitz risks calculated on that CDF estimate. However, this concentration result applies only to a single hypothesis. In contrast, most uniform convergence results in learning theory have concentrated largely on the expected loss (Vapnik, 1999; Vapnik, 2013; Bartlett and Mendelson, 2002). While uniform convergence results are known for several specific risk classes, including the spectral/rank-weighted risks (Khim et al., 2020) and optimized certainty equivalent risks (Lee et al., 2020), no results to date provide uniform convergence guarantees that hold simultaneously over both a hypothesis class and a broad class of risks.

Tackling this problem, we present, to our knowledge, the first uniform convergence guarantee on estimation of the loss CDF. Our bounds rely on appropriate complexity measures of the hypothesis class. In addition to the traditional Rademacher complexity and VC dimension, we propose a new notion of permutation complexity that is especially suited to CDF estimation.

For general risk estimation, we adopt the broader class (subsuming the Lipschitz risks) of Hölder risk functionals, for which closeness in distribution entails closeness in risk. Combined with our uniform convergence guarantees for CDF estimation, this property allows us to establish uniform convergence guarantees for risk estimation of supervised learning models, which hold simultaneously over all hypotheses in the model class and over all Hölder risks.

These results license us to optimize general risks, assuring that for appropriate model classes and given sufficient data, the empirical risk minimizer will indeed generalize and that whichever objective is optimized, all Hölder risk estimates will be close to their true values. Generalization aside, optimizing complex risks is non-trivial. To tackle this problem, we propose a new algorithm for optimizing distortion risks, a subset of Hölder risks that subsumes the spectral risks, including the expectation, CVaR, cumulative prospect theory risks, and others. Our approach extends traditional gradient-based empirical risk minimization methods to handle distortion risks (Denneberg, 1990; Wang, 1996). In particular, we calculate the empirical distortion risk by re-weighting losses based on CDF values and establish convergence guarantees for the proposed optimization method. Finally, we experimentally validate our algorithm, both in settings where uniform convergence results hold and in high-dimensional settings with deep networks.

In summary, we contribute the following:

1. The first uniform convergence result for CDF estimation together with corresponding new complexity measures suited to the task (Section 6.4).
2. Uniform convergence for risk estimation that holds simultaneously for all Hölder risks, yielding learning guarantees for empirical risk minimization for the broad class of distortion risks (Section 6.5).
3. A gradient-based method for minimizing distortion risks that re-weights examples dynamically based on the empirical CDF of losses, and corresponding convergence guarantees (Section 6.6).

4. Experiments confirming the practical usefulness of our learning algorithm (Section 6.7).

6.2 Related Literature

Risk functionals have long been studied in diverse contexts (Sharpe, 1966; Artzner et al., 1999; Rockafellar, Uryasev, et al., 2000; Krokhmal, 2007; Shapiro et al., 2014; Acerbi, 2002; Prashanth et al., 2016; Jie et al., 2018). CVaR, value-at-risk, and mean-variance (Cassel et al., 2018; Sani et al., 2013; Vakili and Zhao, 2015; Zimin et al., 2014) rank among the most widely studied risks. Prashanth et al. (2016) introduces the cumulative prospect theory risks, which have been studied in bandit (Gopalan et al., 2017) and supervised learning settings (Leqi et al., 2019b). Many previous works have tackled the evaluation (Huang et al., 2021; Chandak et al., 2021) and optimization (Torossian et al., 2019; Munos, 2014) of risk functionals.

Recent work on risk-sensitive supervised learning has established the uniform convergence of a single risk functional when losses incurred by the models are bounded (Khim et al., 2020; Lee et al., 2020), or the excess risk of a particular learning procedure in cases where the loss could be unbounded (Holland and Haress, 2021). Collectively, these works have addressed the class of spectral risks (L-risks or rank-weighted risks) that includes the expected value, CVaR and cumulative prospect theory risks (Khim et al., 2020; Holland and Haress, 2021), as well as the class of optimized certainty equivalent risks that includes the expected value, CVaR and entropic risks (Ben-Tal and Teboulle, 1986; Lee et al., 2020) .

To our knowledge, the aforementioned risk-sensitive learning results are considerably narrower: the analyses apply only to smaller *families* of risks and the guarantees hold only for a *single risk functional* (not simultaneously over the family). By contrast, we establish uniform convergence results that hold simultaneously over both a broader class of risks and over an entire model class (constrained by an appropriate complexity measure). The key to our approach is to estimate the CDF of losses and control its sup norm error uniformly over a hypothesis class. CDF estimation is a central topic in learning theory (Devroye et al., 2013). Strong approximation results provide concentration bounds on the Kolmogorov–Smirnov distance (sup norm) between the true and estimated CDF (Massart, 1990). As our uniform convergence results are over a hypothesis class of possibly infinite number of hypotheses, we control the complexity of the hypothesis class using data-dependent complexity notions (e.g., Rademacher complexity) and data-independent complexity notions (e.g., VC dimension) (Alexander, 1984; Vapnik, 2006; Gänssler and Stute, 1979).

6.3 Preliminaries

We use $\mathcal{X}$ to denote the space of covariates, $\mathcal{Y}$ the space of labels, and $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$. Let $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ denote a loss function and $\mathbb{F}$ a hypothesis class, where for any $f \in \mathbb{F}$, $f : \mathcal{X} \to \mathcal{Y}$. The set $\mathbb{F}_\ell$, with elements ℓ_f, denotes the class of functions that are compositions of the loss function ℓ and a hypothesis $f \in \mathbb{F}$, i.e., $\forall z \in \mathcal{Z}, \ell_f(z) = \ell\left(f(x), y\right)$. Furthermore, we use $\ell_f\left(Z\right)$

to denote the random variable of the loss incurred by $f \in \mathbb{F}$ under data $Z = (X, Y)$. For any $n \in \mathbb{N}$, $[n] := \{1, \ldots, n\}$.

We use $\mathcal{U}$ to denote the space of real-valued random variables that admit CDFs. For any $U \in \mathcal{U}$, its CDF is denoted by F_U. A risk functional $\rho : \mathcal{U} \to \mathbb{R}$ is a mapping from a space of real-valued random variables to reals. A risk functional is called law-invariant (or version-independent) if for any pair of random variables $U, U' \in \mathcal{U}$ with the same law ($F_U = F_{U'}$), we have $\rho(U) = \rho(U')$ (Kusuoka, 2001). We work with law-invariant risk functionals in this paper, and with some abuse of notation, we refer to $\rho(F_U)$ and $\rho(U)$ interchangeably.

6.4 Uniform Convergence for CDF Estimation

We begin with an important building block for risk estimation—CDF estimation with uniform convergence guarantees. Given a loss function ℓ and a data set of n labeled data points $\{Z_i\}_{i=1}^n$ where $Z_i = (X_i, Y_i)$, we are interested in estimating the CDF of $\ell_f(Z)$ for all $f \in \mathbb{F}$. We use the unbiased empirical CDF estimator:

$$\widehat{F}(r; f) := \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{\{\ell_f(Z_i) \le r\}}, \tag{6.1}$$

where $\mathbb{E}[\widehat{F}(r; f)] = \mathbb{P}(\ell_f(Z) \le r) = F(r; f)$. To establish the uniform convergence of the estimator, our central goal is to analyze the following quantity:

$$e_n(\mathbb{F}, \ell) = \sup_{f \in \mathbb{F}} \sup_{r \in \mathbb{R}} \left| \widehat{F}(r; f) - F(r; f) \right|. \tag{6.2}$$

In Section 6.4.1, we exploit the special structure of CDF estimation and propose a new notion of permutation complexity that captures the complexity of the hypothesis class used for CDF estimation. In Section 6.4.2, we apply the more classical approach for analyzing uniform convergence that does not exploit any special structure of CDF estimation. Each approach offers a unique perspective and contributes to our understanding of CDF estimation. We highlight that the uniform convergence we provide hold for *any* loss distribution regardless of whether the loss is binary or bounded.

We first introduce notation key to our analysis. The Rademacher complexity in our setting (for a given loss function ℓ) is given as follows:

$$\begin{aligned} \mathcal{R}(n, \mathbb{F}) &= \mathbb{E}_{\mathbb{P},\mathbb{R}} \left[\sup_{f \in \mathbb{F}} \sup_{r \in \mathbb{R}} \frac{1}{n} \left| \sum_{i=1}^{n} \xi_i \mathbb{1}_{\{\ell_f(Z_i)) \le r\}} \right| \right] \\ &= \mathbb{E}_{\mathbb{P},\mathbb{R}} \left[\sup_{f \in \mathbb{F}} \sup_{g \in \mathbb{G}(1)} \frac{1}{n} \left| \sum_{i=1}^{n} \xi_i g(\ell_f(Z_i)) \right| \right], \end{aligned} \tag{6.3}$$

with $\mathbb{R}$ being a Rademacher measure on a set of Rademacher random variables $\{\xi_i\}_{i=1}^n$ and $\mathbb{G}(1) := \{\mathbb{1}_{\{\cdot \le r\}} : \forall r \in \mathbb{R}\}$ is the set of indicator functions parameterized by a real-valued r.

Using McDiarmid's inequality and symmetrization, we obtain the following classical result that bounds $e_n(\mathbb{F}, \ell)$ in terms of the Rademacher complexity. All proofs in this section can be found in Appendix E.2.

Theorem 6.1

Given a hypothesis class $\mathbb{F}$, any loss function $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$, and n samples $\{Z_i\}_{i=1}^n$, we have that with probability at least $1 - \delta$,

$$e_n(\mathbb{F}, \ell) \leq 2\mathcal{R}(n, \mathbb{F}) + \sqrt{\frac{\log(\frac{1}{\delta})}{2n}}.$$

In general, the Rademacher complexity $\mathcal{R}(n, \mathbb{F})$ is hard to obtain. Researchers have come up with different ways to control it for various hypothesis classes, e.g., hypothesis classes with finite VC dimension (Wainwright, 2019). In the following, we discuss how we work with $\mathcal{R}(n, \mathbb{F})$.

6.4.1 Permutation Complexity

We first notice that $\mathcal{R}(n, \mathbb{F})$ depends *jointly* on both the hypothesis class $\mathbb{F}$ and $\mathbb{G}(1)$. A direct approach that follows from the classical statistical learning theory is to work with the function class that combines $\mathbb{F}$ and $\mathbb{G}(1)$, which we provide more details in Section 6.4.2. In this section, we propose a new way of thinking about $\mathcal{R}(n, \mathbb{F})$. By exploiting the special structure of CDF estimation (the structure of $\mathbb{G}(1)$), we uncover that $\mathcal{R}(n, \mathbb{F})$ can be controlled by *only* the complexity of the hypothesis class $\mathbb{F}$ (or $\mathbb{F}_\ell$ with elements ℓ_f). In order to do so, we first introduce the notion of *permutation complexity*. This complexity measure is data-dependent and enables us to work with $\mathcal{R}(n, \mathbb{F})$ by disentangling the complexity of $\mathbb{F}$ (or $\mathbb{F}_\ell$) from that of $\mathbb{G}(1)$.

For a measurable space $\mathcal{V}$, let $\zeta : \mathcal{V} \to \mathbb{R}$ denote a measurable function and $\{v_i\}_{i=1}^n$ denote a set of n points in $\mathcal{V}$. Satisfying the conditions of selection and maximum theorems (Guide, 2006, Chapter 17), a permutation function $\pi : [n] \to [n]$ in the space $\Pi(n)$ of all permutation of size n exists, and permutes the indices of $\{v_i\}_{i=1}^n$ such that $\zeta(v_{\pi(1)}) \leq \zeta(v_{\pi(2)}) \leq \ldots \leq \zeta(v_{\pi(n)})$. We note that the permutation function can depend on the specific data points $\{v_i\}_{i=1}^n$ and the function ζ of interests. In the following definitions, we consider a function class $\mathbb{J}$ of functions $\zeta : \mathcal{V} \to \mathbb{R}$ and a probability measure μ on $(\mathcal{V}, \sigma(\mathcal{V}))$, where $\sigma(\mathcal{V})$ denotes the σ-algebra generated by $\mathcal{V}$.

Definition 6.1: Permutation Complexity

The instance-dependent permutation complexity of $\mathbb{J}$ at n data points $\{v_i\}_{i=1}^n$, denoted as $\mathcal{N}_\Pi(\mathbb{J}, \{v_i\}_{i=1}^n)$, is the minimum number of permutation functions $\pi \in \Pi(n)$ needed to sort elements of $\{\zeta(v_i)\}_{i=1}^n$, $\forall \zeta \in \mathbb{J}$. The permutation complexity of $\mathbb{J}$ at n (random) data points $\{V_i\}_{i=1}^n$ with measure μ is

$$\mathcal{N}_\Pi(n, \mathbb{J}, \mu) := \mathbb{E}_\mu \left[\mathcal{N}_\Pi(\mathbb{J}, \{V_i\}_{i=1}^n)\right].$$

To obtain a better understanding of permutation complexity, we provide the permutation complexity of monotone real-valued functions.

Lemma 6.1

When $\mathcal{V}$ is the space of reals, $\mathbb{J}$ is a set of real-valued non-decreasing functions on $\mathcal{V}$, the permutation complexity $\mathcal{N}_\Pi(n, \mathbb{J}, \mu) = 1$.

Proof. For a set of real-valued points $\{v_i\}_{i=1}^n$, we can construct a permutation function π such that $v_{\pi(1)} \leq v_{\pi(2)} \leq \ldots \leq v_{\pi(n)}$. Since all functions in $\mathbb{J}$ are non-decreasing, for any $\zeta \in \mathbb{J}$, we have $\{v_i\}_{i=1}^n$ such that $\zeta(v_{\pi(1)}) \leq \zeta(v_{\pi(2)}) \leq \ldots \leq \zeta(v_{\pi(n)})$. Thus, $\mathcal{N}_\Pi(\mathbb{J}, \{v_i\}_{i=1}^n) = 1$ and $\mathcal{N}_\Pi(n, \mathbb{J}, \mu) = 1$ for any measure μ. □

Remark 6.1. *This result implies that the permutation complexity of threshold functions $\mathbb{G}(1)$ is one.*

Mapping the definition to our setting, the function class $\mathbb{J}$ of interest is $\mathbb{F}_\ell$ with elements ℓ_f. The data points $V_i = Z_i = (X_i, Y_i)$ and the measure $\mu = \mathbb{P}$ (the probability measure for Z).

An immediate observation is that when $|\mathbb{F}|$ is finite, we only need at most $|\mathbb{F}|$ permutation functions to sort $\{\ell_f(z_i)\}_{i=1}^n$ (one permutation function for each $f \in \mathbb{F}$), i.e., $\mathcal{N}_\Pi(\mathbb{F}_\ell, \{z_i\}_{i=1}^n) \leq |\mathbb{F}|$. For the special case of binary classification, where $\mathcal{Y} = \{0, 1\}$ and the loss function ℓ is the $0/1$ loss, a coarse upper bound on the permutation complexity when $\mathbb{F}$ has finite VC dimension $\nu(\mathbb{F})$ is $\mathcal{N}_\Pi(n, \mathbb{F}_\ell, \mathbb{P}) \leq (n+1)^{\nu(\mathbb{F})}$. This is due to Sauer's Lemma: there are at most $(n+1)^{\nu(\mathbb{F})}$ ways of labeling the data, which suggests that we need at most $(n+1)^{\nu(\mathbb{F})}$ permutation functions. However, as one may have noticed, the number of permutation functions needed may be (much) smaller than this number. For example, consider a binary classification setting where we have 3 data points and $\mathbb{F}$ is large enough such that all 2^3 possible losses $(000, 001, \ldots)$ can be incurred. In such a case, we only need 4 permutation functions, since loss sequences that are non-decreasing (or non-increasing) can share the same permutation function, e.g., for loss sequences $111, 011, 001, 000$, we can use the same permutation function $\pi(i) = i, \forall i \in [3]$. We note that the permutation complexity is defined for not just binary-valued function classes. Precisely characterizing the permutation complexity for different combinations of function classes, data distributions and loss functions is of future interest.

Theorem 6.2

For any hypothesis class $\mathbb{F}$ and loss function $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$, we have that

$$\mathcal{R}(n, \mathbb{F}) \leq \sqrt{\frac{\log(4\mathcal{N}_\Pi(n, \mathbb{F}_\ell, \mathbb{P}))}{2n}}.$$

Theorem 6.2 indicates that, despite $\mathcal{R}(n, \mathbb{F})$ depending on the supremum over both the function class $\mathbb{F}$ and $\mathbb{G}(1)$ where $\mathbb{G}(1)$ is an infinite set, $\mathcal{R}(n, \mathbb{F})$ can be controlled by just the complexity of $\mathbb{F}_\ell$. When the permutation complexity $\mathcal{N}_\Pi(n, \mathbb{F}_\ell, \mathbb{P})$ is polynomial in the number of samples n, we obtain that $\mathcal{R}(n, \mathbb{F}) = O(\sqrt{\log(n)/n})$. The immediate consequence of Theorem 6.2 when the hypothesis class $\mathbb{F}$ is finite is provided below.

Corollary 6.1. *For a finite hypothesis class $\mathbb{F}$,*

$$\mathcal{R}(n, \mathbb{F}) \leq \sqrt{\frac{\log(4|\mathbb{F}|)}{2n}}.$$

Corollary 6.1 suggests that the generalization bound of CDF estimation follows the same rate as classical generalization bound of the expected loss for finite function classes, i.e., $O(\sqrt{\log(|\mathbb{F}|)/n})$.

6.4.2 Classical Approach

In this section, we present a more classical approach for analyzing the uniform convergence without exploiting the specific structure of CDF estimation. As we have noted before, $\mathcal{R}(n, \mathbb{F})$ depends on both $\mathbb{F}$ and $\mathbb{G}(1)$. In this approach, we directly work with the function class that combines $\mathbb{F}$ and $\mathbb{G}(1)$: for a given loss function ℓ, we define

$$\mathbb{H} := \{h : \mathcal{Z} \to \{0, 1\} : \\ h(z; r) = \mathbb{1}_{\{\ell_f(z) \leq r\}}, f \in \mathbb{F}, r \in \mathbb{R}\}.$$

We note that even when $\mathbb{F}$ is finite, $\mathbb{H}$ is an infinite set. The Rademacher complexity (6.3) can be re-written as

$$\mathcal{R}(n, \mathbb{F}) = \mathbb{E}_{\mathbb{P},\mathbb{R}} \left[\sup_{h \in \mathbb{H}} \frac{1}{n} \left| \sum_{i=1}^{n} \xi_i h(Z_i) \right| \right].$$

Lemma 6.2

(Wainwright, 2019, Lemma 4.14) Let $|\mathbb{H}(\{z_i\}_{i=1}^n)|$ denote the maximum cardinality of the set $\mathbb{H}(\{z_i\}_{i=1}^n) = \{(h(z_1), \ldots, h(z_n)) : h \in \mathbb{H}\}$, where $n \in \mathbb{N}$ is fixed and $\{z_i\}_{i=1}^n$ can be any data collection for $z_i \in \mathcal{Z}$. Then, we have

$$\mathcal{R}(n, \mathbb{F}) \leq 2\sqrt{\frac{\log(|\mathbb{H}(\{z_i\}_{i=1}^n)|)}{n}}.$$

Since $\mathbb{H}$ consists of binary functions, for any data collection $\{z_i\}_{i=1}^n$, the set $\mathbb{H}(\{z_i\}_{i=1}^n)$ is finite. When $\mathbb{F}$ is finite, we obtain that $|\mathbb{H}(\{z_i\}_{i=1}^n)| \leq (n+1)|\mathbb{F}|$. This is true since for a given data collection $\{z_i\}_{i=1}^n$ and hypothesis $f \in \mathbb{F}$, after sorting $\{\ell_f(z_i)\}_{i=1}^n$, we have at most $(n+1)$ ways of labeling them using the indicator functions $\mathbb{1}_{\{\cdot \leq r\}}$ for $r \in \mathbb{R}$. Thus, if we directly apply Lemma 6.2, we obtain that the Rademacher complexity $\mathcal{R}(n, \mathbb{F})$ is on the order of $O(\sqrt{(\log|\mathbb{F}| + \log n)/n})$, which has an extra $\log(n)$ term in the numerator compared to our bound in Corollary 6.1. In general, when $\mathbb{H}$ has finite VC dimension $\nu(\mathbb{H})$, we obtain that $\mathcal{R}(n, \mathbb{F}) \leq O(\sqrt{\nu(\mathbb{H}) \log(n+1)/n})$. However, this result is far from being sharp. Using more advanced techniques, e.g., chaining and Dudley's entropy integral (Wainwright, 2019), one can remove the extra $\log(n)$ factor on the numerator, and obtain that $\mathcal{R}(n, \mathbb{F}) \leq O(\sqrt{\nu(\mathbb{H})/n})$ (for more details, see Wainwright (2019, Example 5.24)). As a consequence, when $|\mathbb{F}| < \infty$, we have $\nu(\mathbb{H}) \leq \log(|\mathbb{F}|)$ and thus obtain that $\mathcal{R}(n, \mathbb{F}) \leq O(\sqrt{\log(|\mathbb{F}|)/n})$ which is at the same rate as our bound in Corollary 6.1.

6.5 Uniform Convergence for Hölder Risk Estimation

Using our results in Section 6.4, we show uniform convergence for a broad class of risk functionals—Hölder risks. As illustrated in Section 6.6, uniform convergence for a single risk functional provides grounding for learning models through minimizing the empirical risk. In addition to uniform convergence for a single risk, we also provide uniform convergence results hold for a collection of risks simultaneously. The second result is important due to the following reasons: Although models are trained to optimize a single risk objective, evaluating their performance under multiple risks can give a holistic assessment of their behavior—a task we call *risk assessment*. For example, models minimizing CVaR at different α levels may have different tradeoffs with their expected loss, and monitoring the progress of both objectives throughout the training process can inform choice of the best model. In addition, when given a set of models obtained under different learning mechanisms, one may want to compare them in terms of different risks. To this end, using our results on uniform convergence for CDF estimation (Section 6.4), we demonstrate how a collection of models may be assessed under many risks simultaneously, with estimation errors of the same order as the CDF estimation error.

6.5.1 Hölder Risk Functionals

We begin by introducing a new class of risks—the Hölder risk functionals—that includes many popularly studied risks and generalizes the notion of Lipschitz risk functionals (Huang et al., 2021) and Hölder continuous functionals in Wasserstein distance (Bhat and LA, 2019).

Definition 6.2

Let d denote a quasi-metric[a] on the space of CDFs. A risk functional ρ is $L(\rho, p, d)$ Hölder on a space of real-valued random variables $\mathcal{U}$ if there exist constants $p > 0$ and $L(\rho, p, d) > 0$ such that for all $U, U' \in \mathcal{U}$ with CDF F_U and $F_{U'}$ respectively, the following holds:

$$|\rho(F_U) - \rho(F_{U'})| \leq L(\rho, p, d) d(F_U, F_{U'})^p.$$

[a] Quasi-metrics are defined in Appendix E.1.

The class of Hölder risk functionals subsumes many other risk functional classes. In particular, Lipschitz risk functionals (Huang et al., 2021) are $L(\cdot, 1, \mathcal{L}_\infty)$ Hölder on bounded random variables. As a direct result, distortion risk functionals with Lipschitz distortion functions (Denneberg, 1990; Wang, 1996; Wang et al., 1997; Balbás et al., 2009; Wirch and Hardy, 1999; Wirch and Hardy, 2001), cumulative prospect theory risks (Prashanth et al., 2016; Leqi et al., 2019b), variance, and linear combinations of aforementioned functionals are all Hölder on bounded random variables. In addition, the optimized certainty equivalent risks (Lee et al., 2020) and spectral risks (Khim et al., 2020; Holland and Haress, 2021) recently studied in risk-sensitive supervised learning literatures, are Lipschitz (hence Hölder) on the space of bounded random variables. We provide proofs and further details in Appendix E.1.

To be more specific, we present a subset of Hölder risk functionals—distortion risks with

Lipschitz distortion functions—that consists of many well-studied risks including the expected value, CVaR and cumulative prospect theory risks. When the loss is non-negative, the distortion risk of $\ell_f(Z)$ is defined to be:

$$\rho(F(\cdot; f)) = \int_0^\infty g(1 - F(r; f))dr, \tag{6.4}$$

where the distortion function $g : [0, 1] \to [0, 1]$ is non-decreasing with $g(0) = 0$ and $g(1) = 1$. In the case of expected value, the distortion function g is 1-Lipschitz. For CVaR at level α, the distortion function g is $\frac{1}{\alpha}$-Lipschitz. For more details, we refer the readers to Huang et al. (2021). In the following, we show uniform convergence for estimating Hölder risks using our proposed estimator (Section 6.5.2) and develop optimization procedures to minimize distortion risks (Section 6.6).

6.5.2 Uniform Convergence for Risk Estimation

For a given hypothesis $f \in \mathbb{F}$ and loss function ℓ, we estimate the risk $\rho(F(\cdot; f))$ using the CDF estimator $\widehat{F}(\cdot; f)$ by plugging it in the functional of interest: $\rho(\widehat{F}(\cdot; f))$. Many existing risk estimators in the supervised learning literatures, including the traditional empirical risk and estimators used for estimating spectral risks (Khim et al., 2020) and optimized certainty equivalent risks (Lee et al., 2020) can be viewed as examples of the above estimator.

Leveraging the uniform convergence results of the CDF estimator, we present uniform convergence result of the proposed risk estimator. The uniform convergence holds both over the hypothesis class $\mathbb{F}$ and over a set of Hölder risk functionals. As the Hölder class contains a large set of popularly studied risks, our result demonstrates that models can be assessed under many risks without loss of statistical power. This is formalized in Theorem 6.3 below.

Theorem 6.3

For a hypothesis class $\mathbb{F}$, a bounded loss function ℓ, and $\delta \in (0, 1]$, if $\mathbb{P}(e_n(\mathbb{F}, \ell) \leq \epsilon) \geq 1-\delta$, then with probability $1 - \delta$, for all $\rho \in \mathbb{T}$, we have

$$\sup_{f \in \mathbb{F}} |\rho(F(\cdot; f)) - \rho(\widehat{F}(\cdot; f))| \leq L(\rho, 1, \mathcal{L}_\infty)\epsilon,$$

where $\mathbb{T}$ is the set of $L(\cdot, 1, \mathcal{L}_\infty)$ Hölder risk functionals on the space of bounded random variables.

In cases where the CDF estimation error ϵ is of order $O(1/\sqrt{n})$, we can estimate the set of Hölder risks $\mathbb{T}$ for all hypotheses in $\mathbb{F}$ at rate $O(1/\sqrt{n})$. Because our result is uniform over both the hypothesis class $\mathbb{F}$ and the risk functional class $\mathbb{T}$, it is a generalization of existing uniform convergence results that are uniform over $\mathbb{F}$, but for a single risk functional (Khim et al., 2020; Lee et al., 2020).

In Appendix E.3, we provide similar uniform convergence results where the set of risk functionals are Hölder smooth in Wasserstein distance. As a direct consequence to Theorem 6.3, uniform convergence of a single Hölder risk, e.g., a distortion risk (6.4), is given below.

	VGG-11	GoogLeNet	ShuffleNet	Inception	ResNet-18
Accuracy	69.022%	69.772%	69.356%	69.542%	69.756%
$\mathbb{E}[\ell_f]$	1.261	1.283	1.360	1.829	1.247
$\text{CVaR}_{.05}(\ell_f)$	1.327	1.350	1.431	1.925	1.313
$\mathbb{E}[\ell_f] + 0.5\text{Var}(\ell_f)$	5.215	4.376	6.718	14.416	5.353
$\text{HRM}_{.3,.4}(\ell_f)$	1.374	1.336	1.542	2.214	1.382
$\text{HRM}_{.2,.8}(\ell_f)$	1.233	1.239	1.344	1.845	1.225

Table 6.1: Risks for different ImageNet classification models evaluated on the validation set. $\ell_f(Z)$ is the cross-entropy loss for each model f. For simplicity, we omitted the arguments Z in the table. CVaR_α is the expected value of the top 100α percent losses. $\text{HRM}_{a,b}$ is the cumulative prospect theory risk defined in Leqi et al. (2019b). All results are rounded to 3 digits.

Corollary 6.2. *For a hypothesis class $\mathbb{F}$, a bounded loss function $\ell : \mathcal{Y} \times \mathcal{Y} \to [0, D]$, and $\delta \in (0, 1]$, if $\mathbb{P}(e_n(\mathbb{F}, \ell) \leq \epsilon) \geq 1 - \delta$, then with probability $1 - \delta$, we have*

$$\sup_{f \in \mathbb{F}} |\rho(F(\cdot; f)) - \rho(\widehat{F}(\cdot; f))| \leq L\epsilon,$$

where ρ is a distortion risk with $\frac{L}{D}$-Lipschitz distortion function.

Remark 6.2. *As an example, consider a binary classification setting where the loss function ℓ is the $0/1$ loss and the hypothesis class $\mathbb{F}$ is finite, using our results in Corollary 6.2, we obtain that the the generalization error for the expected value is $O(\sqrt{\log(|\mathbb{F}|)/n})$ and the generalization error for the CVaR is $O(\sqrt{\log(|\mathbb{F}|)/\alpha^2 n})$, which are of the same rates (with better constants) as the ones in Lee et al. (2020).*

6.6 Empirical Risk Minimization

For a single risk functional, one may want to learn models that optimize it. Our uniform convergence results for risk estimation license us to learn models that minimize the population risk through Empirical Risk Minimization (ERM). We denote the population and empirical risk minimizers as follows:

$$f_\star \in \arg\min_{f \in \mathbb{F}} \rho(F(\cdot; f)), \;\; \widehat{f}_\star \in \arg\min_{f \in \mathbb{F}} \rho(\widehat{F}(\cdot; f)). \tag{6.5}$$

The excess risk of the empirical risk minimizer $\widehat{f}_\star$ can be bounded by

$$\begin{aligned}
&\rho(F(\cdot; \widehat{f}_\star)) - \rho(F(\cdot; f_\star)) = \rho(F(\cdot; \widehat{f}_\star)) - \rho(\widehat{F}(\cdot; \widehat{f}_\star)) \\
&\quad + \rho(\widehat{F}(\cdot; \widehat{f}_\star)) - \rho(\widehat{F}(\cdot; f_\star)) + \rho(\widehat{F}(\cdot; f_\star)) - \rho(F(\cdot; f_\star)) \\
&\leq 2 \sup_{f \in \mathbb{F}} |\rho(\widehat{F}(\cdot; f)) - \rho(F(\cdot; f))|.
\end{aligned}$$

We study ERM when the loss function is non-negative and the risk functional of interest is a distortion risk with a Lipschitz distortion function (6.4). Such distortion risk functionals consist many well-studied risks, including the expected value, CVaR, cumulative prospect theory risks, and other spectral risks (Bäuerle and Glauner, 2021). Using Corollary 6.2, we obtain that when $e_n(\mathbb{F}, \ell) = O(1/\sqrt{n})$ and the loss ℓ is bounded, the excess risk of $\widehat{f}_\star$ is $O(1/\sqrt{n})$.

We consider settings where the hypothesis class $\mathbb{F}$ is a class of parameterized functions, e.g., linear models and neural networks and use $\Theta \subseteq \mathbb{R}^d$ to denote the set of parameters. For a hypothesis $f \in \mathbb{F}$ parameterized by $\theta \in \Theta$, we denote $F(\cdot; f)$ and $\ell_f(z_i)$ by F_θ and $\ell_\theta(i)$, respectively. Similarly, we use $\theta_\star$ and $\widehat{\theta}_\star$ for referring to $f_\star$ and $\widehat{f}_\star$. As in Section 6.4.1, we use $\pi_\theta : [n] \to [n]$ to denote the permutation function such that $\ell_\theta(\pi_\theta(i))$ is the i-th smallest loss under the current model θ and the fixed dataset $\{z_i\}_{i=1}^n$. Using the CDF estimator $\widehat{F}_\theta$ (6.1), the *empirical distortion risk* $\rho(\widehat{F}_\theta)$ can be re-written as

$$\sum_{i=1}^{n} g\left(1 - \frac{i-1}{n}\right) \cdot \left(\ell_\theta(\pi_\theta(i)) - \ell_\theta(\pi_\theta(i-1))\right),$$

where for all $\theta \in \Theta$, we set $\ell_\theta(\pi_\theta(0)) := 0$ since the losses are non-negative.

To employ first-order methods for minimizing the empirical distortion risk, it is natural to first identify when $\rho(\widehat{F}_\theta)$ is differentiable.

Lemma 6.3

If $\{\ell_\theta(z_i)\}_{i=1}^n$ are Lipschitz continuous in $\theta \in \Theta$, then for all $i \in [n]$, $\ell_\theta(\pi_\theta(i))$, i.e., the i-th smallest loss, is Lipschitz continuous in θ and $\rho(\widehat{F}_\theta)$ is differentiable in θ almost everywhere.

When $\rho(\widehat{F}_\theta)$ (and $\ell_\theta(\pi_\theta(i))$) is differentiable, the gradient $\nabla_\theta \rho(\widehat{F}_\theta)$ can be written as

$$\sum_{i=1}^{n} \left(g\left(1 - \frac{i-1}{n}\right) - g\left(1 - \frac{i}{n}\right)\right) \cdot \nabla_\theta \ell_\theta(\pi_\theta(i)). \tag{6.6}$$

When $g(x) = x$, the distortion risk is the same as the expected loss and we recover the gradient for the traditional empirical risk.

To avoid the non-differentiable points, we add a small noise to the gradient descent steps. By doing so, we ensure that the descent steps will end up in differentiable points almost surely. Choose initial point $\theta_1 \in \Theta$. At iteration t, the parameter is updated as follows

$$\theta_{t+1} \leftarrow \theta_t - \eta \left(\nabla_\theta \rho(\widehat{F}_\theta) + w_t\right), \tag{6.7}$$

where η is the learning rate, $\nabla_\theta \rho(\widehat{F}_\theta)$ is given in (6.6) and w_t is sampled from a d-dimensional Gaussian with mean 0 and variance $\frac{1}{d}$.

In general, even when the loss function ℓ is convex in the parameter θ, the empirical distortion risk $\rho(\widehat{F}_\theta)$ may not be convex in θ. In Corollary 6.3, we show local convergence of θ_t obtained through following (6.7).

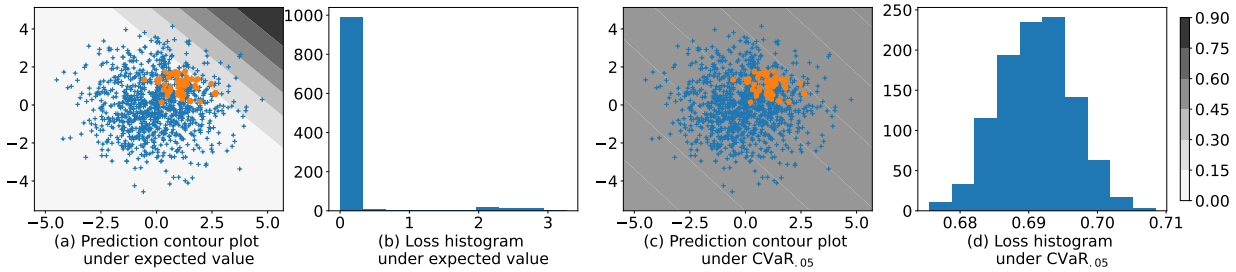
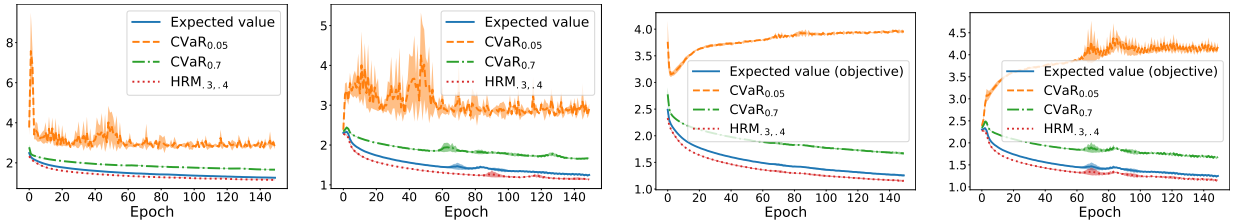


Figure 6.1: Prediction (predicted probability of a covariate being labeled as 1) contours and loss histograms of two models learned under the expected loss and the $\text{CVaR}_{.05}$ objective, respectively. The blue pluses and orange dots represent two classes. The loss distribution for the expected loss model has extremely high values for a small subset of the covariates.



(a) Training objective values

(b) Testing objective values

(c) Training risk evaluations

(d) Testing risk evaluations

Figure 6.2: Performance of VGG-16 models trained under expected loss, $\text{CVaR}_{.05}$, $\text{CVaR}_{.7}$ and $\text{HRM}_{.3,.4}$. In Figures 6.2a and 6.2b, each model is trained and evaluated on the same objective. In Figures 6.2c and 6.2d, we only train one model under the expected loss but report all four objectives of that model. All results are averaged over 5 runs.

Corollary 6.3. *If $\{\ell_\theta(z_i)\}_{i=1}^n$ are Lipschitz continuous and $\rho(\widehat{F}_\theta)$ is β-smooth in θ, then the following holds almost surely when the learning rate in (6.7) is $\eta = \frac{1}{\beta\sqrt{T}}$:*

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\|\nabla_\theta \rho(\widehat{F}_{\theta_t})\|^2\right] \leq \frac{2\beta}{\sqrt{T}}\left(\rho(\widehat{F}_{\theta_1}) - \rho(\widehat{F}_{\theta_\star}) + \frac{1}{2\beta}\right).$$

Corollary 6.3 demonstrates that the average gradient magnitude over T iterations shrinks as T goes to infinity, suggesting that performing gradient descent by following (6.7) will converge to an approximate stationary point.

6.7 Experiment

In our experiments, we demonstrate the efficacy of our proposed estimator for risk estimation and proposed learning procedure for obtaining risk-sensitive models. In Section 6.7.1, we work with a risk assessment setting where we simultaneously inspect a finite set of models in terms of multiple risks. In Section 6.7.2, we show the performance of empirical risk minimization under

various distortion risks. After showing that the classifier learned under different risk objectives behave differently in a toy example, we learn risk-sensitive models for CIFAR-10.

6.7.1 Risk Assessment on ImageNet Models

We perform risk assessments on pretrained Pytorch models for ImageNet classification. In particular, we choose models with similar accuracy (both reported on the official Pytorch website and confirmed by us) on the validation set (50,000 images) for the ImageNet classification challenge (Russakovsky et al., 2015). The models are VGG-11 (Simonyan and Zisserman, 2014), GoogLeNet (Szegedy et al., 2015), ShuffleNet (Ma et al., 2018), Inception (Szegedy et al., 2016) and ResNet-18 (He et al., 2016) and the accuracy of these models evaluated on the validation set are around 69% (Table 6.1). By assessing the risks of models with similar accuracy, we highlight how models with similar performance under traditional metrics (e.g., accuracy) could have different risk performances. For example, though Inception has similar accuracy compared to other models, its loss variance is much higher compared to others, which may be detrimental in settings where high-varying performance is not preferred. We also evaluated the CVaR of these models under different α's (Figure E.1 in Appendix E.5). Our theoretical results suggest that all these evaluations hold simultaneously across the risk functionals and models of interest with the error being $O(\sqrt{\log|\mathbb{F}|/n})$ ($|\mathbb{F}| = 5$ in this experiment). In addition to showcasing the power of our theoretical results, this example demonstrates how model assessments under multiple risk notions provide a better understanding of model behaviors.

6.7.2 Empirical Distortion Risk Minimization

To illustrate the difference among models learned under different risk objectives, we first present a toy example for comparing models learned under the expected loss objective and the CVaR objective respectively. We then show the efficacy of our proposed optimization procedure through training deep neural networks on CIFAR-10. In both cases, the models are learned by following (6.7). For more details on these experiments, we refer the readers to Appendix E.5.

Toy Example In the toy example, we work with a binary classification task where the covariates are 2-dimensional. In Figure 6.1, the blue pluses and orange dots represent two classes, respectively. We have learned logistic regression models to minimize the expected loss and the $\text{CVaR}_{.05}$ (expected value of the top 5% losses) through minimizing their empirical risks. The loss distribution along with the prediction contours of the two classifiers showcase the difference between the two models. In particular, the model learned under expected loss suffers high loss for a small subset of the covariates while the model learned under $\text{CVaR}_{.05}$ have all losses concentrated around a small value. Indicated by the (uniform) grey color in the contour plot, the predictions (predicted probability of a covariate being labeled as 1) for the $\text{CVaR}_{.05}$ model are around 0.5. In contrast, the predictions for the expected loss model spread across a wide range between 0 and 1.

CIFAR-10 We have trained VGG-16 models on CIFAR-10 through minimizing the empirical risks for expected loss, $\text{CVaR}_{.05}$, $\text{CVaR}_{.7}$ and $\text{HRM}_{.3,.4}$ (Leqi et al., 2019b) using the gradient

descent step presented in (6.7). The models are trained over 150 epochs and the learning rate is chosen to be 0.005. As shown in Figure 6.2a and 6.2b, in general, the objective values are decreasing over the epochs during training and testing. In addition, we observe that minimizing the empirical risk for expected loss does not necessarily imply minimizing other risks, e.g., $\text{CVaR}_{.05}$ (Figure 6.2c). These results suggest the efficacy of our proposed optimization procedure for minimizing distortion risks.

6.8 Discussion

We have presented a principled framework, including analytic tools and algorithms for risk-sensitive learning and assessment that: (1) obtains the empirical CDF; (2) estimates the risks of interest through plugging in the empirical CDF; and (3) minimizes the empirical risk (for risk-sensitive learning). Our theoretical results on the uniform convergence of the proposed risk estimators hold simultaneously over a hypothesis class (constrained by an appropriate complexity measure) and over Hölder risks. The key building block for these results is the uniform convergence of the CDF estimator.

There are multiple future directions of our work. First, we hope to more precisely characterize the permutation complexity (under various hypothesis classes). Second, our gradient descent procedure (6.7) requires sorting all losses. An important next step would be to allow minibatches (sorting only a small subset of losses) when minimizing empirical distortion risks. Third, as shown in Figure 6.1, models learned under different risk objectives behave distinctly. Characterizing these model behaviors theoretically and empirically, and understanding the trade-offs among these objectives is crucial for building future models.

Chapter 7

Median Optimal Treatment Regimes

7.1 Introduction

Optimal treatment regime methods aim to learn policies that map subject covariates to a decision, typically with the goal of maximizing a population criterion. The task of assigning treatment decisions based on individual characteristics is common and crucially important in many disciplines, with applications ranging from personalized medicine to loan approval to educational program assignment. By now there is a large literature on theory, methods, and applications of optimal treatment regimes (Murphy, 2003; Robins, 2004; Laber et al., 2014; Chakraborty, 2013; Schulte et al., 2014; Laan and Luedtke, 2014; Kosorok and Laber, 2019); we mostly refer to this and related work for more general background and details.

As more individualized policies are deployed in practice, it is of particular importance to examine the target value for which the treatment regimes are optimized. Traditionally, the mean outcome in the population has been used as the criterion. An important reference in this stream for our purposes is Laan and Luedtke (2014), which studies doubly robust estimation of the mean outcome under both a known policy and the unknown optimal policy, showing how nonparametric efficiency bounds can be achieved in both cases. This is part of a larger literature on semiparametric efficiency theory and doubly robust estimation (also known as one-step or debiased or double machine learning estimation) (Newey, 1990; Bickel et al., 1993; Vaart, 2002; van der Laan and Robins, 2003; Tsiatis, 2007; Kennedy, 2016; Chernozhukov et al., 2018).

However, compared to the median, means can be a poor and/or sensitive measure of the centrality of a distribution, for example in the presence of skew or contamination (Huber et al., 1967; Casella and Berger, 2002; Huber, 2004). As a result, a recent literature has evolved studying alternative target values for learning policies. Linn et al. (2017), Wang et al. (2018), and Luedtke, Chambaz, et al. (2020) propose the marginal quantile to be the criterion in settings where the outcome distribution is skewed or the tail of the outcome distribution is of interest. For simplicity we refer to these approaches as marginal median-based, since the issues we detail with marginal median objectives are the same as those for generic quantiles. In presenting the marginal quantile optimal treatment regime, Linn et al. (2017) also considers a marginal

cumulative distribution function related target value where the goal is to find a policy that maximizes the probability of the outcome being above a given threshold. Qi et al. (2019) uses conditional value-at-risk (Rockafellar and Uryasev, 2002) on the outcome distribution as the criterion to ensure the policy to be risk-averse. Unfortunately, in addition to not having a closed-form optimal policy, the marginal median approaches yield optimal treatment decisions for subjects with covariates $X = x$ that depend on outcomes of *other subjects* with different covariates $X \neq x$. We view this as allowing a form of unfairness, which will be discussed in more detail shortly.

Motivated by these concerns, in this paper we propose a treatment regime which assigns treatment to those subjects who have a higher *conditional median* outcome under treatment versus control, and which is optimal with respect to a new objective we call the Average Conditional Median Effect (ACME). Crucially, the optimal policy for the ACME has two key properties, which in general do not both hold for either mean or marginal median optimal treatment regimes:

- "Within-group robustness": Within a group, i.e., for subjects with the same observed covariates, the treatment decision is based on the median of the conditional outcome distribution, and thus the policy is robust to outliers (e.g., when a small fraction of the group has extreme outcomes), unlike the mean optimal treatment regime.
- "Across-group fairness": Across groups, i.e., for subjects with different observed covariates, the treatment decision for a given group cannot depend on outcomes of a different group, unlike the marginal median optimal treatment regime.

Remark 7.1. *Here we use the term "fair" rather loosely; in contrast there has been lots of recent work defining fairness more formally (Dwork et al., 2012; Hardt et al., 2016b; Chouldechova and Roth, 2020; Mehrabi et al., 2019). We also acknowledge that other definitions of fairness could be plausible in our setup, and really only use the above for a convenient shorthand. We describe in much more detail why we refer to the second property as fairness-related in an example in Section 7.4.3.*

Our main contributions are as follows. We reflect on existing standard target values and their optimal treatment regimes and propose the median optimal treatment regime[1], that assigns treatment to an individual based on their *conditional* median treatment effect (Section 7.3). A new measure of policy value, namely the average conditional median effect, is defined. The proposed regime promotes within-group robustness and across-group fairness, in comparison to the mean optimal and marginal median optimal policies (Section 7.4). In Section 7.5, we establish the local asymptotic minimax bound for estimating the ACME, and construct a doubly robust-style estimator along with a simple algorithm that achieves the bound under mild conditions (Section 7.6). To learn the median optimal treatment regime, we propose a new doubly robust-style estimator for the Conditional Median Treatment Effect (CMTE) in Section 7.7. Finally, we use numerical simulations to show finite-sample properties of the estimator and illustrate the algorithm using a dataset from a randomized clinical trial on HIV patients (Section 7.8).

[1]We use *marginal* median optimal treatment regimes to refer to policies that maximize marginal median values.

7.2 Preliminaries

We are given independent and identically distributed (iid) samples $Z_i = (X_i, A_i, Y_i)$ drawn from a distribution $\mathbb{P}$ where $X_i \in \mathcal{X} \subseteq \mathbb{R}^d$ are covariates, $A_i \in \{0, 1\}$ is a binary treatment assignment and $Y_i \in \mathcal{Y}$ is a continuous real-valued outcome of interest. We use Y^a to denote the potential outcome under treatment $A = a$ (Rubin, 1974). Throughout we let $m(V \mid w) = \inf\{v \in \mathbb{R} : \mathbb{P}(V \leq v \mid W = w) \geq 1/2\}$ denote the median of a generic random variable V given $W = w$. To simplify the presentation, we introduce the following notation for components of the distribution $\mathbb{P}$:

$$\begin{aligned}
\pi_a(x) &= \mathbb{P}(A = a \mid X = x), \\
F_a(y \mid x) &= \mathbb{P}(Y \leq y \mid X = x, A = a), \\
f_a(y \mid x) &= \frac{d}{dy} F_a(y \mid x), \\
m_a(x) &= m\left(Y \mid X = x, A = a\right), \\
F_{a,m}(x) &= F_a(m_a(x) \mid X = x), \\
f_{a,m}(x) &= f_a(m_a(x) \mid X = x), \\
\sigma_a(x) &= \sqrt{\text{Var}(Y \mid X = x, A = a)}.
\end{aligned}$$

We note that $\pi_1(x)$ is called the propensity score, i.e., chance of receiving treatment given covariates, and $\pi_0(x) = 1 - \pi_1(x)$. We assume $F_a(y \mid x)$ to be absolutely continuous and hence $f_a(y \mid x)$ exists. Throughout the paper, we refer to $F_a(y \mid x)$ as the conditional cumulative distribution function, $f_a(y \mid x)$ as the conditional density, $m_a(x)$ as the conditional median, $f_{a,m}(x)$ as the conditional density at the conditional median, and $\sigma_a(x)$ as the conditional standard deviation. We assume the following standard conditions for identifying causal effects:

Assumption 7.1. *The following causal assumptions hold*

$$\begin{aligned}
&(\textit{Consistency}) && Y = AY^1 + (1 - A)Y^0, \\
&(\textit{Positivity}) && \mathbb{P}\left(\epsilon \leq \pi_1(X) \leq 1 - \epsilon\right) = 1 \textit{ for some } \epsilon > 0, \\
&(\textit{Exchangeability}) && A \perp\!\!\!\perp Y^a \mid X \textit{ for } a = 0, 1.
\end{aligned}$$

Additional Notation We use $\mathbb{P}\widehat{f} := \int \widehat{f}(z) d\mathbb{P}(z)$ to denote the expectation conditioned upon the randomness of $\widehat{f}$ and $\mathbb{P}_n f := \mathbb{P}_n\{f(Z)\} = \frac{1}{n}\sum_{i=1}^n f(Z_i)$. We denote the Euclidean norm for a real-valued vector $x \in \mathbb{R}^d$ to be $\|x\|_2$ and the $L_2(\mathbb{P})$ norm of a function f to be $\|f\| := \left(\int f(z)^2 d\mathbb{P}(z)\right)^{1/2}$. Finally, $a \lesssim b$ is equivalent to $a \leq Cb$ for some positive constant C.

7.3 Proposed Target Policy & Value

In this section, we first review common target values and their optimal policies in the current literature (Section 7.3.1). In Section 7.3.2, we introduce a new policy and a new measure of a

policy's value, based on conditional medians. We show that under Assumption 7.1 these policies and values are identified.

7.3.1 Standard Policies & Values

Let $d : \mathcal{X} \to \{0, 1\}$ denote a deterministic policy that assigns a treatment based on input covariates and $\mathcal{D}$ be the set of all such measurable functions. Traditionally, most optimal treatment regime research (Murphy, 2003; Zhang et al., 2012) has focused on finding the policy that maximizes the mean outcome

$$\mathbb{E}[Y^d] := \mathbb{E}[d(X)Y^1 + (1 - d(X))\, Y^0],$$

which under Assumption 7.1 is equivalent to $\mathbb{E}[Y^d] = \mathbb{E}_X\left[\mu_d(X)\right]$, where $\mu_d(X) := d(X)\mu_1(X) + (1 - d(X))\mu_0(X)$ and $\mu_a(X) := \mathbb{E}[Y \mid X, A = a]$. When the goal is to maximize the overall population mean $\mathbb{E}[Y^d]$, it is well-known that the optimal policy is given by

$$d^*_{\text{MEAN}}(X) := \mathbb{1}\{\mu_1(X) > \mu_0(X)\}.$$

This policy simply treats only those subjects whose conditional mean treatment effect is positive.

Remark 7.2. *Often, the mean optimal policy d^*_{MEAN} is motivated by the fact that it maximizes $\mathbb{E}[Y^d]$; however, we view its primary motivation as coming from the fact that it maximizes the* conditional *value $\mathbb{E}[Y^d \mid X = x]$ for all $x \in \mathcal{X}$, i.e., it is mean optimal for subjects of any type. From this perspective, the* marginal *value $\mathbb{E}[Y^d]$ is just a one-number summary of the overall performance of a policy, across a heterogeneous population of subjects with different covariates. As we discuss further in subsequent sections, we view the maximization of the conditional value as more fundamental and fair in the across-group sense introduced in Section 7.1 and detailed further in Section 7.4.3.*

While the mean is a valuable measure of centrality in many cases, it is sensitive to outliers and so can be an inappropriate target value when some have extreme responses to treatment. This has led to recent work maximizing the marginal median

$$m(Y^d) := m\left(d(X)Y^1 + (1 - d(X))\, Y^0\right).$$

In the general case, marginal quantile-based values are proposed where the goal is to find the optimal policy with respect to a quantile of the outcome (Linn et al., 2017; Wang et al., 2018; Kallus et al., 2019). Under Assumption 7.1, we obtain that

$$\begin{aligned} m(Y^d) &= \inf\left\{m : \mathbb{P}(Y^d \leq m) \geq 1/2\right\} \\ &= \inf\left\{m : \int \mathbb{P}(Y^d \leq m \mid X = x) d\mathbb{P}(x) \geq 1/2\right\} \\ &= \inf\left\{m : \int \mathbb{P}(Y \leq m \mid A = d(x), X = x) d\mathbb{P}(x) \geq 1/2\right\}. \end{aligned}$$

From here on we use $m(Y^d)$ to refer to the identified quantity above, with the understanding that it corresponds the counterfactual marginal median only under Assumption 7.1.

Unfortunately, as illustrated in Proposition 7.2, the marginal median objective allows specific subjects' optimal treatment assignments to be determined by *different* subjects, with different covariate values. We see this as a violation of across-group fairness. For example, if in some population one group suddenly started responding more to treatment, then under the marginal median objective, the optimal treatment for other groups could change – even if their response to treatment did not.

Unlike the mean optimal policy $d^*_{\text{MEAN}}(x)$, in general, the optimal policy under the marginal median objective (which we denote as $d^*_{\text{MME}}(x)$) cannot be viewed as maximizing a parameter defined upon the conditional outcome distributions $[Y \mid X = x, A = a]$ for $a \in \{0, 1\}$, and does not have a closed form that depends only on the conditional outcome distributions.

7.3.2 Median Optimal Treatment Regimes

As mentioned in Remark 7.2, we believe conditional values are generally more fundamental and useful than marginal values. Therefore here we propose a policy d^*_{ACME} that maximizes the conditional median $m(Y^d \mid X = x)$ for all $x \in \mathcal{X}$. This implies that

$$d^*_{\text{ACME}}(X) := \mathbb{1}\{m_1(X) > m_0(X)\}. \tag{7.1}$$

Using the conditional median instead of the conditional mean to measure centrality of the conditional outcome distributions, compared to the mean optimal policy d^*_{MEAN}, d^*_{ACME} is more *robust* to outliers. Unlike the marginal median optimal policy d^*_{MME}, for all $x \in \mathcal{X}$, our proposed policy $d^*_{\text{ACME}}(x)$ is more *individualized*, in the sense that optimal treatment assignments for subjects with $X = x$ do not depend on outcomes of different subjects with $X \neq x$. We summarize the overall performance of our proposed conditional median optimal policy (also denoted as median optimal treatment regime and median optimal policy) d^*_{ACME} across groups of subjects with different covariates using the average conditional median effect, which we introduce below.

Definition 7.1

Given a policy $d \in \mathcal{D}$, we define the average conditional median effect (ACME) as

$$\mathbb{E}_X[m(Y^d \mid X)] := \mathbb{E}_X\left[m\left(d(X)Y^1 + (1 - d(X))\, Y^0 \mid X\right)\right]. \tag{7.2}$$

Remark 7.3. *Traditionally a treatment "effect" is a* contrast *between (distributions of) potential outcomes; for simplicity we call the ACME parameter an effect even though it involves counterfactuals under only one treatment policy, and so is not a contrast.*

Identification Under Assumption 7.1, we have that

$$\begin{aligned}
\mathbb{E}_X[m(Y^d \mid X)] &= \mathbb{E}_X[d(X)m(Y^1 \mid A = 1, X) + (1 - d(X))m(Y^0 \mid A = 0, X)] \\
&= \mathbb{E}_X[d(X)m(Y \mid A = 1, X) + (1 - d(X))m(Y \mid A = 0, X)] \\
&= \mathbb{E}_X\left[m_d(X)\right],
\end{aligned}$$

where $m_d(X) := d(X)m_1(X) + (1 - d(X))\, m_0(X)$. For the rest of the paper, without further specification, we assume Assumption 7.1 to hold.

As illustrated at the beginning of Section 7.3.2, the ACME is used to summarize the performance of the conditional median optimal policy d^*_{ACME}. We show here that, just as d^*_{MEAN} optimizes $\mathbb{E}[\mathbb{E}[Y^d \mid X]]$, d^*_{ACME} optimizes the ACME $\mathbb{E}[m(Y^d \mid X)]$, i.e., for all $d \in \mathcal{D}$,

$$\begin{aligned}\mathbb{E}_X[m(Y^d \mid X)] &= \mathbb{E}_X[d(X)m_1(X) + (1 - d(X))m_0(X)] \\ &= \mathbb{E}_X[m_0(X) + (m_1(X) - m_0(X))\, d(X)] \\ &\leq \mathbb{E}_X[m_0(X) + (m_1(X) - m_0(X))\, d^*_{\text{ACME}}(X)].\end{aligned}$$

Remark 7.4. *We note that in the case where the polices can only be measurable functions of a subset of covariates $V \subseteq X$, the optimal treatment regime for ACME is not as straightforward as the one for the mean outcome, due to the nonlinearity of the median. Developing the V-specific optimal treatment regime for ACME is an avenue for future work.*

7.4 Policy Comparisons

In this section, we compare the three policies d^*_{MEAN}, d^*_{MME} and d^*_{ACME}. In Section 7.4.1, we give a simple condition under which the mean and median optimal treatment regimes are equivalent, and point out that the mean optimal policy is not necessarily robust within groups (defined in Section 7.1). In Section 7.4.2, we illustrate that the marginal median optimal treatment regime does not ensure across-group fairness (defined in Section 7.1). In Section 7.4.3, we use a motivating example to summarize the differences among the three policies.

7.4.1 d^*_{MEAN} and d^*_{ACME}

To characterize the difference between the mean optimal treatment regime d^*_{MEAN} and our proposed median optimal treatment regime d^*_{ACME}, we begin by stating a simple condition that ensures the two to be equivalent: $d^*_{\text{MEAN}}(x) = d^*_{\text{ACME}}(x)$ if and only if the conditional mean and median treatment effects are always the same sign. One sufficient condition for the two policies to be equivalent is thus given below.

Proposition 7.1

If the conditional mean and median treatment effects are separated from zero relative to the conditional outcome standard deviation, in the sense that

$$|\mu_1(x) - \mu_0(x)| > \sigma_1(x) + \sigma_0(x) \;\text{ and }\; |m_1(x) - m_0(x)| > \sigma_1(x) + \sigma_0(x),$$

then $d^*_{\text{MEAN}}(x) = d^*_{\text{ACME}}(x)$.

An important note about the mean optimal policy d^*_{MEAN} is that it does not necessarily satisfy the within-group robustness mentioned in Section 7.1, in that it is sensitive to outliers. For

example, as illustrated in Section 7.4.3, its decision can be very affected by a small subgroup that benefits a lot from the treatment, even if the majority of the subgroup is harmed by the decision. In contrast, the median remains unchanged as long as the amount of mass below and above it remains unchanged, while the mean is determined by how the masses are distributed, e.g., the mean can be arbitrarily high (or low) by moving a small fraction of the mass above (or below) the median to extreme values (Huber et al., 1967; Casella and Berger, 2002; Huber, 2004). Further comparisons between the mean optimal and our proposed median optimal treatment regime are given in Section 7.4.3.

7.4.2 d^*_{MME} and d^*_{ACME}

To illustrate ideas, here we consider a simple Gaussian model to compare the marginal median optimal treatment regime d^*_{MME} and our proposed median optimal treatment regime d^*_{ACME}. This provides a counterexample showing that $d^*_{\text{MME}}(x)$ is determined not just by the conditional outcome distribution $[Y \mid X = x, A = a]$ for $a \in \{0, 1\}$ but also by the conditional outcome distribution at other (different) covariate values. That is, even if the conditional distribution $[Y \mid X = x, A = a]$ for $a \in \{0, 1\}$ is unchanged, the marginal median optimal policy $d^*_{\text{MME}}(x)$ can be different depending on $[Y \mid X = x', A = a]$ for $x' \neq x$. On the other hand, our median optimal policy $d^*_{\text{ACME}}(x)$ is fair across groups, in that the optimal decision remains the same so long as the conditional distribution for $X = x$ is unchanged.

Specifically consider a data-generating process where

$$X \sim \text{Bernoulli}(1/2)$$
$$Y \mid X = x, A = a \sim N\left(\mu_a(x), \sigma_a^2(x)\right)$$

for some unspecified (for now) mean $\mu_a(x)$ and standard deviation $\sigma_a(x)$. For any policy $d \in \mathcal{D}$, the ACME is $\mathbb{E}[m_d(X)] = (\mu_d(0) + \mu_d(1))/2$, and the marginal median value $m(Y^d)$ satisfies $\Phi\left(\frac{m(Y^d)-\mu_d(0)}{\sigma_d(0)}\right) + \Phi\left(\frac{m(Y^d)-\mu_d(1)}{\sigma_d(1)}\right) = 1$, which implies that

$$m(Y^d) = \frac{\sigma_d(1)\mu_d(0) + \sigma_d(0)\mu_d(1)}{\sigma_d(0) + \sigma_d(1)},$$

i.e., it is a standard deviation-weighted average of outcome regressions (where each regression is weighted by the standard deviation of the other group). In the next proposition we give a counterexample showing how, somewhat surprisingly, the marginal median optimal policy for subjects with $X = 1$, for example, can depend on the means and variances for different subjects with $X = 0$.

Proposition 7.2

Suppose outcomes are higher and more variable under treatment for $X = 1$, i.e.,

$$\mu_1(1) > \mu_0(1) \text{ and } \sigma_1(1) > \sigma_0(1),$$

and the variances differ more than the means in the sense that $\frac{\mu_1(1)}{\mu_0(1)} < \frac{\sigma_1(1)}{\sigma_0(1)}$. Then there

exist conditional distributions Q_a and $\overline{Q}_a$ for $[Y \mid X = 0, A = a]$ where $a \in \{0, 1\}$ such that

$$\mathbb{1}\{\mu_1(1) > \mu_0(1)\} = d^*_{\text{MME}}(1; Q_0, Q_1) \neq d^*_{\text{MME}}(1; \overline{Q}_0, \overline{Q}_1) = \mathbb{1}\{\mu_1(1) \leq \mu_0(1)\}$$

where $d^*_{\text{MME}}(x; Q_0, Q_1)$ and $d^*_{\text{MME}}(x; \overline{Q}_0, \overline{Q}_1)$ are the marginal median optimal policies when $[Y \mid X = x, A = a]$ follows distributions Q_a and $\overline{Q}_a$ respectively.

Remark 7.5. *When $\sigma_1(0) = \sigma_0(0) = \sigma_1(1) = \sigma_0(1)$, it follows that $m(Y^d) = \mathbb{E}[m_d(X)]$ for all $d \in \mathcal{D}$ and $d^*_{ACME}(x) = \mathbb{1}\{\mu_1(x) > \mu_0(x)\}$ is an optimal policy with respect to $m(Y^d)$.*

Remark 7.6 (Distribution shift). *Importantly, we also note that among the three optimal policies, only the marginal median optimal policy d^*_{MME} varies under different covariate distribution $\mathbb{P}(X)$. This is crucial when distribution shift is a concern, which is common in many current setups (Quionero-Candela et al., 2009; Mo et al., 2020). Therefore, under distribution shift between training and testing times, the marginal median optimal policy d^*_{MME} could be suboptimal at testing time even if optimal at training. For example, if during test time $\mathbb{P}(X = 0) = 0$ and $\mathbb{P}(X = 1) = 1$, then in the case when $d^*_{MME}(1) = \mathbb{1}\{\mu_1(1) \leq \mu_0(1)\}$ during training and $\mu_1(1) \neq \mu_0(1)$, d^*_{MME} is not optimal (with respect to the marginal median) during test time since $m(Y^d) = \mu_d(1)$. Unlike d^*_{MME}, our proposed median optimal treatment regime d^*_{ACME} remains the same under covariate shifts.*

As illustrated in Proposition 7.2, the optimal decision for a specific subject of covariate x with respect to the marginal median depends on not just its own conditional distribution $[Y \mid X = x, A = a]$ for $a \in \{0, 1\}$ but also the conditional distributions at other covariate values. On the other hand, ACME allows the optimal decision for x to be only influenced by its own conditional outcome distribution at that x. In Section 7.4.3, we give an illustration that summarizes the differences among the three policies.

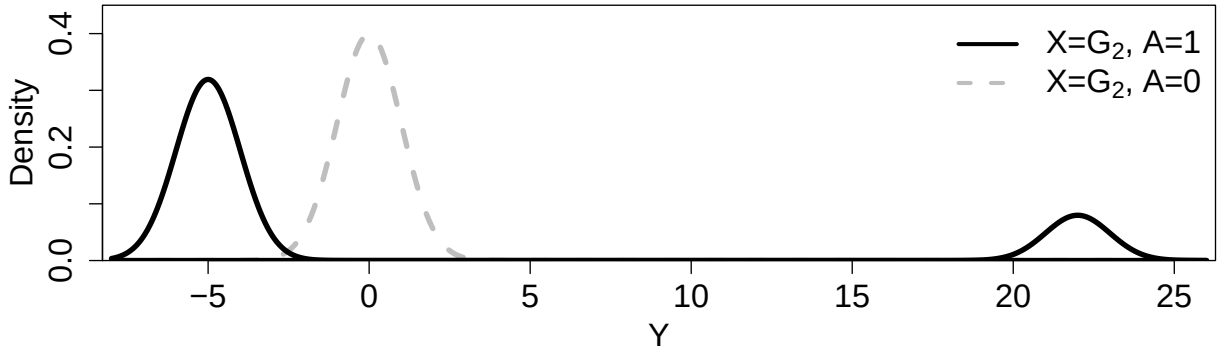


Figure 7.1: The density of $[Y \mid X = G_2, A = a]$ in Case I and Case II.

7.4.3 Simple Illustration Comparing d^*_{MEAN}, d^*_{MME}, and d^*_{ACME}

Here we give a simple illustration that showcases the differences among the three policies, and among the corresponding values they maximize. We will see how the ACME optimal policy d^*_{ACME} exhibits both within-group robustness and across-group fairness, while the more standard policies d^*_{MEAN} and d^*_{MME} may not. We note that although in this illustrative example, the covariate X is chosen to be discrete, none of the above characterization of median optimal treatment regimes and the theory we develop for estimating and evaluating such regimes require the covaraite to be discrete. In particular, while within-group robustness may be more interpretable in settings where there are multiple treatment outcomes under the same covariate, across-group robustness is a property that does not require it. That is, a treatment decision of one subject should ideally not depend on the outcome distribution of another subject with different covariates, regardless of whether there are more than one individual with the same covariate.

Consider a population of two groups: one with college degrees (G_1) and the other without (G_2). The binary treatment A indicates whether a job training program is assigned to the individual and the outcome Y is the subjects' improvement in income. In Case I, suppose the groups and outcomes follow the distribution:

$$\begin{aligned} X &\sim \text{Bernoulli}(1/2) \\ Y \mid X = G_1, A = a &\sim \mathcal{N}(9, 1) \\ Y \mid X = G_2, A = a &\sim (1-a)\mathcal{N}(0,1) + a(.2\,\mathcal{N}(22,1) + .8\,\mathcal{N}(-5,1)) \end{aligned}$$

In other words, there is no treatment effect for subjects with a college degree, whose outcome distribution is always the same Gaussian. However, under control subjects without a college degree have outcomes that follow a Gaussian centered at 0, while under treatment the outcomes follow a Gaussian mixture. The outcome distribution for G_2 is given in Figure 7.1. When examining the optimal treatment under each value, we find that both the mean optimal policy d^*_{MEAN} and marginal median optimal policy d^*_{MME} treat all subjects without a college degree (Table 7.1), while d^*_{ACME} does not treat them. Compared to d^*_{MEAN}, our proposed median optimal policy d^*_{ACME} is robust against outliers, hence its decisions promotes within-group robustness in the sense that the decisions are less prone to only benefiting a small subgroup, e.g., as in our $A = 1$ case for the group without a college degree. On the other hand, although intuitively the marginal median optimal policy d^*_{MME} also gives robust decisions, it overlooks characteristics of specific subgroups. In particular, treatment decisions for those with $X = x$ can depend on other subjects' conditional outcome distribution with $X \neq x$, violating across-group fairness. This is illustrated by comparing the optimal marginal median decision for subjects without a college degree under Case I and Case II where in Case II the the conditional outcome distribution for subjects with a college degree under treatment and control is changed to

$$Y \mid X = G_1, A = a \sim \mathcal{N}(-4, 2)\,.$$

Surprisingly, as in Proposition 7.2, the marginal median optimal policy d^*_{MME} under Case II is different for individuals without a college degree compared to the d^*_{MME} under Case I, even

though it was only the outcome distribution for subjects with a college degree that changed. This illustrates across-group unfairness of the marginal median value. In contrast, the median optimal decision for individuals without a college degree has remained unchanged.

	Case I		Case II	
Policy	Treat G_1	Treat G_2	Treat G_1	Treat G_2
d^*_{MEAN}	–	✓	–	✓
d^*_{MME}	–	✓	–	
d^*_{ACME}	–		–	

Table 7.1: The table shows the treatment assignments under the mean optimal policy d^*_{MEAN}, the marginal median optimal policy d^*_{MME}, and the median optimal policy d^*_{ACME} in Case I and Case II. Since in both cases, the conditional outcome distributions for G_1 are the same under treatment and control, "–" is used to indicate that the optimal decision for G_1 can be either $a = 0$ or $a = 1$. Though the conditional outcome distributions for G_2 remain unchanged in both cases, the optimal marginal median decision for G_2 has changed.

To summarize, our characterization for the three types of policies shed light on when to use each of them. Depending on the context, one may choose to use the ACME optimal policy d^*_{ACME} because of reasons including that the conditional outcome distribution is skewed, the policy should not decide a subject's treatment based on other subjects' outcome distribution, or the policy should not be influenced by covariate distribution shifts. In addition, as we discuss next, one may also choose the policy to use based on how well one can estimate it. Our discussion above should be interpreted as characterizations of different treatment regimes, and a reminder for us to start critically examining the target value of a treatment regime, instead of a firm conclusion that one should only use median optimal treatment regimes.

7.5 Efficiency Bound

In this section, using semiparametric theory (Newey, 1990; Bickel et al., 1993; Vaart, 2002; van der Laan and Robins, 2003; Tsiatis, 2007; Kennedy, 2016; Dıéaz, 2017; Chernozhukov et al., 2018), we study the nonparametric efficiency bound for estimating the ACME $\psi_d := \mathbb{E}[m_d(X)]$ of a given policy d, given iid samples from distribution $\mathbb{P}$. One of the key tools we will use throughout the paper is the efficient influence function $\phi_d(Z)$ (Corollary 7.2) which serves as a first-order derivative in a von Mises-type expansion of the target parameter (Lemma 7.1).

There are a number of reasons why characterizing efficient influence functions is important. The variance of the efficient influence function gives a minimax lower bound for estimating ψ_d (Theorem 7.1) and so provides a benchmark to compare against when constructing estimators. In particular, as used in Section 7.6, efficient influence functions suggest a doubly robust-style bias-corrected estimator. Doubly robust estimators attain fast parametric rates in nonparametric settings where nuisance functions (e.g., $m_a, \pi_a, f_{a,m}$ in our context) are estimated at slower rates. We begin with presenting the von Mises-type expansion (a distributional Taylor expansion) for

ACME, which is crucial for the efficiency bound (Theorem 7.1) and the estimation guarantees for our proposed doubly robust-style estimator (Theorem 7.2).

Lemma 7.1

Given a policy $d \in \mathcal{D}$, for the average conditional median effect ψ_d defined in (7.2), the following decomposition holds for all distributions $\overline{\mathbb{P}}$ and $\mathbb{P}$:

$$\psi_d(\overline{\mathbb{P}}) - \psi_d(\mathbb{P}) = \int \phi_d(Z; \overline{\mathbb{P}}) d(\overline{\mathbb{P}} - \mathbb{P}) + R_d(\overline{\mathbb{P}}, \mathbb{P}),$$

where $\phi_d(Z; \mathbb{P}) = \xi_d(Z; \mathbb{P}) - \psi_d(\mathbb{P})$,

$$\xi_d(Z; \mathbb{P}) = \frac{Ad(X)}{\pi_1(X)} \frac{1/2 - \mathbb{1}\{Y \leq m_1(X)\}}{f_{1,m}(X)} + \frac{(1-A)(1-d(X))}{\pi_0(X)} \frac{1/2 - \mathbb{1}\{Y \leq m_0(X)\}}{f_{0,m}(X)} + m_d(X), \tag{7.3}$$

and

$$R_d(\overline{\mathbb{P}}, \mathbb{P}) = \int_{\mathcal{X}} d(x) \left(\overline{m}_1(x) - m_1(x) - \frac{\pi_1(x)}{\overline{\pi}_1(x)} \frac{F_{1,\overline{m}}(x) - F_{1,m}(x)}{\overline{f}_{1,\overline{m}}(x)} \right) + (1 - d(x)) \left(\overline{m}_0(x) - m_0(x) - \frac{\pi_0(x)}{\overline{\pi}_0(x)} \frac{F_{0,\overline{m}}(x) - F_{0,m}(x)}{\overline{f}_{0,\overline{m}}(x)} \right) d\mathbb{P}(x). \tag{7.4}$$

$\overline{\pi}_1, \overline{m}_a$ and $\overline{f}_{a,\overline{m}}$ are the propensity score, the conditional median and the conditional density at the conditional median defined under $\overline{\mathbb{P}}$ and $F_{a,\overline{m}}(x) = F_a(\overline{m}_a(x) \mid x)$.

Proofs of the results in this section can be found in Appendix F.1.2. Lemma 7.1 has several important implications. It implies that the efficient influence function of ACME is $\phi_d(Z)$ as shown in Corollary 7.2. The efficient influence function plays a crucial role in the bias term for the plug-in estimator $\psi_d(\widehat{\mathbb{P}})$, which we will correct to obtain the doubly robust-style estimator given later on in (7.8). The decomposition result in (7.4) (along with an alternate expression in (7.5)) is then used to show the convergence rate of the proposed estimator.

Throughout the rest of the paper, we rely on the following model assumption, which ensures that certain boundedness and mild smoothness conditions hold.

Assumption 7.2. *The true distribution $\mathbb{P}$ lies in $\mathcal{P}$ where $\mathcal{P}$ is the statistical model given by*

$$\mathcal{P} := \Big\{ \mathbb{P} \;\Big|\; \exists \epsilon \in (0,1), \mathbb{P}\left(\epsilon < \pi_1(X) \leq 1 - \epsilon\right) = 1,$$
$$\forall a \in \{0,1\}, \exists\, 0 < M_0 \leq M_1, \mathbb{P}\left(M_0 \leq f_{a,m}(X) \leq M_1\right) = 1, \textit{ and}$$
$$\forall a \in \{0,1\}, y \in \mathcal{Y}, f_a(y \mid X) \textit{ is differentiable and L-Lipschitz continuous in } y \textit{ almost surely} \Big\}.$$

The conditions of Assumption 7.2 only require some boundedness of the propensity score and conditional outcome density, as well as some weak smoothness of the conditional density

in y. Importantly, Assumption 7.2 does not require any smoothness of the propensity score or outcome density in X. The condition on $f_{a,m}(X)$ ensures that the density at the conditional median outcome is bounded away from 0 almost surely, ensuring sufficient number of samples for estimating the median.

As written, the reminder term $R_d(\overline{\mathbb{P}}, \mathbb{P})$ in Lemma 7.1 does not appear to be second-order, i.e., involving products of differences of $\overline{\mathbb{P}}$ and $\mathbb{P}$. However, in the next corollary we show that it is in fact second-order, under the model assumption (Assumption 7.2).

Corollary 7.1. *For all distributions $\mathbb{P} \in \mathcal{P}$, given any distribution $\overline{\mathbb{P}}$, the remainder term $R_d(\overline{\mathbb{P}}, \mathbb{P})$ defined in (7.4) is equivalent to*

$$R_d(\overline{\mathbb{P}}, \mathbb{P}) = \mathbb{P}\Bigg\{ \frac{d\,(\overline{m}_1 - m_1)}{\overline{\pi}_1 \overline{f}_{1,\overline{m}}} \left((\overline{\pi}_1 - \pi_1)\,\overline{f}_{1,\overline{m}} + \left(\overline{f}_{1,\overline{m}} - f_{1,m}\right) \pi_1 - (\overline{m}_1 - m_1) \frac{f'_{1,c}\pi_1}{2} \right) + \frac{(1-d)\,(\overline{m}_0 - m_0)}{\overline{\pi}_0 \overline{f}_{0,\overline{m}}} \left((\overline{\pi}_0 - \pi_0)\overline{f}_{0,\overline{m}} + (\overline{f}_{0,\overline{m}} - f_{0,m})\pi_0 - (\overline{m}_0 - m_0) \frac{f'_{0,c}\pi_0}{2} \right) \Bigg\}, \tag{7.5}$$

where $f'_{a,c} \leq L$ is the derivative of $f_a(y \mid x)$ at $y = c_a(x)$ for some value $c_a(x)$ between $m_a(x)$ and $\overline{m}_a(x)$.[2]

The fact that the remainder term is second-order implies that ϕ_d from Lemma 7.1 is the efficient influence function, as stated in the next corollary.

Corollary 7.2. *For a given policy $d \in \mathcal{D}$, the efficient influence function for the average conditional median effect ψ_d is $\phi_d(Z) = \xi_d(Z) - \psi_d$ where ξ_d is defined in (7.3).*

Since the variance of ϕ_d serves as a nonparametric efficiency bound, in the local minimax sense (Vaart, 2002), in the next result we give the exact form of this variance. This shows what factors drive the statistical difficulty in ACME estimation, and demonstrates the difference in efficiency for ACME versus other value measures.

Theorem 7.1

For a given policy $d \in \mathcal{D}$, the nonparametric efficiency bound for estimating ψ_d is given by the variance $\sigma_d^2 := \mathrm{Var}\{\phi_d(Z)\}$ where

$$\sigma_d^2 = \mathbb{E}\left[\frac{d(X)}{4\pi_1(X) f_{1,m}^2(X)} + \frac{1 - d(X)}{4\pi_0(X) f_{0,m}^2(X)} \right] + \mathrm{Var}\left(m_d(X)\right). \tag{7.6}$$

Theorem 7.1 shows how the efficiency bound for the ACME is driven by three main factors:

- the inverse propensity score $1/\pi_1(X)$,
- the inverse conditional density at the conditional median $1/f_{a,m}(X)$, and
- the heterogeneity of the conditional median $m_d(X)$, i.e., $\mathrm{Var}\left(m_d(X)\right)$.

In contrast, recall the nonparametric efficiency bound for the mean value $\mathbb{E}[Y^d]$ (Hahn, 1998,

[2] Since $\mathbb{P} \in \mathcal{P}$, $f'_{a,c}$ exists and is bounded by L almost surely.

Theorem 1):

$$\sigma^2_{d,\text{MEAN}} := \mathbb{E}\left[\frac{d(X)\sigma_1^2(X)}{\pi_1(X)} + \frac{(1-d(X))\sigma_0^2(X)}{\pi_0(X)}\right] + \text{Var}\left(\mu_d(X)\right),$$

instead depends on the heterogeneity of the conditional outcomes $\sigma_a^2(X)$ and the heterogeneity of the conditional mean Var $(\mu_d(X))$ (along with the inverse propensity score).

Importantly, when the conditional distribution $[Y \mid X = x, A = a]$ is heavy-tailed, e.g., the outcome variance $\sigma_a^2(X)$ is very large, then in general the mean value efficiency bound $\sigma^2_{d,\text{MEAN}}$ would be severely affected, while the ACME analog σ_d^2 could still be quite small. However, in settings where $f_{a,m}(x)$ is close to 0 for all $x \in \mathcal{X}$, σ_d^2 may be larger than $\sigma^2_{d,\text{MEAN}}$. (In settings where $f_{a,m}(x)$ is close to 0 only for covariates with low densities, σ_d^2 may still be low.) This is understandable, as one requires samples around the conditional median outcome in order to estimate it. When $f_{a,m}(x)$ is not bounded away from 0, e.g., when Assumption 7.2 is violated, one may not be able to obtain enough samples to estimate the conditional medians and thus should consider using alternative policies, e.g., the mean optimal treatment regimes. In general, if one can estimate $\sigma_a^2(X)$ and $f_{a,m}(X)$ reliably well, one may use these estimates to obtain the estimated efficiency bound to decide between the conditional mean and outcome treatment regimes.

Next we formalize the local minimax lower bound property of the variance σ_d^2.

Corollary 7.3. *(Vaart, 2002, Corollary 2.6) For a given policy $d \in \mathcal{D}$ and any estimator $\widehat{\psi}_d$ learned using n iid samples from $\mathbb{P}$, it follows that*

$$\inf_{\delta>0} \liminf_{n\to\infty} \sup_{TV(\overline{\mathbb{P}},\mathbb{P})<\delta} n\, \mathbb{E}_{\overline{\mathbb{P}}}\left[\left(\widehat{\psi}_d - \psi_d\left(\overline{\mathbb{P}}\right)\right)^2\right] \geq \sigma_d^2,$$

where $TV(\overline{\mathbb{P}}, \mathbb{P})$ is the total variation distance between $\overline{\mathbb{P}}$ and $\mathbb{P}$ and σ_d^2 is defined in (7.6).

Corollary 7.3 directly follows from Vaart (2002, Corollary 2.6). It shows that without further assumptions, the asymptotic local minimax mean squared error of any estimator scaled by n can be no smaller than σ_d^2. In Section 7.6, under mild conditions, we construct doubly robust-style estimators that achieve the local minimax lower bound shown in Corollary 7.3.

7.6 Estimation of the ACME

In this section, we propose efficient estimators of the ACME that utilize the efficient influence function given in Corollary 7.2. We start by presenting a simple plug-in estimator, which is a sample analogue of ψ_d for a given policy $d \in \mathcal{D}$. In Section 7.6.1, we present the doubly robust-style estimator for evaluating the ACME of a fixed given policy, as well as the optimal policy, and which improves on the simple plug-in. Specifically, under mild conditions, we show that in both cases the doubly robust-style estimator achieves the local minimax lower bound shown in Corollary 7.3. Finally, we present an algorithm for constructing the doubly robust-style estimator (Section 7.6.2).

Let $Z^n = (Z_1, \ldots, Z_n)$ and $Z^{n,0} = (Z_1^0, \ldots, Z_n^0)$ denote two independent samples. For a given policy $d \in \mathcal{D}$, the most natural estimator for estimating ψ_d is the plug-in estimator, i.e.,

$$\widehat{\psi}_{d,\text{pi}} = \mathbb{P}_n \left\{ d(X)\widehat{m}_1(X) + (1 - d(X))\, \widehat{m}_0(X) \right\}, \tag{7.7}$$

where $\widehat{m}_1$ and $\widehat{m}_0$ are learned using a separate sample $Z^{n,0}$ from the sample Z^n used for taking the sample average $\mathbb{P}_n$. While the plug-in estimator is intuitive, in general it inherits the slow convergence rate from the nuisance functions (Bickel et al., 1993; Vaart, 2002). As suggested by the von Mises-type decomposition (Lemma 7.1), a natural estimator that improves upon the plug-in estimator is given in (7.8).

Remark 7.7. *For a given set of iid data, we can obtain Z^n and $Z^{n,0}$ by randomly splitting the data in half. In general, to obtain full sample size efficiency, one can split the data in folds and perform cross-fitting (Bickel and Ritov, 1988; Robins et al., 2008; Zheng and Laan, 2010; Chernozhukov et al., 2018), i.e., repeat the learning procedure for each split and average the results. Throughout this section, to ease notation, we present the analyses and results for the learning procedure with a single sample split, which can be easily extended to averages of multiple independent splits. If one wants to avoid sample splitting, our results would still hold under appropriate empirical process conditions (e.g., the nuisance functions taking values in Donsker classes).*

7.6.1 Doubly Robust-Style Estimator of the ACME

To improve on the potential deficiencies of the simple plug-in above, we propose the doubly robust-style estimator $\widehat{\psi}_{d,\text{dr}}$:

$$\widehat{\psi}_{d,\text{dr}} = \mathbb{P}_n \left\{ \frac{\mathbb{1}\{A = d(X)\}}{A\widehat{\pi}_1(X) + (1 - A)\widehat{\pi}_0(X)} \frac{1/2 - \mathbb{1}\{Y \leq \widehat{m}_A(X)\}}{\widehat{f}_{A,\widehat{m}}(X)} + \widehat{m}_d(X) \right\}, \tag{7.8}$$

where the nuisance functions $\widehat{\pi}_1$, $\widehat{m}_a$ and $\widehat{f}_{a,\widehat{m}}$ are learned using the separate sample $Z^{n,0}$, and $\widehat{\pi}_0(X) = 1 - \widehat{\pi}_1(X)$. Importantly, this estimator uses the efficient influence function (Corollary 7.2) to correct the bias of the plug-in.

Remark 7.8. *As will be shown in Theorem 7.2, our estimator is not strictly doubly robust in the usual sense (i.e., its consistency requires consistent estimation of the conditional median). Nonetheless, we still call it doubly robust-style because it does have what is arguably the most crucial property of doubly robust estimators: its error involves* second-order *products, and so is "doubly small".*

In Section 7.6.1, we study the asymptotic convergence rate of $\widehat{\psi}_{d,\text{dr}}$ for a given fixed policy d, showing it can converge at parametric rates even when nuisance functions are estimated nonparametrically, and follows a straightforward asymptotic normal distribution. Then in Section 7.6.1, we show that under mild conditions, our estimator $\widehat{\psi}_{d,\text{dr}}$ still exhibits $\sqrt{n}$-convergence and asymptotic normality for the optimal ACME value, even after plugging in the learned policy $\mathbb{1}\{\widehat{m}_1(X) > \widehat{m}_0(X)\}$, under a margin condition.

ACME of a Fixed Policy

We begin with the case when a fixed policy d is given and we aim to estimate the ACME of it. This can be useful if we have a specified class of policies, which we aim to optimize over; this includes the case where we use an independent sample or split to learn a policy, and then condition on that sample when estimating the value using another. In such cases, under mild assumptions, we show in Theorem 7.2 that the error of $\widehat{\psi}_{d,\text{dr}}$ is the sum of a centered sample average term which is asymptotically normal and a term containing products of nuisance estimation errors.

Theorem 7.2

Assume

1. $\mathbb{P}\left(\epsilon \leq \widehat{\pi}_1(X) \leq 1-\epsilon\right) = 1$ for some $\epsilon \in (0,1)$,
2. $\mathbb{P}(M_0 \leq \widehat{f}_{a,\widehat{m}}(X) \leq M_1) = 1$ for $a \in \{0,1\}$ and $0 < M_0 \leq M_1$, and
3. $\|\widehat{\xi}_d - \xi_d\| = o_{\mathbb{P}}(1)$.

Then,

$$\widehat{\psi}_{d,\text{dr}} - \psi_d = \left(\mathbb{P}_n - \mathbb{P}\right)\xi_d(\mathbb{P}) + O_{\mathbb{P}}\left(\sum_{a=0}^{1} \|\widehat{m}_a - m_a\| \left(\|\widehat{\pi}_a \widehat{f}_{a,\widehat{m}} - \pi_a f_{a,m}\| + \|\widehat{m}_a - m_a\|\right) + o_{\mathbb{P}}(1/\sqrt{n})\right).$$

Importantly, Theorem 7.2 implies that $\widehat{\psi}_{d,\text{dr}}$ attains a faster convergence rate than its nuisance estimators and can be asymptotically normal even when these nuisance functions only satisfy nonparametric sparsity, smoothness or other assumptions. We detail this further in the next corollary.

Corollary 7.4. *Given $d \in \mathcal{D}$, under assumptions in Theorem 7.2, and the following two conditions:*

1. $\|\widehat{m}_a - m_a\| = o_{\mathbb{P}}(n^{-1/4})$, *and*
2. *For $a \in \{0,1\}$, $\|\widehat{\pi}_a \widehat{f}_{a,\widehat{m}} - \pi_a f_{a,m}\| = O_{\mathbb{P}}(n^{-1/4})$,*

we have that $\widehat{\psi}_{d,dr}$ is root-n consistent and asymptotically normal with:

$$\sqrt{n}(\widehat{\psi}_{d,dr} - \psi_d) \rightsquigarrow \mathcal{N}\left(0, \sigma_d^2\right).$$

Remark 7.9. *Note that one sufficient condition for $\|\widehat{\pi}_a \widehat{f}_{a,\widehat{m}} - \pi_a f_{a,m}\| = O_{\mathbb{P}}(n^{-1/4})$ is that $\|\widehat{\pi}_a - \pi_a\| = O_{\mathbb{P}}(n^{-1/4})$ and $\|\widehat{f}_{a,\widehat{m}} - f_{a,m}\| = O_{\mathbb{P}}(n^{-1/4})$ separately.*

Corollary 7.4 shows that $\widehat{\psi}_{d,\text{dr}}$ can achieve the nonparametric efficiency bound given in Corollary 7.3. For example, when π_a, m_a and $f_{a,m}$ are d-dimensional functions in a Hölder smooth class with smoothness parameter being α, β and γ respectively (i.e., their partial derivatives up to order α, β, γ exist and are Lipschitz) and are estimated with squared error $n^{-2\alpha/(2\alpha+d)}$, $n^{-2\beta/(2\beta+d)}$ and $n^{-2\gamma/(2\gamma+d)}$, then the conditions in Corollary 7.4 will be satisfied when $\alpha, \beta, \gamma \geq d/2$, i.e., when the smoothness is greater than half the dimension.

ACME of the Optimal Policy

Instead of evaluating a fixed policy d, one may want to evaluate the ACME of the truly optimal (but unknown) policy $d^*_{\text{ACME}} = \mathbb{1}\{\gamma(X) > 0\}$ where $\gamma(X) := m_1(X) - m_0(X)$ is the conditional median treatment effect (CMTE). This would give a benchmark for the best possible value that could be achieved by any policy; any improvements would have to come by way of changing the population or changing the treatment. A natural estimator for the optimal policy would simply plug in an estimate of γ: $\widehat{d}^*_{\text{ACME}}(X) = \mathbb{1}\{\widehat{\gamma}(X) > 0\}$. Here, the policy $\widehat{d}^*_{\text{ACME}}$ is learned through the same sample $Z^{n,0}$ as the one used for estimating the nuisance functions (7.8). For estimating the CMTE γ, in addition to the plug-in $\widehat{m}_1 - \widehat{m}_0$, a doubly robust-style estimator will be discussed in Section 7.7. An immediate question to ask is whether the estimator $\widehat{\psi}_{\widehat{d}^*_{\text{ACME}},\text{dr}}$ for the learned policy $\widehat{d}^*_{\text{ACME}}$ can still exhibit similar asymptotic convergence guarantees as $\widehat{\psi}_{d,\text{dr}}$ for a fixed policy d. Our goal in this subsection is to answer this question. To simplify notation, we use $\psi_{d^*} := \psi_{d^*_{\text{ACME}}}$ and $\widehat{\psi}_{\widehat{d}^*,\text{dr}} := \widehat{\psi}_{\widehat{d}^*_{\text{ACME}},\text{dr}}$.

Importantly, the ACME of the optimal policy is a non-smooth functional, because of its dependence on the indicator function $\mathbb{1}\{\gamma(X) > 0\}$. This challenge also arises in the mean value setting (Chakraborty et al., 2010; Laber and Murphy, 2011; Hirano and Porter, 2012; Laan and Luedtke, 2014; Laber et al., 2014; Luedtke and Laan, 2016). One solution is to incorporate a margin condition (Tsybakov et al., 2004; Luedtke and Laan, 2016), as follows.

Assumption 7.3 (Margin Condition)**.** *For some $\alpha > 0$ and all $t > 0$, we have*

$$\mathbb{P}\left(|\gamma(X)| \leq t\right) \leq c(t)^{\alpha},$$

for some constant $c > 0$ such that $ct \leq 1$.

The exponent α in the margin condition characterizes the mass such that $m_1(X)$ and $m_0(X)$ are close. The lower α is, the weaker the margin condition is, i.e., the more mass the distribution of $\gamma(X)$ is allowed to have near zero. Similar margin conditions are used in classification literature (Tsybakov et al., 2004) and optimal treatment regimes (Luedtke and Laan, 2016). Note that $\alpha = 0$ encodes no assumption, allowing $\gamma(X) = 0$ almost surely, while $\alpha = 1$ would hold as long as $\gamma(X)$ is continuously distributed with bounded density. The margin condition therefore provides a characterization on how hard the optimal decision problem is—intuitively, when $\gamma(X)$ is near zero, it is very hard to distinguish which subjects will benefit from treatment, while when $\gamma(X)$ is very different from zero, this is easy to distinguish. Under the margin condition, we show in Theorem 7.3 that the error of $\widehat{\psi}_{\widehat{d}^*,\text{dr}}$ is similar to that of the error of $\widehat{\psi}_{d^*,\text{dr}}$ for the truly optimal policy d^*.

Theorem 7.3

Assume

1. $\mathbb{P}\left(\epsilon \leq \widehat{\pi}_1(X) \leq 1-\epsilon\right) = 1$ for some $\epsilon \in (0,1)$,
2. $\mathbb{P}(M_0 \leq \widehat{f}_{a,\widehat{m}}(X) \leq M_1) = 1$ for $a \in \{0,1\}$ and $0 < M_0 \leq M_1$, and
3. $\|\widehat{\xi}_{\widehat{d}^*} - \xi_{d^*}\| = o_{\mathbb{P}}(1)$.

Then, under Assumption 7.3, we have that

$$\widehat{\psi}_{\widehat{d}^*,\mathrm{dr}} - \psi_{d^*} = (\mathbb{P}_n - \mathbb{P})\,\xi_{d^*}$$
$$+ O_{\mathbb{P}}\Bigg(\sum_{a=0}^{1} \|\widehat{m}_a - m_a\| \left(\|\widehat{\pi}_a \widehat{f}_{a,\widehat{m}} - \pi_a f_{a,m}\| + \|\widehat{m}_a - m_a\|\right) + \|\widehat{\gamma} - \gamma\|_\infty^{1+\alpha} + o_{\mathbb{P}}(1/\sqrt{n})\Bigg).$$

Unlike the error of $\widehat{\psi}_{d,\mathrm{dr}}$ for a fixed policy d presented in Theorem 7.2, the error of $\widehat{\psi}_{\widehat{d}^*,\mathrm{dr}}$ depends on $\|\widehat{\gamma} - \gamma\|_\infty^{1+\alpha}$. When α gets higher, the error of $\widehat{\gamma}$ plays less of a role, as captured in Corollary 7.5. Under the conditions in Corollary 7.5 we obtain asymptotic normality of the estimator and thus can construct asymptotic confidence intervals as illustrated in our experiments in Section 7.8.2.

Corollary 7.5. *Under the assumptions of Theorem 7.3, and the following three conditions:*

1. $\|\widehat{m}_a - m_a\| = o_{\mathbb{P}}(n^{-1/4})$,
2. $\|\widehat{\gamma} - \gamma\|_\infty = o_{\mathbb{P}}\left(n^{-1/(2(1+\alpha))}\right)$,
3. *For* $a \in \{0,1\}$, $\|\widehat{\pi}_a \widehat{f}_{a,\widehat{m}} - \pi_a f_{a,m}\| = O_{\mathbb{P}}(n^{-1/4})$,

we have that $\widehat{\psi}_{\widehat{d}^*,dr}$ *is root-n consistent and asymptotically normal with:*

$$\sqrt{n}(\widehat{\psi}_{\widehat{d}^*,dr} - \psi_{d^*}) \rightsquigarrow \mathcal{N}\left(0, \sigma_{d^*}^2\right).$$

As shown in Corollary 7.5, the error of $\widehat{\psi}_{\widehat{d}^*,\mathrm{dr}}$ shares the same convergence guarantee as $\widehat{\psi}_{d^*,\mathrm{dr}}$ under the extra assumption that the margin condition holds and the estimation error of $\widehat{\gamma}$ is $o_{\mathbb{P}}\left(n^{-1/(2(1+\alpha))}\right)$. This suggests that when the CMTE γ can be estimated well, or when the margin condition holds in a strong sense, the doubly robust-style estimator when based on the estimated policy $\widehat{d}^*$ behaves as if the true optimal policy d^* was instead plugged in. In Section 7.6.2, we give an algorithm that describes the estimation procedure for our proposed estimator.

7.6.2 Construction of the Estimators

The construction of the estimator contains two main steps: nuisance training and policy evaluation (with cross-fitting an additional possible step). Let (D_1, D_2, D_3) denote three independent samples of n observations of (X_i, A_i, Y_i).

Step 1 Nuisance training:

(a) Use D_1 to construct propensity score estimates $\widehat{\pi}_1$.

(b) Use D_1 to construct conditional median estimates $\widehat{m}_a$ for $a \in \{0, 1\}$, for example using quantile regression.

(c) Use D_2 to construct conditional density estimates at the estimated conditional median $\widehat{f}_{a,\widehat{m}}$, for example by regressing $\frac{1}{h}K\left(\frac{Y-\widehat{m}_a(X)}{h}\right)$ on X among those with $A = a$, for K a standard kernel function.

Step 2 Evaluate the value of the given fixed policy d or the learned optimal policy $\widehat{d}^*(X) = \mathbb{1}\{\widehat{\gamma}(X) > 0\}$ on D_3 through the estimator presented in (7.8).

Step 3 Cross-fitting (Optional): Repeat Step 1 and 2 two times by using the dataset in the order (D_1, D_3, D_2), and (D_2, D_3, D_1). Use the average of the resulting estimators as the final estimate of the ACME of the evaluated policy. We focus on three folds for simplicity, but alternatives using $k > 3$ folds are also possible.

In Section 7.8.2, we use an example to illustrate this estimation procedure. In the above we suggested quantile regression and a particular conditional density estimation procedure, but our convergence rate results show that any generic method could be used, as long as it satisfied the high-level L_2 error conditions listed there.

Remark 7.10. *In the above we use sample splitting to avoid empirical process conditions. Specifically, we split D_1 from D_2 in order to avoid conditions when estimating the conditional density $f_{a,m}$, and we split D_3 in order to avoid conditions in doing bias correction using the estimated influence function. This splitting could be omitted for if Donsker-type conditions are deemed acceptable.*

7.7 Estimation of the CMTE & Policy Learning

In this section, we propose efficient estimators for the conditional median treatment effect $\gamma(X) = m_1(X) - m_0(X)$, and use it to construct the median optimal treatment regime. Such an estimator is of independent interest, as it suggests ways to characterize the heterogeneity of the median treatment effect. After providing the doubly robust-style estimator $\widehat{\gamma}_{\text{dr}}$ in Section 7.7.1, we show how its estimation error is connected to the performance of the corresponding median optimal policy estimator $\widehat{d}^*_{\text{dr}}(X) = \mathbb{1}\{\widehat{\gamma}_{\text{dr}}(X) > 0\}$ (Section 7.7.2). Unlike existing work on estimating the marginal and conditional quantile treatment effect (Chernozhukov and Hansen, 2005; Dıéaz, 2017; Firpo, 2007; Fortin et al., 2011; Frölich and Melly, 2013; Machado and Mata, 2005; Melly, 2005; Rothe, 2010), our proposed nonparametric estimator for the CMTE relies on pseudo-outcome regression.

7.7.1 Doubly Robust-Style Estimator of the CMTE

Following the conditional average treatment effect estimation procedure proposed in Kennedy (2020), we construct the doubly robust-style estimator $\widehat{\gamma}_{\text{dr}}$ as follows: Let (D_1, D_2, D_3) denote three independent samples of n observations of $Z_i = (X_i, A_i, Y_i)$.

Step 1 Nuisance training: Same as Step 1 in Section 7.6.2.

Step 2 Pseudo-outcome regression: Use D_3 to construct the pseudo-outcome

$$\widehat{g}(Z) = \widehat{m}_1(X) - \widehat{m}_0(X) + \frac{A - \widehat{\pi}_1(X)}{\widehat{\pi}_1(X)\widehat{\pi}_0(X)} \frac{\frac{1}{2} - \mathbb{1}\{Y \leq \widehat{m}_A(X)\}}{\widehat{f}_{A,\widehat{m}}(X)},$$

and regress it on covariates X to obtain $\widehat{\gamma}_{\text{dr}}$. The regression estimator $\widehat{\mathbb{E}}_n$ is given by

$$\widehat{\gamma}_{\text{dr}}(x) = \widehat{\mathbb{E}}_n\{\widehat{g}(Z)|X = x\} = \sum_{i=1}^{n} w_i(x; X^n)\widehat{g}(Z_i), \tag{7.9}$$

where the weights $w_i(x; X^n)$ are learned using X^n in sample D_3. Examples of such linear smoothers $\widehat{\mathbb{E}}_n$ include kernel estimators, linear, ridge, local polynomial and RKHS regression, some random forests (e.g., Mondrian and kernel forests), as well as weighted combinations of aforementioned methods (Wasserman, 2006).

Step 3 Cross-fitting (Optional): Repeat Step 1 and 2 two times by using the dataset in the order (D_1, D_3, D_2), and (D_2, D_3, D_1). Use the average of the resulting estimators as the final $\widehat{\gamma}_{\text{dr}}$.

In a recent line of work (Nie and Wager, 2021; Kennedy, 2020), the conditional treatment effect estimators are commonly compared with an oracle estimator $\widetilde{\gamma}$ that has access to the true nuisance functions. We define the oracle estimator in our setting as follows.

Definition 7.2: Oracle

Denote $g(Z)$ to be the pseudo-outcome that depends on the true nuisance fucntions:

$$g(Z) = m_1(X) - m_0(X) + \frac{A - \pi_1(X)}{\pi_1(X)\pi_0(X)} \frac{1/2 - \mathbb{1}\{Y \leq m_A(X)\}}{f_{A,m}(X)}.$$

Given independent and identically distributed samples $\{X_i, O_i\}_{i=1}^n$ where $O_i = \gamma(X_i) + \varepsilon_i$ and ε_i is a mean-zero noise defined to be $\varepsilon_i = g(Z_i) - \gamma(X_i)$, the oracle is given by

$$\widetilde{\gamma}(x) = \widehat{\mathbb{E}}_n\{O|X = x\}.$$

In other words, the oracle $\widetilde{\gamma}$ regresses O on the covariates X using the same linear smoother $\widehat{\mathbb{E}}_n$ as the one used in $\widehat{\gamma}_{\text{dr}}$ (7.9). The performance of the oracle depends directly on the complexity (e.g., smoothness) of γ itself, since the outcome O is the sum of $\gamma(X)$ and a mean-zero noise.

We illustrate in the following theorem that the mean squared error of $\widehat{\gamma}_{\text{dr}}$ can be upper bounded by the oracle error incurred by $\widetilde{\gamma}$ and products of nuisance errors. This allows the CMTE to be estimated at a faster rate even when the nuisance estimates are obtained at slower rates.

Theorem 7.4

Let $\widehat{\gamma}_{\text{dr}}$ and $\widetilde{\gamma}$ denote the conditional median treatment effect estimator and oracle defined as above. Define the weighted norm $\|\cdot\|_w$ and squared-weighted norm $\|\cdot\|_{w^2}$ by

$$\|v\|_w^2 = \|v(x)\|_w^2 = \sum_{i=1}^n \frac{|w_i(x;X^n)|}{\sum_j |w_j(x;X^n)|} \int |v(z)|^2 d\mathbb{P}(z|X_i),$$
$$\|v\|_{w^2}^2 = \|v(x)\|_{w^2}^2 = \sum_{i=1}^n \frac{|w_i(x;X^n)|^2}{\sum_j |w_j(x;X^n)|^2} \int |v(z)|^2 d\mathbb{P}(z|X_i).$$

Assume

1. $\mathbb{P}\left(\epsilon \le \widehat{\pi}_1(X) \le 1-\epsilon\right) = 1$ for some $\epsilon \in (0,1)$,
2. $\mathbb{P}(M_0 \le \widehat{f}_{a,\widehat{m}}(X) \le M_1) = 1$ for $a \in \{0,1\}$ and $0 < M_0 \le M_1$.
3. $\text{Var}\{g(Z)|X=x\} \ge \sigma_{\min}^2$ for all $x \in \mathcal{X}$.

Then, for all $x \in \mathcal{X}$, we have

$$(\widehat{\gamma}_{\text{dr}}(x) - \gamma(x))^2 \lesssim (\widetilde{\gamma}(x) - \gamma(x))^2 + b(x) + O_{\mathbb{P}}\left(\|\widehat{g} - g\|_{w^2}^2 \mathbb{E}[(\widetilde{\gamma}(x) - \gamma(x))^2]\right), \quad (7.10)$$

where $b(x) = (\sum_{i=1}^n |w_i(x;X^n)|)^2 (\sum_{a=0}^1 \|\widehat{m}_a - m_a\|_w^2 (\|\widehat{\pi}_a \widehat{f}_{a,\widehat{m}} - \pi_a f_{a,m}\|_w + \|\widehat{m}_a - m_a\|_w))^2$.

Remark 7.11. *When the pseudo-outcome estimator $\widehat{g}$ is consistent in the squared-weighted norm, i.e., $\|\widehat{g} - g\|_{w^2} = o_{\mathbb{P}}(1)$, we have the third term in (7.10) to be $o_{\mathbb{P}}(\mathbb{E}[(\widetilde{\gamma}(x) - \gamma(x))^2])$.*

The assumptions of Theorem 7.4 require $g(Z)$ to have variation for all $X = x$. Our upper bound on the squared error of $\widehat{\gamma}_{\text{dr}}$ contains three terms: the first and last terms depend on the error of the oracle $\widetilde{\gamma}$, while the middle term relies on products of nuisance errors with respect to the weighted norm. When $\sum_j |w_j(x;X^n)| = 1$, the weighted norm $\|\cdot\|_w$ weighs the nuisance errors by $w_i(x;X^n)$ in a similar way to how the linear smoother $\widehat{\mathbb{E}}_n$ weighs the pseudo-outcomes. As an example, when the linear smoother $\widehat{\mathbb{E}}_n$ (e.g., nearest neighbor estimator) relies only on local information, these weighted norms ensure that the error at x is also weighed by only its local information. In the more general case, the weights for the weighted norms are normalized so that they sum up to 1. As shown in Stone (1977) and Györfi et al. (2002), $\sum_{i=1}^n |w_i(x;X^n)| = O_{\mathbb{P}}(1)$ is a sufficient condition to ensure $\widehat{\mathbb{E}}_n$ to be weakly universally consistent. In such cases, if $\|\widehat{g} - g\|_{w^2} = o_{\mathbb{P}}(1)$, then the squared error of $\widehat{\gamma}_{\text{dr}}$ deviates from the error of the oracle $\widetilde{\gamma}$ by products of nuisance errors, suggesting that the CMTE can be estimated at a faster rate compared to the nuisance functions. For a more detailed discussion on sufficient and necessary conditions on $w_i(x;X^n)$ for ensuring the consistency of $\widehat{\mathbb{E}}_n$, we refer the readers to (Stone, 1977).

7.7.2 Policy Learning

We briefly touch on the problem of learning the median optimal treatment regime. There are many approaches for policy learning. Among them, the two canonical ways are (1) empirical

value maximization where the goal is to directly find the optimal policy through maximizing the empirical value of policies in a fixed policy class (Zhao et al., 2012; Zhang et al., 2013; Athey and Wager, 2021); and (2) estimating the conditional treatment effect $\gamma(X)$ first and then plugging it into the closed form of the optimal treatment regime $d^*(X) = \mathbb{1}\{\gamma(X) > 0\}$ (Murphy, 2003; Robins, 2004; Laan and Luedtke, 2014). The second approach is closely related to plug-in classifiers (Audibert and Tsybakov, 2007), since deciding on the optimal treatment can be viewed as a binary classification task. Another related line of work is robust policy learning (Xiao et al., 2019; Zhang et al., 2021). Using a model-based approach, Xiao et al. (2019) learns conditional quantile treatment regimes through robust regression. On the other hand, Zhang et al. (2021) uses a heuristic two-stage nonparametric approach for robust policy learning. For a more comprehensive review on policy learning, we refer the readers to Athey and Wager (2021).

Using the doubly robust-style estimator $\widehat{\gamma}_{\text{dr}}$ presented in (7.9), we adopt the second approach and construct the median optimal treatment regime through $\widehat{d}^*_{\text{dr}}(X) = \mathbb{1}\{\widehat{\gamma}_{\text{dr}}(X) > 0\}$. We notice that the error of the policy $\widehat{d}^*_{\text{dr}}$ is bounded by the error of $\widehat{\gamma}_{\text{dr}}$. Under the assumption that for all $x \in \mathcal{X}$, either $|\gamma(x)| > \delta$ for some $\delta > 0$ or $\gamma(x) = 0$, we obtain that

$$\begin{aligned}\mathbb{E}[\mathbb{1}\{\widehat{d}^*_{\text{dr}}(x) \neq d^*(x)\}] &\leq \mathbb{P}(|\gamma(x)| \leq |\widehat{\gamma}_{\text{dr}}(x) - \gamma(x)|) \\ &\lesssim \mathbb{E}[|\widehat{\gamma}_{\text{dr}}(x) - \gamma(x)|] \leq \sqrt{\mathbb{E}[(\widehat{\gamma}_{\text{dr}}(x) - \gamma(x))^2]}.\end{aligned}$$

The first inequality follows from Lemma F.1 and the second inequality holds since if $\gamma(x) = 0$, then $d^*(x)$ can be either 1 or 0, suggesting that $\widehat{d}^*_{\text{dr}}(x)$ will always be optimal. The error of the learned policy $\widehat{d}^*_{\text{dr}}$ can be upper bounded by the square root of the expected squared error given in Theorem 7.4. The assumption used here is stronger than the margin condition. It is of future interest to relax such assumption and study the performance of $\widehat{d}^*_{\text{dr}}$ in more flexible settings.

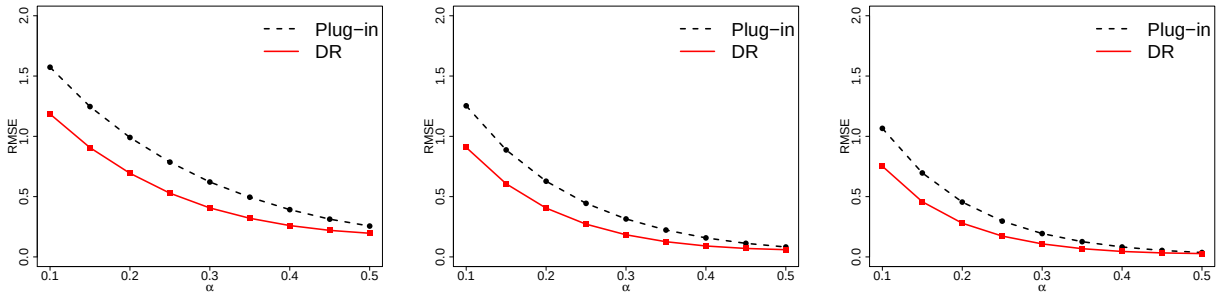
7.8 Experiments

7.8.1 Numerical Simulation

To explore the finite-sample properties of the estimator, we simulate from the following data generating process:

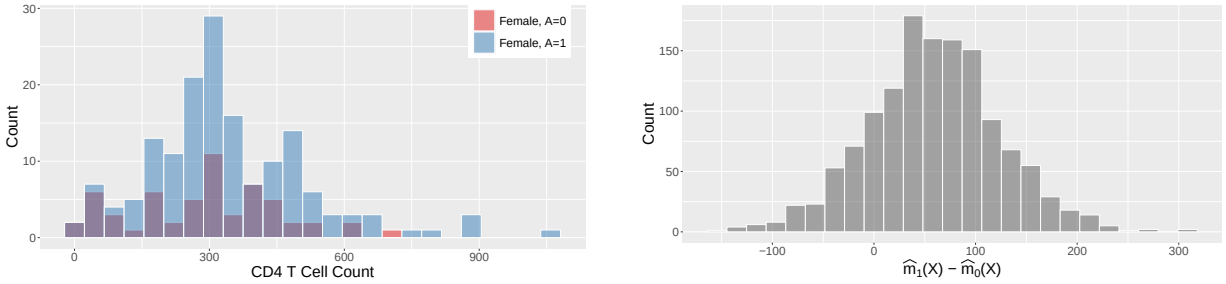
$$\begin{aligned}X &\sim \mathcal{N}(0, \mathbf{I}_5), \\ \text{logit}\{\pi_1(X)\} &= X^\top \beta \text{ where } \beta = (.2, .2, .2, .2, .2), \\ Y|X, A &\sim \text{Lognormal}(X^\top \beta + A, .25).\end{aligned}$$

We note that there is heterogeneity of the conditional median treatment effects across covariates X, i.e., $m_1(X) - m_0(X) = \exp(X^\top \beta + 1) - \exp(X^\top \beta) = (e-1)\exp(X^\top \beta)$. To inspect the rate of convergence of the estimator in terms of the nuisance estimation error, we constructed the nuisance estimators through the following procedure: $\widehat{\pi}_1(X) = \text{expit}\{\text{logit}(\pi_1(X)) + \epsilon_{1,n}\}$, $\widehat{m}_a(X) = m_a(X) + \epsilon_{2,n}$, and $\widehat{f}_{a,\widehat{m}}(X) = f_{a,m}(X) + \epsilon_{3,n}$ where $\epsilon_{1,n}, \epsilon_{2,n}, \epsilon_{3,n}$ are independent samples drawn from $\mathcal{N}(n^{-\alpha}, n^{-2\alpha})$. This construction ensures that the root mean square errors



(a) $n = 100$ **(b)** $n = 1000$ **(c)** $n = 5000$

Figure 7.2: Root mean square error of the doubly robust-style estimator (7.8) and the plug-in estimator (7.7) for settings described in Section 7.8.1. The sample size n varies in each subplot. Within a subplot, the x-axis is α, which controls the estimation error of the nuisances.



(a) Histogram of the outcomes for females. **(b)** Histogram of $\widehat{m}_1(X_i) - \widehat{m}_0(X_i)$.

Figure 7.3: 7.3a is the histogram of the outcomes for females under treatment and control. The difference in mean outcome under treatment versus control is 49, while the difference in median outcomes is only 1. 7.3b is the histogram of $\widehat{m}_1(X_i) - \widehat{m}_0(X_i)$ where $\widehat{m}_a$ is estimated using quantile regression forests described in Section 7.8.2.

of $\widehat{\pi}_a$, $\widehat{m}_a$ and $\widehat{f}_{a,\widehat{m}}$ are of order $O(n^{-\alpha})$. The policy used for evaluation is $d(X) = \mathbb{1}\{X_1 > 0\}$. Results shown in Figure 7.2 are averaged over 1000 rounds.

As we have seen in Figure 7.2, the doubly robust-style estimator converges faster compared to the plug-in estimator. In general, the plug-in estimator is more intuitive but often times, not optimal. When $n = 5000$, we see the estimation error of the doubly robust-style estimator is close to 0 when $\alpha = .25$, in line with what our theory suggests in Corollary 7.4.

7.8.2 Application: ACTG 175

We illustrate our proposed methods using the ACTG 175 dataset given in the R package `speff2trial` (Juraska et al., 2012). The data are from a randomized clinical trial in a population of adults with HIV type I. The treatment is binary where $A = 0$ stands for only using zidovudine as the therapy and $A = 1$ represents combination therapies. The outcome of interest is CD4 T cell count after 96 ± 5 weeks. Covariates X include baseline CD4 and CD8 T cell count, age, weight, Karnofsky score, indicators for race, gender, hemophilia, homosexual activity, drug

use, whether symptomatic, and previous zidovudine and antiretroviral use. Since the treatment is randomized, the propensity score is known to be $\pi_1(x) = .75$ for all subjects. The number of observations is $n = 1342$, after excluding subjects with missing outcomes.

Figure 7.3a shows a histogram of outcomes under treatment versus control for females. The presence of skewed responses in both histogram suggests that the median may be a more useful measure than the mean for characterizing both the treatment and the control conditional outcome distributions and for determining the optimal treatment in this study. In fact, although the mean outcome in this group is quite different under treatment versus control (mean CD4 count is 341 for treated females, but only 292 for controls), the medians are similar (313 for treated, 312 for control).[3]

Therefore we apply our proposed methods to estimate the value of the median optimal policy, and the value of a few competing policies. Specifically we use the estimator described in Section 7.6.2, splitting the sample into thirds (D_1, D_2 and D_3) and using cross-fitting. In D_1 the conditional median estimate $\widehat{m}_a$ is obtained with quantile regression forests via the package `quantregForest` (Meinshausen, 2006) (recall the propensity score is known and so does not need to be estimated here). D_2 is then used to construct the density estimate $\widehat{f}_{a,\widehat{m}}$, which we estimated by regressing a Gaussian kernel-weighted outcome centered at $\widehat{m}_a$ on X using the `randomForest` R package. The bandwidth h was chosen using Silverman's rule (Silverman, 1986). We considered estimating the ACME value of five policies using D_3: the observational policy $d(X_i) = A_i$, a plug-in median optimal policy $d(X_i) = \mathbb{1}\{\widehat{m}_1(X_i) > \widehat{m}_0(X_i)\}$, the treat-all policy $d(X_i) = 1$, the treat-none policy $d(X_i) = 0$, and a plug-in mean optimal policy $d(X_i) = \mathbb{1}\{\widehat{\mu}_1(X_i) > \widehat{\mu}_0(X_i)\}$ where the regression functions $\widehat{\mu}_a$ are also estimated via random forests.

Figure 7.4 shows the estimated values with the proposed doubly robust-style estimator $\widehat{\psi}_{d,\text{dr}}$, as well as the plug-in $\widehat{\psi}_{d,\text{pi}}$, for reference. 95% confidence intervals for $\widehat{\psi}_{d,\text{dr}}$ are obtained with the usual Wald interval based on the empirical variance $\widehat{\psi}_{d,\text{dr}} \pm 1.96\widehat{\sigma}_d/\sqrt{n}$ where $\widehat{\sigma}_d$ is the sample standard deviation of the influence function estimates. The results show the median optimal policy gives the highest value (342.40 with 95% CI 8.20), with the mean optimal policy and treat-all policy close behind. The treat-none policy does substantially worse than even the observational (random assignment) policy. Figure 7.3b shows a histogram of the conditional median treatment effects $\widehat{m}_1(X_i) - \widehat{m}_0(X_i)$, indicating a modest amount of effect heterogeneity.

7.9 Discussion

In this paper, we proposed a treatment policy based on conditional median treatment effects, and a new ACME value measure which is maximized for this policy. Importantly, our proposed approach avoids both within-group lack of robustness issues with the mean, as well as across-group unfairness issues with the marginal median. We argue that optimal treatment policies

[3]Connection to the margin condition: We note that the reported medians are only conditioned upon a single covariate (not the entire covariates), thus the the difference between these two values cannot be interpreted as the margin $\gamma(x)$. In addition, we cannot use a particular $\gamma(x)$ to infer the margin condition since the margin condition depends on $\gamma(x)$ for all $x \in \mathcal{X}$.

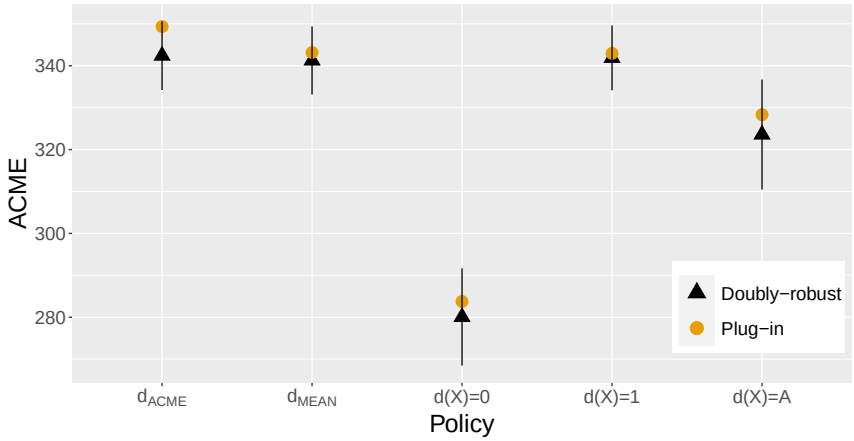


Figure 7.4: The results for the ACTG 175 data analysis. Black triangles show the estimated ACME values of five different policies, computed with our proposed estimator (7.8), and the black lines show the estimated 95% confidence intervals. For comparison, the yellow circles are estimates from the plug-in estimator (7.7).

should be defined in terms of the *conditional* distribution of outcomes, given covariates, rather than the marginal, in order to avoid across-group unfairness. We study nonparametric efficiency bounds and propose provably optimal doubly robust-style estimators, showing their finite-sample properties in simulations and in an illustration analyzing effects of combination therapy in treating HIV. We would generally argue that the mean is most useful as a measure of centrality when it is close to the median, and otherwise the median should be preferred; this suggests median effects should at least be used more widely than they are at present.

There are many opportunities for future related work. In addition to developing V-specific median optimal treatment regimes and analyzing $\widehat{d}^*_{\text{dr}}$ in more general settings, one could also consider more general (conditional) quantile optimal treatment regimes that replace conditional medians by conditional quantiles. All of this could also be adapted to numerous other settings, including continuous or time-varying treatments, or mediation problems, or settings where treatments may be confounded so that sensitivity analysis or instrumental variables may be used, etc.

Part III

Conclusion

Chapter 8

Conclusion

Building artificial intelligence systems from a human-centered perspective is an increasingly urgent subject, as large-scale machine learning models are deployed to interact with people daily. It is a thrilling time to work in this area as we are shaping the very form of artificial intelligence that will interact with many generations in the future. In this thesis, we ask the basic question:

> *Are there some guidelines one could use when building machine learning systems from a human-centered perspective?*

We propose a three-step guideline that involves (**S.1**) identifying the people and their characteristics we want to center the machine learning systems around; (**S.2**) modeling these characteristics in a reliable manner; and (**S.3**) incorporating these models into the design of the learning algorithm. To ground this guideline, we illustrate these steps in applications including recommender systems and decision-support systems. For recommender systems, we focus on users' evolving preferences, model them as dynamical systems (Chapter 2), and design online learning algorithms when facing such preference dynamics (Chapter 4, 5). For decision-support systems, we choose the decision-makers' risk preferences to be the core characteristics of concern, incorporate these preferences into the objective function of the machine learning models (Chapter 3), and propose a general recipe for learning models under diverse risk preferences (Chapter 6, 7).

While we have only covered two applications in this thesis, our proposed guideline can be used to establish human-centered machine learning research in a variety of settings, including mechanism design (Koster et al., 2022) and fine-tuning large language models (Stiennon et al., 2020). The development of this guideline provides a common ground for positioning different human-centered machine learning research, since if following the guideline, we will be clear about which people and their corresponding characteristics are the focus of the study, and articulate which aspect of the learning problem—the modeling of humans or the design of the machine learning algorithm in face of these human models or both—our research tackles.

In contextualizing our proposed guideline, we use insights and methodology from social sciences including behavioral economics and psychology to conduct human subject studies, model human preferences, and provide theoretical guarantees for different learning settings.

Our work is an example showcasing the interdisciplinary nature of human-centered machine learning research. We briefly discuss the interdisciplinarity of this field before outlining a future vision of human-centered machine learning.

Interdisciplinarity. Human-centered machine learning is naturally interdisciplinary. While its main focus is to build learning-based systems, the focal point for building these systems is human-centered, which deviates from the focus of traditional backbone disciplines—statistics and optimization—of machine learning. Thus, it is of crucial need for us to learn from and collaborate with researchers from other fields who have experience in taking a human-centered perspective. To make effective progress across different disciplines, we need to identify the role of each discipline in human-centered machine learning research. As one way of figuring this out, we may use the three-step guideline proposed in this thesis to specify the support that machine learning researchers need from other disciplines. Take human-computer interaction (HCI) as an example. On **S.1**, using human subject studies (e.g., human-in-the-loop evaluation of existing machine learning systems), HCI research can provide machine learning researchers with a more comprehensive understanding of which group of people and what their most relevant characteristics are for building the next generation of the system. On the hand, with respect to **S.2**, HCI research may help test the modeling assumptions on human preferences and behaviors that existing machine learning algorithms rely on. There are also aspects of human-centered machine learning that we have not touched upon here. We note that since the main focus of this thesis is to design learning algorithms from a human-centered perspective. we have not touched upon topics including the design of user interfaces for machine learning systems that some HCI research has focused on.

Future of Human-Centered Machine Learning

We are at the beginning of establishing how we could build machine learning systems from a human-centered perspective. There are many future directions that one could take, with the more concrete ones outlined at the end of each chapter. At a high level, there are three major directions that we envision the field to pursue.

The first direction centers around understanding the human factors that one should account for when developing machine learning models. Ideally, this should be uncovered by first hypothesizing what these factors could be and then conducting human-subject studies in either an experimental or an observational setting. For example, a human-subject study may be used to evaluate existing machine-learning systems and to figure out how the ideal system should behave. This line of research requires us to have an effective communication channel between the HCI and the machine learning community. With that, findings from the HCI community can inform machine learning researchers about the problem settings that require human-centered machine learning algorithm design; and machine learning researchers could provide machine learning models for HCI researchers to analyze.

The second direction is to build realistic, useful, potentially large-scale yet interpretable models for human preferences and behaviors. These models could be used to simulate human behaviors and feedback when building new machine learning systems and evaluating the impact

of existing systems. Currently, most theoretical research in analyzing machine learning systems from a human-centered perspective uses stylized models that capture a particular aspect of the people of interest. While these stylized models can help us highlight the algorithmic challenges in developing machine learning systems that interact with users of certain characteristics, it may be hard to understand the practical insights these models bring. On the other hand, recent economics and psychology research has proposed to use large language generative models to simulate human responses and behaviors. These models may also not be ideal as we do not have a clear path to verify how close the responses given by these models are to the ones given by actual human subjects of interest. The lack of guarantees for these models necessarily indicates that we cannot provide guarantees for the evaluation of machine learning systems conducted using these models. Building the ideal human preference and behavior model requires us to have representative and large-scale preference and behavior data from diverse groups of people. The diversity is crucial here as when using these human models to facilitate human-centered machine learning research, the role of the human subjects of interest may range from the users to the designers to the evaluators of the model. Once we have the ideal human models, we can build new personalized learning systems to interact with them and be evaluated by them.

Finally, although we have not touched upon this topic in the thesis, one growing need in the field is to develop public policies and supporting technologies for regulating and auditing models from a human-centered perspective. For example, in the context of recommender systems, we may require new auditing methods to inspect their impacts on the information received by the users; for large-language models, we may need new model unlearning techniques to fulfill the users' data deletion needs.

Appendices

Appendix A

Dynamic Preference: Additional Details

A.1 Algorithms

We provide additional details on algorithms presented in Section 2.3.1. Recall that $R(t)$ denote the reward obtained at time t and $a(t)$ denote the arm pulled at time t. Denote $\mathcal{T}_{k,t} = \{t' \in [t-1] : a(t') = k\}$. As we discussed in our main paper, the reward $R(t)$ obtained by pulling arm $a(t)$ is given by the enjoyment score the user provides after reading the recommended comic. Thus, for each arm (i.e., each comic genre), the reward distribution corresponds to user's enjoyment score distribution (or in other words, user's preference) towards that genre.

- UCB: For $t \in \{1, \ldots, K\}$, we have $a(t) = t$. For $t \in \{K+1, \ldots, T\}$, we have

$$a(t) \in \arg\max_{k \in [K]} \frac{1}{|\mathcal{T}_{k,t}|} \sum_{t \in \mathcal{T}_{k,t}} R(t) + \sqrt{\frac{2 \log t}{|\mathcal{T}_{k,t}|}}.$$

- TS: For each arm $k \in [K]$, their prior distribution is set to be a Dirichlet distribution with parameters being $(1, 1, 1, 1, 1, 1, 1, 1, 1)$, given that the rewards are categorical with 9 values. At time t, for each arm k, we obtain a virtual reward $\tilde{R}(k, t)$ by sampling from the corresponding Dirichlet distribution first and then use that sample to sample from a multinomial distribution. Then, $a(t) \in \arg\max_{k \in [K]} \tilde{R}(k, t)$. Upon a reward $R(t)$ is received by pulling arm k, the parameter of arm k's Dirichlet distribution is updated: the $R(t)$-th entry of the parameter is increased by 1. We have chosen the Dirichlet and multinomial distribution as the prior distribution and likelihood function because of their conjugacy, which ensures that the posterior distribution is tractable. This choice of the prior distribution and likelihood function has not utilized the fact that the reward is ordinal. One can potentially use a likelihood function that captures the ordinal structure of the rewards, but it may result in difficulties in obtaining the posterior distribution since there may not be a conjugate prior for this new likelihood function.

- ETC: For $t \in \{1, \ldots, \lfloor 0.5 \cdot (T^{2/3}) \rfloor\}$, $a(t) = t \bmod K$. For $t \in \{\lceil 0.5 \cdot (T^{2/3}) \rceil, \ldots, T\}$,

$$a(t) \in \arg\max \frac{1}{|\mathcal{T}_{k,t}|} \sum_{t \in \mathcal{T}_{k,t}} R(t).$$

- ε-Greedy: For $t \in \{1, \ldots, K\}$, $a(t) = t$. For $t \in \{K+1, \ldots, T\}$, with probability 0.9,

$$a(t) \in \arg\max \frac{1}{|\mathcal{T}_{k,t}|} \sum_{t \in \mathcal{T}_{k,t}} R(t).$$

  With probability 0.1, $a(t)$ is randomly selected from the K arms.

A.2 Holm's Sequential Bonferroni Procedure

When testing K hypotheses (in our case, we have one hypothesis for each arm), we adopt Holm's Sequential Bonferroni Procedure to control the family-wise error rate. To ensure that the probability of falsely rejecting *any* null hypothesis to be at most α, we perform the following procedure: Suppose that we are given m sorted p-values $p_1, \ldots, p_m$ from the lowest to the highest for hypothesis $H_1, \ldots, H_m$. For $i \in [m]$, if $p_i < \alpha/(m+1-i)$, then reject H_i and move on to the next hypothesis; otherwise, exit the process (and we cannot reject the rest of the hypotheses). As an illustration, in Table 2.2, since we are testing 5 hypothesis (one for each arm) at the same time and we have set the overall α level to be 0.1, to reject all null hypotheses, we require the lowest to the highest p-values to be no bigger than $0.02, 0.04, 0.06, 0.08, 0.1$. In other words, setting α at level 0.1 suggests that the probability of falsely rejecting *any* null hypothesis is at most 0.1. Under Bonferroni's correction, when rejecting the null hypothesis with the lowest p-value, the corrected alpha level is $\alpha/K = 0.02$.

A.3 Background Survey

We asked the study participants the following four questions:

- What is your age?
- How often do you read comic strips (e.g. newspaper comics, Sunday comics, ...)? Please pick an option that best describes you.
- Out of the following genres of comic strips, which genre do you find to be the most familiar?
- Out of the below genres of comic strips, which genre do you like the most?

The provided data are plotted as histograms in Figure A.1.

A.4 Implementation Details

A.4.1 Platform implementation

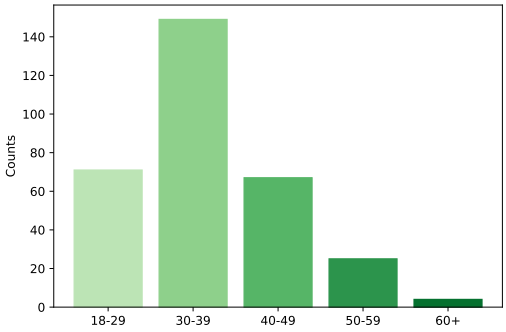
As described in Section 2.3.3, our platform consists of two main components: the participant-facing web interface and server backend. In this section, we provide more in-depth implementation details about both platform components.

Web interface. The participant-facing web interface is written using standard web development tools (HTML/CSS/Javascript) as well as the Flask web framework. When participants are performing the comic rating portion of the study, their responses are submitted in the background. Once the response is successfully recorded by the server, the images, rating slider bar, and button orderings are updated in-place so that users are not distracted by full page refreshes after each comic. The web interface also has a experimenter-defined variable to enforce a minimum time that must elapse between submissions. This ensures that participants spend an adequate amount of time on each comic; there is also a server-side mechanism to handle duplicate submissions if the button is clicked multiple times in quick succession. Once the comic reading portion is complete, the web interface will automatically transition to the post-study survey after users have read the end-of-study message.

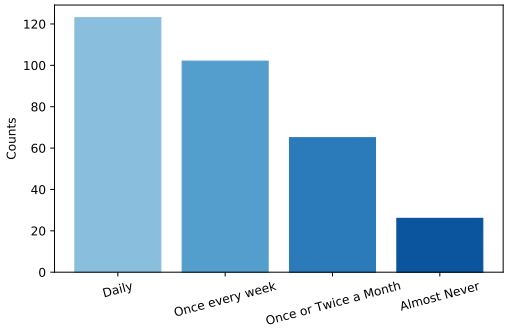
Server. We use the Flask web framework to implement the server and MongoDB for the database backend. The MAB algorithms are implemented in Python with a common interface that defines `get_arm()` and `update_arm()` functions, which can easily be extended to include other MAB algorithms that share this interface. The user accounts used by participants, the comic rating responses, post-study survey responses, and attention check answers are each stored in a separate database. All experiment configurations are prepared prior to launching the study. Experiments are configured using JSON files that enable both local and global control over experiment parameters, such as the number of comics to read, or the post-study survey and attention check questions asked. Participants can only access their assigned experiment, using credentials that are given when they register on the website. As participants are completing tasks in the study, each submitted task response is recorded in the database. The server can use the stored data to reload the session where it left off if the browser is closed or refreshed. This state includes that required for the algorithms, which are all executed on the server and require up-to-date parameters at each timestep of the study. Lastly, we host the server on Amazon Elastic Compute Cloud (EC2) using a `t2.xlarge` instance with 4 vCPUs and 16GB of memory.

A.4.2 MTurk implementation

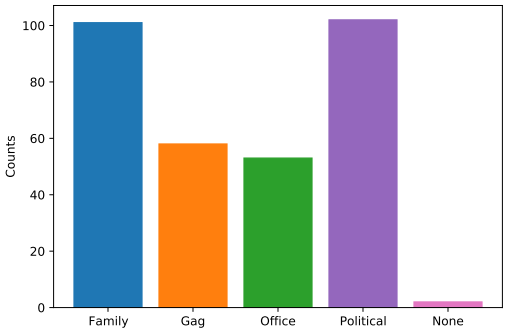
The study is split into two separate tasks on the MTurk platform. Participants must sign up for both tasks in order to proceed with the study. Once they have completed the initial background survey with a valid completion code, they must register their MTurk ID on our website. If their MTurk ID is found among the completed Qualtrics survey entries (retrieved through the MTurk query API), they are then assigned an account with a pre-loaded configuration. We also assign a unique survey code to each participant in the event to ensure that all parts of the study are



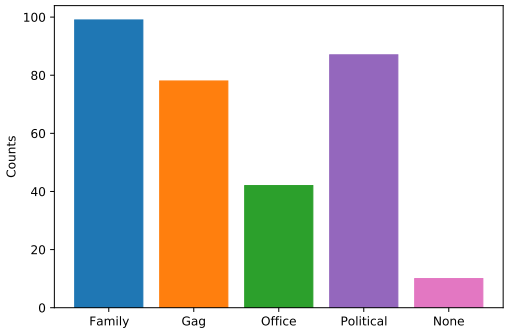
(a) Participant age distribution.



(b) How often participants read comics.



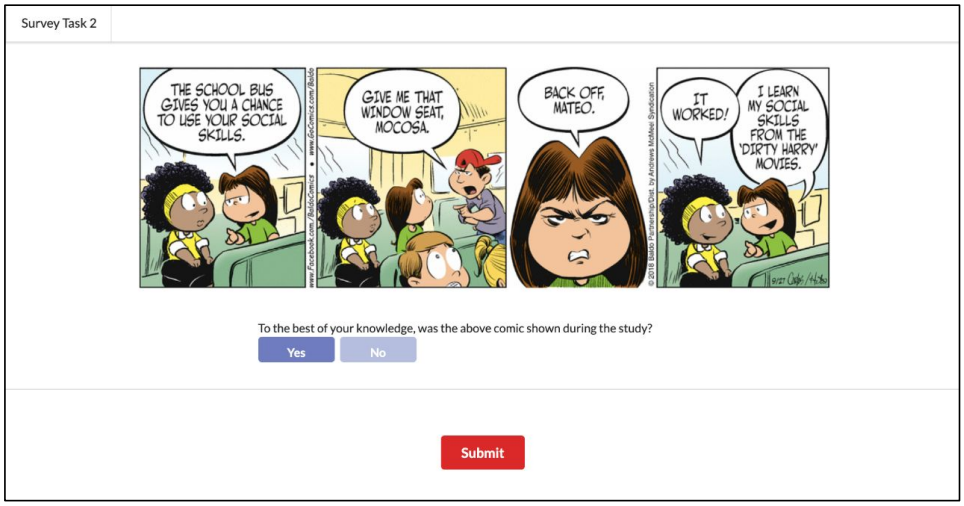
(c) Participants' most familiar category of comic.



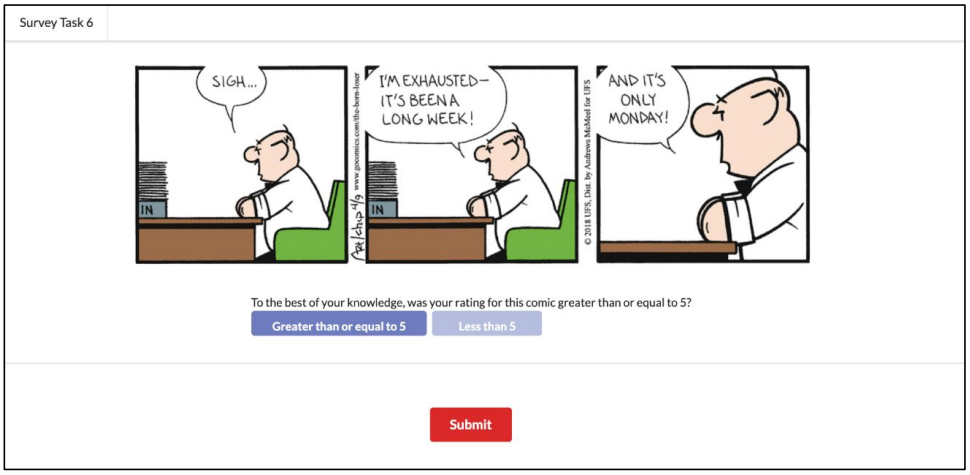
(d) Participants' most liked category of comic.

Figure A.1: Histograms of demographic information collected from the background survey.

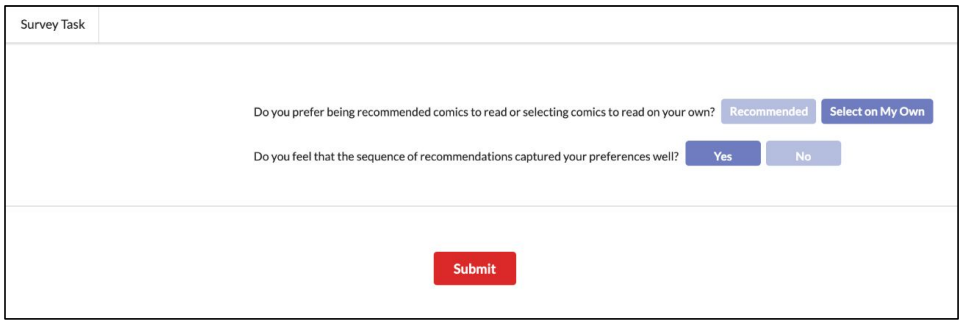
properly completed. Submissions that contain reused survey codes from other participants are considered invalid, and therefore not included in the final dataset.



(a) Testing whether the participant remembers if the given comic was in the study.



(b) Testing whether the participant remembers rating the given comic positively in the study.



(c) Final questions asked in the post-study survey.

Figure A.2: An example of questions contained in the post-study survey.

Appendix B

Human-Algined Risk Minimization: Additional Details

B.1 Proofs

Proof of Proposition 1. $w(F(x))$ satisfies: (i) non-decreasing since both $w(\cdot)$ and $F(\cdot)$ are non-decreasing; (ii) $\lim_{x\to+\infty} w(F(x)) = w\left(\lim_{x\to+\infty} F(x)\right) = 1$, $\lim_{x\to-\infty} w(F(x)) = w\left(\lim_{x\to-\infty} F(x)\right) = 0$; and (iii) right continuous since both $w(\cdot)$ and $F(\cdot)$ are continuous. Thus, as shown in Theorem 1.2.2 in (Durrett, 2019), $w(F(x))$ is a cumulative distribution function.

Proof of Lemma 3.1. First, notice that $w_{\text{POLY}}(P(E))' = \lim_{\Delta\to 0}\big(w_{\text{POLY}}(P(E)+\Delta; a, b) - w_{\text{POLY}}(P(E); a, b)\big)/\Delta$. Denote $P(E)$ to be y. Then, $w_{\text{POLY}}(y)' = \frac{3(3-3b)}{a^2-a+1}\big(y^2 - \frac{2(a+1)}{3}y + \frac{a}{3}\big) + 1 = \frac{3(3-3b)}{a^2-a+1}\big((y-\frac{a+1}{3})^2 + (\frac{a}{3} - \frac{(a+1)^2}{9})\big) + 1$. Let $u = |y - \frac{a+1}{3}|$. Then, $w'(y) = \frac{3(3-3b)}{a^2-a+1}\big(u^2 + (\frac{a}{3} - \frac{(a+1)^2}{9})\big) + 1 = f(u)$. $f'(u) = \frac{6(3-3b)}{a^2-a+1}u \geq 0$. Therefore, $f(u)$ is a monotonically increasing function of u. Thus, the lemma follows.

Proof of Lemma 3.2. Recall that we are given a family of models $\{\mathcal{M}_\theta \mid \theta \in [0,1]\}$ whose losses $\{\ell(\theta) \mid \theta \in [0,1]\}$ are parameterized by θ. For all $\theta \in [0,1]$, $\ell(\theta)$ follows a Bernoulli distribution:

$$P\left(\ell(\theta) = 1 - \left(\frac{1-\theta}{\theta}\right)^{1/2}\right) = \theta, P\left(\ell(\theta) = 1 + \left(\frac{\theta}{1-\theta}\right)^{1/2}\right) = 1-\theta.$$

The mean and variance of the losses are the same, i.e. $\mathbb{E}[\ell(\theta)] = 1$ and $\text{Var}(\ell(\theta)) = 1$. Thus, the skewness of $\ell(\theta)$ is

$$\text{Skewness}(\ell(\theta)) = \frac{2\theta - 1}{\sqrt{\theta(1-\theta)}}.$$

Denote $\alpha_\theta = 1 - \left(\frac{1-\theta}{\theta}\right)^{1/2}$, $\beta_\theta = 1 + \left(\frac{\theta}{1-\theta}\right)^{1/2}$ and $\alpha_\theta \leq 1 \leq \beta_\theta$.

$$R_H(\ell(\theta); w) = \alpha_\theta(w(\theta) - w(0)) + \beta_\theta(w(1) - w(\theta)) = \alpha_\theta w(\theta) + \beta_\theta(1 - w(\theta)).^{1}$$

Then, for different polynomial form of CPT probability weighting function w_1 and w_2,

$$R_H(\ell(\theta); w_1) - R_H(\ell(\theta); w_2) = (\beta_\theta - \alpha_\theta)(w_2(\theta) - w_1(\theta)).$$

Suppose that the probability weighting functions w_1, w_2 have the parametric form suggested by Equation 3.3 with parameter a_1, b_1 and a_2, b_2 respectively. If $a_1 = a_2 = \frac{1}{2}$ and $b_1 < b_2$, then $w_1(\theta) > w_2(\theta)$ on $[0, .5)$ and $w_1(\theta) < w_2(\theta)$ on $(.5, 1]$.

- If $0 \leq \theta < .5$, then $R_H(\theta; w_1) < R_H(\theta; w_2)$.
- If $.5 < \theta \leq 1$, then $R_H(\theta; w_2) < R_H(\theta; w_1)$.

Proof of Lemma 3.3. First, notice that $w_{\text{IT}}(0) = 0$ and $w_{\text{IT}}(1) = 1$. The fixed point of w_{IT} is $\frac{1}{2}$ because $w_{\text{IT}}(1/2) = 1/2$. Since $w'_{\text{IT}}(y) = -1/2 \ln(y \cdot (1-y)) > 0$ for all $x \in [0, 1]$, w_{IT} is monotonically increasing. Notice that $w''_{\text{IT}}(y) = \frac{2y-1}{2y(1-y)}$. Since $w''_{\text{IT}}(y) < 0$ for all $y \in [0, \frac{1}{2})$ and $w''_{\text{IT}}(y) > 0$ for all $y \in (\frac{1}{2}, 1]$, w'_{IT} is monotonically decreasing on $[0, \frac{1}{2})$ and monotonically increasing on $(\frac{1}{2}, 1]$. Thus, $w_{\text{IT}} \in \overline{\mathcal{W}}_{\text{CPT}}$.

Connection between w_{IT} and w_{POLY}. Let $y = F(\ell)$ and $w'_{\text{IT}}(y) = \frac{dw_{\text{IT}}(y)}{dy}$.

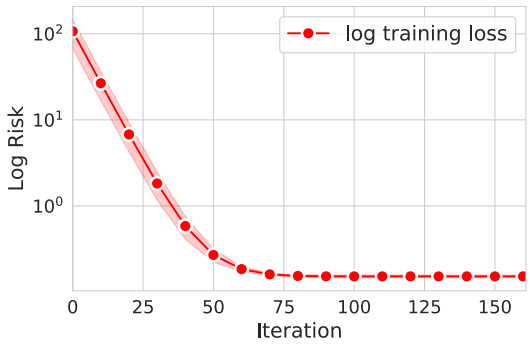
$$\begin{aligned}
w_{\text{POLY}}(y; 1/2, \ln 2) &= 4y^3 - 6y^2 + 3y + \left(-4 \ln 2 y^3 + 6 \ln 2 y^2 - 2 \ln 2 y\right) \\
&= 4y^3 - 6y^2 + 3y + g(y) \\
&\approx 4y^3 - 6y^2 + 3y + g(1/2) + g'(1/2)(y - 1/2) \\
&= 4y^3 - 6y^2 + 3y + (\ln 2y - \ln 2/2) \\
&= w_{\text{IT}}(1/2) + w'_{\text{IT}}(1/2)(y - 1/2) + \frac{w''_{\text{IT}}(1/2)}{2}(y - 1/2)^2 + \frac{w'''_{\text{IT}}(1/2)}{6}(y - 1/2)^2 \\
&\approx w_{\text{IT}}(y).
\end{aligned}$$

The first approximation is done by the first order Taylor expansion of $g(y)$ around $1/2$ and the second approximation is done through the third order Taylor expansion of w_{IT} around $1/2$.

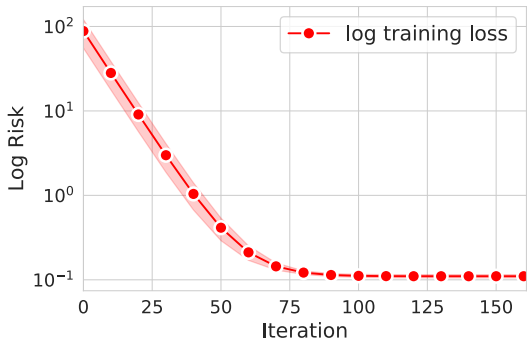
B.2 Optimization of EHRM

Optimally optimizing empirical human-aligned risk is an interesting open question. However, the heuristic approach described in Section 3.3 performs relatively well in the experiments. Figure B.1 and B.2 show the empirical human risk (at training time) of experiments in Section 3.5.1 and 3.5.3 respectively.

[1] The loss distribution is discrete in this case. We have used the CPT-weighted rank-dependent utility of a discrete random variable to obtain the human risk (Diecidue and Wakker, 2001).

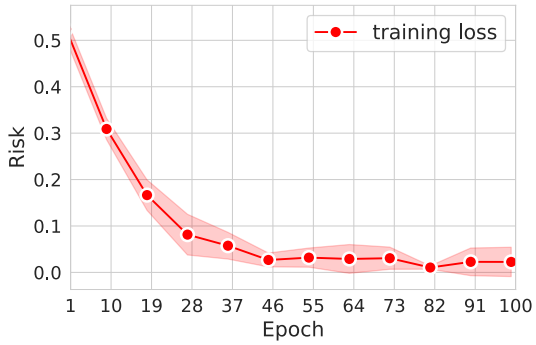


(a) $a = .5, b = .4$



(b) $a = .5, b = .8$

Figure B.1: Using a fixed learning rate .05 and optimization method described in Section 3.3, empirical human risk of the experiments in Section 3.5.1 converge within 100 iterations.



(a) $a = .5, b = .3$

Figure B.2: Using the optimization method described in Section 3.5.3, empirical human risk of the experiment (Section 3.5.3) converges in 100 epochs.

B.3 Fairness Metrics

Denote true positive rate as TPR, false positive rate as FPR, false negative rate as FNG, covariate as $X \in \mathcal{X}$ and label as $Y \in \{0, 1\}$. As suggested by (Bellamy et al., 2018), we define the below fairness metrics in terms of the privileged group $G_1 \subseteq \mathcal{X}$ and unprivileged group $G_2 \subseteq \mathcal{X}$:

1. Statistical Parity Difference: $P(Y = 1|X \in G_2) - P(Y = 1|X \in G_1)$.
2. Disparate Impact: $\frac{P(Y=1|X\in G_2)}{P(Y=1|X\in G_1)}$.
3. Equal Opportunity Difference: $\text{TPR}(G_2) - \text{TPR}(G_1)$.
4. Average Odds Difference: $\frac{1}{2}(\text{FPR}(G_2) - \text{FPR}(G_1) + (\text{TPR}(G_2) - \text{TPR}(G_1)))$.
5. Theil Index: $\frac{1}{n}\sum_{i=1}^{n} \frac{b_i}{\mu} \ln(\frac{b_i}{\mu})$ where $b_i = \widehat{Y}_i - Y_i + 1$ and $\mu = \frac{1}{n}\sum_{i=1}^{n} b_i$. $\widehat{Y}_i$ is the prediction of X_i and n is the number of samples.
6. False Negative Rate Difference: $\text{FNR}(G_2) - \text{FNR}(G_1)$.

B.4 Model Configuration

The model configuration for the gender classification task (Section 3.5.3) is as follows: 3 convolutional layers (with number of output channels $(6, 16, 16)$ respectively, kernel size $(5, 5, 6)$ respectively and a 2×2 max-pooling on the outputs of the first layer), followed by two fully connected layers (the first has 120 hidden units and the second is the output layer with 2 output units); all activation functions are ReLU and all convolutional layers use stride 1.

Appendix C

Rebounding Bandits: Additional Details

C.1 Integer Linear Programming Formulation

The bilinear integer program of (4.4) admits the following equivalent linear integer programming formulation:

$$
\begin{aligned}
\max_{u_{k,t}, z_{k,t,i}} \quad & \sum_{k \in [K]} \sum_{t \in [T]} b_k u_{k,t} - \lambda_k \sum_{i=0}^{t-1} \gamma_k^{t-i} z_{k,t,i} \\
\text{s.t.} \quad & \sum_{k \in [K]} u_{k,t} = 1, \qquad \forall t \in [T], \\
& z_{k,t,i} \leq u_{k,i}, \;\; z_{k,t,i} \leq u_{k,t}, \;\; u_{k,i} + u_{k,t} - 1 \leq z_{k,t,i}, \;\; \forall k \in [K], t \in [T], i \in \{0, \ldots, t-1\}, \\
& u_{k,t} \in \{0,1\}, \;\; u_{k,0} = 0, \qquad \forall k \in [K], t \in [T], \\
& z_{k,t,i} \in \{0,1\}, \qquad \forall k \in [K], t \in [T], i \in \{0, \ldots, t-1\}.
\end{aligned}
$$

C.2 Proofs and Discussion of Section 4.4

C.2.1 Proof of Lemma 4.1

Proof. When the expected rewards of all arms are the same, we know that the arm with the lowest index will be chosen and thus the first K pulls will be $\pi_1 = 1, \ldots, \pi_K = K$. We will complete the proof through induction. Suppose that the greedy pull sequence is periodic with $\pi_1 = 1, \ldots, \pi_K = K$ and $\pi_{t+K} = \pi_t$ until time $h > K$. We define k' to be h mod K and n to be $(h - k')/K$. We will show that $\pi_{h+1} = 1$ if $\pi_h = K$ and $\pi_{h+1} = \pi_h + 1$ otherwise. When $k' = 0$ (i.e., $\pi_h = K$), all arms have been pulled exactly n times as of time h. By the induction assumption, we know that $u_{1,1:h-K} = u_{2,2:h-K+1} = \ldots = u_{K,K:h}$, which implies that last time when each arm is pulled, all of them have the same expected rewards, i.e.,

$$\mu_{1,h-K+1}(u_{1,0:h-K}) = \mu_{2,h-K+2}(u_{2,0:h-K+1}) = \cdots = \mu_{K,h}(u_{K,0:h-1}).$$
$$\text{Moreover, } u_{1,h-K+1:h} = (1, \underbrace{0, \cdots 0}_{\text{K times}}), \quad u_{2,h-K+1:h} = (1, \underbrace{0, \cdots 0}_{\text{K-1 times}}), \quad \cdots, \quad u_{K,h:h} = (1).$$

Therefore, by (4.3), at time $h + 1$, arm 1 has the highest expected reward and will be chosen. In the case where $k' > 0$ (i.e., $\pi_h = k'$), we let $h' \coloneqq h - k'$. We have that $\mu_{1,h'-K+1}(u_{1,0:h'-K}) = \ldots = \mu_{K,h}(u_{K,0:h'-1})$ and $s = s_{1,h'-K+1}(u_{1,0:h'-K}) = \ldots = s_{K,h'}(u_{K,0:h'-1}) \leq \frac{\gamma^K}{1-\gamma^K}$. Then, at time $h + 1$, the satiation level for the arms will be $s_{k,h+1}(u_{k,0:h}) = \gamma^{k'-k+1}\left(1 + \gamma^K s\right)$ for all $k \leq k'$ and $s_{k,h+1}(u_{k,0:h}) = \gamma^{K-k+k'+1}s$ for all $k > k'$. Thus, the arm with the lowest satiation level will be $\pi_{h+1} = k' + 1 = \pi_h + 1$, since $s_{k'+1,h+1}(u_{k'+1,0:h}) < s_{1,h+1}(u_{1,0:h})$. Consequently, the greedy policy will select arm $\pi_h + 1$ at time $h + 1$. □

C.2.2 Proof of Theorem 4.1

Proof. First, when $T \leq K$, greedy policy is optimal since its cumulative expected reward is Tb. So, we consider the case of $T > K$. Assume for contradiction that there exists another policy $\pi^o_{1:T}$ that is optimal and is not greedy, i.e., $\exists t \in [T], \pi^o_t \notin \arg\max_{k\in[K]} b - \lambda s^o_{k,t}$ where $s^o_{k,t}$ denotes the satiation level of arm k at time t under the policy $\pi^o_{1:T}$. We will construct a new policy $\pi^n_{1:T}$ that obtains a higher cumulative expected reward than $\pi^o_{1:T}$. Throughout the proof, we use $s^n_{k,t}$ to denote the satiation levels for the new policy.

We first note two illustrative facts to give the intuition of the proof.

Fact 1: Any policy $\pi^o_{1:T}$ that does not pick the arm with the lowest satiation level (i.e., highest expected reward) at the last time step T is not optimal.
Proof of Fact 1: In this case, the policy $\pi^n_{1:T} = (\pi^o_1, \ldots, \pi^o_{T-1}, \pi_T)$ where $\pi_T \in \arg\max_{k\in[K]} b - \lambda s^o_{k,T}$ will obtain a higher cumulative expected reward.

Fact 2: If a policy $\pi^o_{1:T}$ picks the lowest satiation level for the final pull π^o_T but does not pick the arm with the lowest satiation level at time $T-1$, we claim that $\pi^n_{1:T} = (\pi^o_1, \ldots, \pi^o_{T-2}, \pi^o_T, \pi^o_{T-1}) \neq \pi^o_{1:T}$ obtains a higher cumulative expected reward.
Proof of Fact 2: First, note that $\pi^o_{T-1} \neq \pi^o_T$ because otherwise π^o_{T-1} is the arm with the lowest

satiation level at $T-1$. Moreover, at time $T-1$, $\pi^o_T \in \arg\min_k s^o_{k,T-1}$ has the smallest satiation, since if not, then there exists another arm $k \neq \pi^o_T$ and $k \neq \pi^o_{T-1}$ that has a smaller satiation level than π^o_T at time $T-1$. In that case, π^o_T will not be the arm with the lowest satiation at time T, which is a contradiction. Then, we deduce $s^o_{\pi^o_{T-1},T-1} > s^o_{\pi^o_T,T-1}$. Combining this with $\pi^o_{T-1} \neq \pi^o_T$, we arrive at

$$G_T(\pi^n_{1:T}) - G_T(\pi^o_{1:T}) = \lambda(1-\gamma)\left(s^o_{\pi^o_{T-1},T-1} - s^o_{\pi^o_T,T-1}\right) > 0.$$

For the general case, given any policy $\pi^o_{1:T}$ that is not a greedy policy, we construct the new policy $\pi^n_{1:T}$ that has a higher cumulative expected reward through the following procedure:

1. Find $t^* \in [T]$ such that for all $t > t^*$, $\pi^o_t \in \arg\max_{k\in[K]} b - \lambda s^o_{k,t}$ and $\pi^o_{t^*} \notin \arg\max_{k\in[K]} b - \lambda s^o_{k,t^*}$. Further, we know that $\pi^o_{t^*+1} \in \arg\max_{k\in[K]} b - \lambda s^o_{k,t^*}$, using the same reasoning as the above example, i.e., otherwise $\pi^o_{t^*+1} \notin \arg\max_{k\in[K]} b - \lambda s^o_{k,t^*+1}$. To ease the notation, we use k_1 to denote $\pi^o_{t^*}$ and k_2 to denote $\pi^o_{t^*+1}$.
2. For the new policy, we choose $\pi^n_{1:t^*+1} = (\pi^o_1, \ldots, \pi^o_{t^*-1}, k_2, k_1)$. Let $A^o_{t_1,t_2}$ denote the set $\{t' : t^*+2 \leq t' \leq t_2, \pi^o_{t'} = \pi^o_{t_1}\}$. $A^o_{t_1,t_2}$ contains a set of time indices in between t^*+2 and t_2 when arm $\pi^o_{t_1}$ is played under policy $\pi^o_{1:T}$. We construct the following three sets $T_A := \{t : t^*+2 \leq t \leq T, |A^o_{t^*,t}| < |A^o_{t^*+1,t}|\}$, $T_B := \{t : t^*+2 \leq t \leq T, |A^o_{t^*,t}| > |A^o_{t^*+1,t}|\}$ and $T_C := \{t : t^*+2 \leq t \leq T, |A^o_{t^*,t}| = |A^o_{t^*+1,t}|\}$. For time $t \geq t^*+2$, we consider the following three cases:

Case I. $T_B = \varnothing$, which means that at any time t in between t^*+2 and T, arm k_1 is played more than arm k_2 from t^*+2 to t. In this case, the new policy follows $\pi^n_{t^*+2:T} = \pi^o_{t^*+2:T}$.

Case II. $T_A = \varnothing$, which means that at any time t in between t^*+2 and T, arm k_2 is played more than arm k_1 from t^*+2 to t. In this case, the new policy satisfies: for all $t \geq t^*+2$, 1) $\pi^n_t = \pi^o_t$ if $\pi^o_t \neq k_1$ and $\pi^o_t \neq k_2$; 2) $\pi^n_t = k_2$ if $\pi^o_t = k_1$; and 3) $\pi^n_t = k_1$ if $\pi^o_t = k_2$.

Case III. $T_A \neq \varnothing$ and $T_B \neq \varnothing$. Then, starting from t^*+2, if $t \in T_A$, π^n_t follows the new policy construction in Case I, i.e., $\pi^n_t = \pi^o_t$. If $t \in T_B$, π^n_t follows the new policy construction in Case II. Finally, for all $t \in T_C$, define $t'_{A,t} = \max_{\substack{t'\in T_A:\\ t'<t}} t'$ and $t'_{B,t} = \max_{\substack{t'\in T_B:\\ t'<t}} t'$. If $t'_{A,t} > t'_{B,t}$, then π^n_t follows the new policy construction as Case I. If $t'_{A,t} < t'_{B,t}$, π^n_t follows the new policy construction as Case II. We note that $t'_{A,t} \neq t'_{B,t}$ since $T_A \cap T_B = \varnothing$.

When $T_A = \varnothing$ and $T_B = \varnothing$, we know that k_1 and k_2 are not played in $\pi^o_{t^*+2:T}$. In this case, the new policy construction can follow either Case I or Case II. To complete the proof, we state some facts first:

- From t^*, the expected rewards collected by the policies $\pi^o_{1:T}$ and $\pi^n_{1:T}$ only differ at times when arm k_1 or arm k_2 is played.
- $\pi^n_{1:t^*+1}$ obtains a higher cumulative expected reward than $\pi^o_{1:t^*+1}$.

- At time t^*+2, the new policy follows that $s^n_{k_1,t^*+2} = \gamma+\gamma^2 s^o_{k_1,t^*}$ and $s^n_{k_2,t^*+2} = \gamma^2+\gamma^2 s^o_{k_2,t^*}$. On the other hand, the old policy has $s^o_{k_1,t^*+2} = \gamma^2 + \gamma^2 s^o_{k_1,t^*}$ and $s^o_{k_2,t^*+2} = \gamma + \gamma^2 s^o_{k_2,t^*}$.

Let $N_{k_1} := \{t : t^*+2 \le t \le T, \pi^o_t = k_1\}$ and $N_{k_2} := \{t : t^*+2 \le t \le T, \pi^o_t = k_2\}$ denote the sets of time steps when k_1 and k_2 are played in $\pi^o_{1:T}$. For a given satiation level x at time t' together with the time steps the arm is pulled N_k, we have that at time $t \ge t'$, the arm has satiation level $g_{N_k}(x,t,t') = \gamma^{t-t'}x + \sum_{N_{k,i}<t} \gamma^{t-N_{k,i}}$ where $N_{k,i}$ is the i-th smallest element in N_k.

In Case I, the difference of the cumulative expected rewards between the two policies satisfies:

$$\begin{aligned} G_T(\pi^n_{1:T}) - G_T(\pi^o_{1:T}) &> \sum_{i=1}^{|N_{k_2}|} -\lambda g_{N_{k_2}}(s^n_{k_2,t^*+2}, N_{k_2,i}, t^*+2) + \lambda g_{N_{k_2}}(s^o_{k_2,t^*+2}, N_{k_2,i}, t^*+2) \\ &\quad + \sum_{j=1}^{|N_{k_1}|} -\lambda g_{N_{k_1}}(s^n_{k_1,t^*+2}, N_{k_1,j}, t^*+2) + \lambda g_{N_{k_1}}(s^o_{k_1,t^*+2}, N_{k_1,j}, t^*+2) \\ &= \lambda\left(s^o_{k_2,t^*+2} - s^n_{k_2,t^*+2}\right)\sum_{i=1}^{|N_{k_2}|}\gamma^{N_{k_2,i}-(t^*+2)} + \lambda\left(s^o_{k_1,t^*+2} - s^n_{k_1,t^*+2}\right)\sum_{j=1}^{|N_{k_1}|}\gamma^{N_{k_1,j}-(t^*+2)} > 0, \end{aligned}$$

where we have used the fact that $s^o_{k_2,t^*+2} - s^n_{k_2,t^*+2} = -\left(s^o_{k_1,t^*+2} - s^n_{k_1,t^*+2}\right) > 0$, $|N_{k_2}| \ge |N_{k_1}|$ and for all $j \in [|N_{k_1}|]$, $N_{k_2,j} < N_{k_1,j}$. In Case II, similarly, we have that

$$\begin{aligned} G_T(\pi^n_{1:T}) - G_T(\pi^o_{1:T}) &> \sum_{j=1}^{|N_{k_1}|} -\lambda g_{N_{k_1}}(s^n_{k_2,t^*+2}, N_{k_1,j}, t^*+2) + \lambda g_{N_{k_1}}(s^o_{k_1,t^*+2}, N_{k_1,j}, t^*+2) \\ &\quad + \sum_{i=1}^{|N_{k_2}|} -\lambda g_{N_{k_2}}(s^n_{k_1,t^*+2}, N_{k_2,i}, t^*+2) + \lambda g_{N_{k_2}}(s^o_{k_2,t^*+2}, N_{k_2,i}, t^*+2) \\ &= \lambda\left(s^o_{k_1,t^*+2} - s^n_{k_2,t^*+2}\right)\sum_{j=1}^{|N_{k_1}|}\gamma^{N_{k_1,j}-(t^*+2)} + \lambda\left(s^o_{k_2,t^*+2} - s^n_{k_1,t^*+2}\right)\sum_{i=1}^{|N_{k_2}|}\gamma^{N_{k_2,i}-(t^*+2)} > 0, \end{aligned}$$

since $s^o_{k_1,t^*+2} - s^n_{k_2,t^*+2} = -\left(s^o_{k_2,t^*+2} - s^n_{k_1,t^*+2}\right) > 0$, $|N_{k_2}| \le |N_{k_1}|$ and for all $i \in [|N_{k_2}|]$, $N_{k_1,i} < N_{k_2,i}$.

Finally, for Case III, the new policy construction is a mix of Case I and Case II. We represent the time interval $[t^*+2, T]$ to be $[t^*+2, T] = [t_{i_1,s_1}, t_{i_1,e_1}] \cup [t_{i_2,s_2}, t_{i_2,e_2}] \cup \cdots \cup [t_{i_M,s_M}, t_{i_M,e_M}]$ where $t^*+2 = t_{i_1,s_1} \le \ldots \le t_{i_M,s_M} = T$, $\cap_{m=1}^M [t_{i_m,s_m}, t_{i_m,e_m}] = \varnothing$ and $M-1$ is the number of new policy construction switches happen in between t^*+2 and T. We say that a new policy construction switch happens at time t if the policy construction follows Case I at time $t-1$ but follows Case II at time t or vice versa. Each $i_m \ne i_{m-1}$ can take values I or II, representing which policy construction rule is used between the time period t_{i_m,s_m} and t_{i_m,e_m}. For any time index set V, we use the notation $V[t_{i_m,s_m}, t_{i_m,e_m}] \coloneqq \{t \in V : t_{i_m,s_m} \le t \le t_{i_m,e_m}\}$.

We notice that at any switching time t_{i_m,s_m}, the number of previous pulls of arm k_1 and k_2 from time $t_{i_{m-1},s_{m-1}}$ to $t_{i_{m-1},e_{m-1}}$ are equivalent, which is denoted by $l_m = |N_{k_1}[t_{i_m,s_m}, t_{i_m,e_m}]| =$

$|N_{k_2}[t_{i_m,s_m}, t_{i_m,e_m}]|$ for all $m < M$. From our analysis of Case I and Case II, we know that to show that $\pi^n_{1:T}$ obtains a higher cumulative expected reward, it suffices to prove: for all $m < M$ such that

$$s^o_{k_2,t_{i_m,s_m}} - s^n_{k_2,t_{i_m,s_m}} = -\left(s^o_{k_1,t_{i_m,s_m}} - s^n_{k_1,t_{i_m,s_m}}\right) > 0,$$
$$s^o_{k_1,t_{i_m,s_m}} - s^n_{k_2,t_{i_m,s_m}} = -\left(s^o_{k_2,t_{i_m,s_m}} - s^n_{k_1,t_{i_m,s_m}}\right) > 0,$$

we have

$$s^o_{k_2,t_{i_{m+1},s_{m+1}}} - s^n_{k_2,t_{i_{m+1},s_{m+1}}} = -\left(s^o_{k_1,t_{i_{m+1},s_{m+1}}} - s^n_{k_1,t_{i_{m+1},s_{m+1}}}\right) > 0,$$
$$s^o_{k_1,t_{i_{m+1},s_{m+1}}} - s^n_{k_2,t_{i_{m+1},s_{m+1}}} = -\left(s^o_{k_2,t_{i_{m+1},s_{m+1}}} - s^n_{k_1,t_{i_{m+1},s_{m+1}}}\right) > 0.$$

We will establish these facts in Lemma C.1. Finally, we note that the above required conditions are held at time $t_{i_1,s_1} = t^* + 2$. □

Lemma C.1

Let $N_k[t_s, t_e]$ denote the set of time steps when arm k is pulled in between (and including) time t_s and t_e under policy $\pi^o_{1:T}$. Let $s^o_{k,t}$ and $s^n_{k,t}$ represent the satiation level of arm k at time t when following the policy $\pi^o_{1:T}$ and $\pi^n_{1:T}$, respectively. For two different arms k_1 and k_2, suppose that at time t_s we have

$$s^o_{k_2,t_s} - s^n_{k_2,t_s} = -\left(s^o_{k_1,t_s} - s^n_{k_1,t_s}\right) > 0,$$
$$s^o_{k_1,t_s} - s^n_{k_2,t_s} = -\left(s^o_{k_2,t_s} - s^n_{k_1,t_s}\right) > 0.$$

Further, suppose that from time t_s to t_e, $\pi^n_{1:T}$ follows either Case I (or Case II) of new policy construction (see proof of Theorem 4.1 for their definitions); and at time $t'_s = t_e + 1$, the new policy construction for $\pi^n_{1:T}$ has switched to Case II (or Case I if Case II is used from t_s to t_e). Then at time t'_s, we have that

$$s^o_{k_2,t'_s} - s^n_{k_2,t'_s} = -\left(s^o_{k_1,t'_s} - s^n_{k_1,t'_s}\right) > 0,$$
$$s^o_{k_1,t'_s} - s^n_{k_2,t'_s} = -\left(s^o_{k_2,t'_s} - s^n_{k_1,t'_s}\right) > 0.$$

Proof of Lemma C.1. Following the definition in the proof of Theorem 4.1, given that at time t_s, arm k has satiation s, let $g_{N_k[t_s,t_e]}(s, t'_s, t_s)$ denote the satiation level of arm k at time t'_s after being pulled at the time steps in the set $N_k[t_s, t_e]$. Let $N_{k,i}[t_s, t_e]$ be the i-th smallest element in the set $N_k[t_s, t_e]$. From the definition of the new policy construction given in the proof of Theorem 4.1, we also know that (1) $N := |N_{k_1}[t_s, t_e]| = |N_{k_2}[t_s, t_e]|$; (2) if Case I is applied in between t_s and t_e, we have that for all $i \in [N]$, $N_{k_2,i}[t_s, t_e] < N_{k_1,i}[t_s, t_e]$; and (3) if Case II is applied in between t_s and t_e, we have that for all $i \in [N]$, $N_{k_2,i}[t_s, t_e] > N_{k_1,i}[t_s, t_e]$.

We first consider the setting when Case I new policy construction is applied, then at time t'_s, we can show that

$$\begin{aligned} s^o_{k_1,t'_s} - s^n_{k_2,t'_s} =& g_{N_{k_1}[t_s,t_e]}\left(s^o_{k_1,t_s}, t'_s, t_s\right) - g_{N_{k_2}[t_s,t_e]}\left(s^n_{k_2,t_s}, t'_s, t_s\right) \\ =& \gamma^{t'_s - t_s}\left(s^o_{k_1,t_s} - s^n_{k_2,t_s}\right) + \sum_{i=1}^{l} \gamma^{t'_s - N_{k_1,i}[t_s,t_e]} - \gamma^{t'_s - N_{k_2,i}[t_s,t_e]} \\ =& \gamma^{t'_s - t_s}\left(s^n_{k_1,t_s} - s^o_{k_2,t_s}\right) + \sum_{i=1}^{l} \gamma^{t'_s - N_{k_1,i}[t_s,t_e]} - \gamma^{t'_s - N_{k_2,i}[t_s,t_e]} \\ =& s^n_{k_1,t'_s} - s^o_{k_2,t'_s} > 0, \end{aligned}$$

where the last inequality has used the fact that when we use Case I construction, we have $N_{k_2,i}[t_s,t_e] < N_{k_1,i}[t_s,t_e]$. Meanwhile, we also have that

$$\begin{aligned} s^o_{k_2,t'_s} - s^n_{k_2,t'_s} =& g_{N_{k_2}[t_s,t_e]}\left(s^o_{k_2,t_s}, t'_s, t_s\right) - g_{N_{k_2}[t_s,t_e]}\left(s^n_{k_2,t_s}, t'_s, t_s\right) \\ =& \gamma^{t'_s - t_s}\left(s^o_{k_2,t_s} - s^n_{k_2,t_s}\right) = -\gamma^{t'_s - t_s}\left(s^o_{k_1,t_s} - s^n_{k_1,t_s}\right) \\ =& -\left(s^o_{k_1,t'_s} - s^n_{k_1,t'_s}\right) > 0. \end{aligned}$$

When Case II new policy construction is applied, then at time t'_s, we get

$$\begin{aligned} s^o_{k_1,t'_s} - s^n_{k_2,t'_s} =& g_{N_{k_1}[t_s,t_e]}\left(s^o_{k_1,t_s}, t'_s, t_s\right) - g_{N_{k_1}[t_s,t_e]}\left(s^n_{k_2,t_s}, t'_s, t_s\right) \\ =& \gamma^{t'_s - t_s}\left(s^o_{k_1,t_s} - s^n_{k_2,t_s}\right) = -\gamma^{t'_s - t_s}\left(s^o_{k_2,t_s} - s^n_{k_1,t_s}\right) \\ =& -\left(s^o_{k_2,t'_s} - s^n_{k_1,t'_s}\right) > 0, \end{aligned}$$

since $s^o_{k_1,t_s} - s^n_{k_2,t_s} > 0$. On the other hand, we have that

$$\begin{aligned} s^o_{k_2,t'_s} - s^n_{k_2,t'_s} =& g_{N_{k_2}[t_s,t_e]}\left(s^o_{k_2,t_s}, t'_s, t_s\right) - g_{N_{k_1}[t_s,t_e]}\left(s^n_{k_2,t_s}, t'_s, t_s\right) \\ =& \gamma^{t'_s - t_s}\left(s^o_{k_2,t_s} - s^n_{k_2,t_s}\right) + \sum_{i=1}^{l} \gamma^{t'_s - N_{k_2,i}[t_s,t_e]} - \gamma^{t'_s - N_{k_1,i}[t_s,t_e]} \\ =& \gamma^{t'_s - t_s}\left(s^n_{k_1,t_s} - s^o_{k_1,t_s}\right) + \sum_{i=1}^{l} \gamma^{t'_s - N_{k_2,i}[t_s,t_e]} - \gamma^{t'_s - N_{k_1,i}[t_s,t_e]} \\ =& s^n_{k_1,t'_s} - s^o_{k_1,t'_s} > 0, \end{aligned}$$

where the last inequality is true because when Case II new policy construction is applied, we have $N_{k_1,i}[t_s,t_e] < N_{k_2,i}[t_s,t_e]$. □

C.2.3 Proof of Proposition 4.1

Proof. If $T \leq K$, a Max K-Cut of $\mathcal{K}_T$ is $\forall k \in [T], P_k = \{k\}$, which is the same as an optimal solution to (4.4). Let $\mathbf{1}\{\cdot\}$ denote the indicator function. When $T > K$, the integer program

in (4.4) is equivalent to

$$
\begin{aligned}
&\max_{\substack{u_{k,t}\in\{0,1\}:\\ \forall t\in[T],\sum_k u_{k,t}=1}} \sum_{k=1}^{K} b u_{k,1} + \sum_{k=1}^{K}\sum_{t=2}^{T}\left(b u_{k,t} - \lambda\sum_{i=1}^{t-1}\gamma^{t-i}u_{k,i}u_{k,t}\right)\\
=&\max_{\substack{P_1,\ldots,P_K\subseteq[T]:\\ \cup_k P_k=[T],\\ \forall k\neq k', P_k\cap P_{k'}=\varnothing}} \sum_{k=1}^{K} b\mathbf{1}\{1\in P_k\} + \sum_{k=1}^{K}\sum_{t=2}^{T}\left(b\mathbf{1}\{t\in P_k\} - \lambda\sum_{i=1}^{t-1}\gamma^{t-i}\mathbf{1}\{i\in P_k\}\mathbf{1}\{t\in P_k\}\right)\\
=&\max_{\substack{P_1,\ldots,P_K\subseteq[T]:\\ \cup_k P_k=[T],\\ \forall k\neq k', P_k\cap P_{k'}=\varnothing}} Tb - \sum_{k=1}^{K}\sum_{\substack{t,i\in P_k:\\ i<t}}\lambda\gamma^{t-i}\\
=&Tb - \sum_{t=2}^{T}\sum_{i=1}^{t-1}\lambda\gamma^{t-i} + \max_{\substack{P_1,\ldots,P_K\subseteq[T]:\\ \cup_k P_k=[T],\\ \forall k\neq k', P_k\cap P_{k'}=\varnothing}} \sum_{k=1}^{K-1}\sum_{k'=k+1}^{K}\sum_{\substack{t\in P_k,\\ i\in P_{k'}:\\ i<t}}\lambda\gamma^{t-i},
\end{aligned}
$$

where the second equality uses the fact $\sum_{k=1}^{K}\mathbf{1}\{t\in P_k\}=1$ for all $t\in[T]$ and the third equality is true because for any $P_1,\ldots P_K$ such that $\forall k\neq k'$, $P_k\cap P_{k'}=\varnothing$ and $\cup_k P_k=[T]$, we have

$$
\text{Total Edge Weights of } \mathcal{K}_T = \sum_{t=2}^{T}\sum_{i=1}^{t-1} e(t,i) = \sum_{\substack{t,i\in[T]:i<t,\\ \exists k\in[K],i,t\in P_k}} e(t,i) \quad + \sum_{\substack{t,i\in[T]:i<t,\\ \forall k\in[K],i,t\notin P_k}} e(t,i).
$$

□

C.2.4 Proof of Theorem 4.2

Proof. Given $\pi^*_{1:T}$ and $\pi^w_{1:T}$, define a set of new policies $\{\tilde{\pi}^i_{1:T}\}_{i=1}^{l-1}$ such that for all i, $\tilde{\pi}^i_{1:T} = (\pi^w_{1:iw}, \pi^*_{iw+1:T})$. Based on this, we have the following decomposition

$$
\begin{aligned}
G_T(\pi^*_{1:T}) - G_T(\pi^w_{1:T}) = &\underbrace{G_T(\pi^*_{1:T}) - G_T(\tilde{\pi}^1_{1:T})}_{A_0} + \left(\sum_{i=1}^{l-2}\underbrace{G_T(\tilde{\pi}^i_{1:T}) - G_T(\tilde{\pi}^{i+1}_{1:T})}_{A_i}\right)\\
&+ \underbrace{G_T(\tilde{\pi}^{l-1}_{1:T}) - G_T(\pi^w_{1:T})}_{A_{l-1}}.
\end{aligned}
$$

To distinguish the past pull sequences of each arm under different policies, we use the following notations: $\mu_{k,t}(u_{k,0:t-1};\pi')$ gives the expected reward of arm k at time t by following pull sequence $\pi'_{1:t-1}$. By the definition of $\pi^w_{1:T}$, we have that

$$
A_0 = \sum_{t=1}^{w}\mu_{\pi^*_t,t}(u_{\pi^*_t,0:t-1};\pi^*) - \mu_{\pi^w_t,t}(u_{\pi^w_t,0:t-1};\pi^w) + \sum_{t=w+1}^{T}\mu_{\pi^*_t,t}(u_{\pi^*_t,0:t-1};\pi^*) - \mu_{\pi^*_t,t}(u_{\pi^*_t,0:t-1};\tilde{\pi}^1)
$$

$$\leq \sum_{t=w+1}^{T} \mu_{\pi_t^*,t}(u_{\pi_t^*,0:t-1};\pi^*) - \mu_{\pi_t^*,t}(u_{\pi_t^*,0:t-1};\tilde{\pi}^1),$$

where the inequality follows from the fact that $\pi_{1:w}^w$ is optimal for (4.4) when $T = w$. Similarly, we obtain that for all $i \in [l-2]$,

$$\begin{aligned} A_i &= \sum_{t=1}^{iw} \underbrace{\mu_{\pi_t^w,t}(u_{\pi_t^w,0:t-1};\pi^w) - \mu_{\pi_t^w,t}(u_{\pi_t^w,0:t-1};\pi^w)}_{=0} + \sum_{t=iw+1}^{(i+1)w} \underbrace{\mu_{\pi_t^*,t}(u_{\pi_t^*,0:t-1};\tilde{\pi}^i) - \mu_{\pi_t^w,t}(u_{\pi_t^w,0:t-1};\pi^w)}_{\leq 0} \\ &\quad + \sum_{t=(i+1)w+1}^{T} \mu_{\pi_t^*,t}(u_{\pi_t^*,0:t-1};\tilde{\pi}^i) - \mu_{\pi_t^*,t}(u_{\pi_t^*,0:t-1};\tilde{\pi}^{i+1}) \\ &\leq \sum_{t=(i+1)w+1}^{T} \mu_{\pi_t^*,t}(u_{\pi_t^*,0:t-1};\tilde{\pi}^i) - \mu_{\pi_t^*,t}(u_{\pi_t^*,0:t-1};\tilde{\pi}^{i+1}). \end{aligned}$$

Finally, we have $A_{l-1} = \sum_{t=(l-1)w+1}^{T} \mu_{\pi_t^*,t}(u_{\pi_t^*,0:t-1};\tilde{\pi}^{l-1}) - \mu_{\pi_t^w,t}(u_{\pi_t^w,0:t-1};\pi^w) \leq 0$. To complete the proof, it suffices to use the fact that for all $i \in \{1,\ldots,l-1\}$,

$$\begin{aligned} \max_{\substack{\pi'_{1:T},\pi_{1:T}:\\ \pi'_{iw+1:T}=\pi_{iw+1:T}}} \sum_{t=iw+1}^{T} \mu_{\pi_t,t}(u_{\pi_t,0:t-1};\pi) - \mu_{\pi_t,t}(u_{\pi_t,0:t-1};\pi') &\leq \sum_{t=0}^{T-iw-1} \overline{\lambda}\overline{\gamma}^t \frac{\overline{\gamma}}{1-\overline{\gamma}} \leq \frac{\overline{\lambda}\overline{\gamma}(1-\overline{\gamma}^{T-iw})}{(1-\overline{\gamma})^2} \\ &\leq \frac{\overline{\lambda}\overline{\gamma}(1-\overline{\gamma}^{T-w})}{(1-\overline{\gamma})^2}, \end{aligned}$$

where the first inequality holds because for any arm, the maximum satiation level discrepancy under two pull sequences (after iw time steps) is $\overline{\gamma}/(1-\overline{\gamma})$ and from time $iw+1$ till time T, the objective will be maximized when the arm with the maximum satiation discrepancy is played all the time. □

C.3 More Discussion on Learning with Unknown Dynamics

As we have noted in Section 4.5, when the learner makes a decision on which arm to pull, the learner does not observe the hidden satiation level the user has for the arms. The POMDP the learner faces can be cast as a fully observable MDP (Appendix C.3.1) where the estimated reward model (Appendix C.3.2) can be used for planning (Appendix C.3.3). In addition to policies that are time-dependent (actions taken by time-dependent policies only depend on the time steps at which they are taken) considered in Section 4.6, we also consider state-dependent policies where the states are continuous.

C.3.1 MDP Setup

We begin with describing the full MDP setup of rebounding bandits, including the state representation and reward function defined in Section 4.5.1. Following (Ortner et al., 2012), at any time $t \in [T]$, we define our state vector to be $x_t = (x_{1,t}, n_{1,t}, x_{2,t}, n_{2,t}, \ldots, x_{K,t}, n_{K,t})$, where $n_{k,t} \in \mathbb{N}$ is the number of steps since arm k is last selected and $x_{k,t}$ is the satiation influenceas of the most recent pull of arm k. Since the most recent pull happens at $t - n_{k,t}$, we have $x_{k,t} = b_k - \mu_{k,t-n_{k,t}} = \lambda_k s_{k,t-n_{k,t}}$. We note that b_k can be obtained when arm k is pulled for the first time since the satiation effect is 0 if an arm has not been pulled before. The initial state is $x_{\text{init}} = (0, \ldots, 0)$. Transitions between two states x_t and x_{t+1} are defined as follows: If arm k is chosen at time t, i.e., $\pi_t = k$, and reward $\mu_{k,t}$ is obtained, then the next state x_{t+1} will be:

A.1 For the pulled arm k, $n_{k,t+1} = 1$ and $x_{k,t+1} = b_k - \mu_{k,t}$.

A.2 For other arms $k' \neq k$, $n_{k',t+1} = n_{k',t} + 1$ if $n_{k',t} \neq 0$ and $n_{k',t+1} = 0$ if $n_{k',t} = 0$. The satiation influence remains the same, i.e., $x_{k',t+1} = x_{k',t}$.

For all $x_t \in \mathcal{X}$ and $k \in [K]$, we have that $\mathbb{E}[x_{k,t}] \leq \overline{\lambda}\overline{\gamma}/(1-\overline{\gamma})$ and $\text{Var}[x_{k,t}] \leq \overline{\lambda}^2\sigma_z^2/(1-\overline{\gamma}^2)$. Hence, for any $\delta \in (0,1)$, $\mathbb{P}\left(\max_{k,t} |x_{k,t}| \geq B(\delta)\right) \leq \delta$, where

$$B(\delta) := \frac{\overline{\lambda}\overline{\gamma}}{1-\overline{\gamma}} + \overline{\lambda}\sigma_z\sqrt{\frac{2\log(2KT/\delta)}{1-\overline{\gamma}^2}}. \tag{C.1}$$

The MDP the learner faces can be described as a tuple $\mathcal{M} := \langle x_{\text{init}}, [K], \{\gamma_k, \lambda_k, b_k\}_{k=1}^K, T\rangle$ of the initial state x_{init}, actions (arms) $[K]$, the horizon T and parameters $\{\gamma_k, \lambda_k, b_k\}_{k=1}^K$. Let $\Delta(\cdot)$ denote the probability simplex. Given $\{\gamma_k, \lambda_k, b_k\}_{k=1}^K$, the expected reward $r : \mathcal{X} \times [K] \to \mathbb{R}$ and transition functions $p : \mathcal{X} \times [K] \times [T] \to \Delta(\mathcal{X})$ are defined as follows:

1. $r : \mathcal{X} \times [K] \to \mathbb{R}$ gives the *expected* reward of pulling arm k conditioned on x_t, i.e., $r(x_t, k) = \mathbb{E}[\mu_{k,t}|x_t]$.[1] If $n_{k,t} = 0$, then $r(x_t, k) = b_k$. If $n_{k,t} \geq 1$, $r(x_t, k) = b_k - \gamma_k^{n_{k,t}} x_{k,t} - \lambda_k\gamma_k^{n_{k,t}}$.
2. When pulling arm k at time t and state x_t, $p\left(x_{t+1}|x_t, k, t\right) = 0$ if x_{t+1} does not satisfy A.1 or A.2. When x_{t+1} fulfills both A.1 and A.2, we consider two cases of x_t. If $n_{k,t} \neq 0$,

[1] By conditioning on x_t, we mean conditioning on the σ-algebra generated by past actions and observed rewards.

then the transition function $p\left(x_{t+1}|x_t, k, t\right)$ is given by the Gaussian density with mean $\gamma_k^{n_{k,t}}(x_{k,t} + \lambda_k)$ and variance $\lambda_k^2\sigma_z^2 \sum_{i=0}^{n_{k,t}-1} \gamma_k^{2i}$, as illustrated in (C.2). If $n_{k,t} = 0$, then $p(x_{t+1}|x_t, k, t) = 1$ since for the first pull of arm k, the obtained reward $\mu_{k,t} = b_k$.

At time t, the learner follows an action $\pi_t : \mathcal{X} \to [K]$ that depends on the state. We use $V^{\pi}_{t,\mathcal{M}} : \mathcal{X} \to \mathbb{R}$ to denote the value function of policy $\pi_{1:T}$ at time t under MDP $\mathcal{M}$: $V^{\pi}_{t,\mathcal{M}}(x_t) = r(x_t, \pi_t(x_t)) + \mathbb{E}_{x_{t+1}\sim p(\cdot|x_t,\pi_t(x_t),t)}[V^{\pi}_{t+1,\mathcal{M}}(x_{t+1})]$ and $V^{\pi}_{T+1,\mathcal{M}}(x) = 0$ for all $x \in \mathcal{X}$. To restate our goal (4.2) in terms of the value function: for an MDP $\mathcal{M}$, we would like to find a policy $\pi_{1:T}$ that maximizes

$$V^{\pi}_{1,\mathcal{M}}(x_{\text{init}}) = \mathbb{E}\left[\sum_{t=1}^{T} r(x_t, \pi_t(x_t)) \middle| x_1 = x_{\text{init}}\right].$$

To simplify the notation, we use π to refer to a policy $\pi_{1:T}$. Given an MDP $\mathcal{M}$, we denote its optimal policy by $\pi^*_{\mathcal{M}}$ and the value function for the optimal policy by $V^*_{t,\mathcal{M}}$, i.e., $V^*_{t,\mathcal{M}}(x) := V^{\pi^*_{\mathcal{M}}}_{t,\mathcal{M}}(x)$.

C.3.2 Exploration and Estimation of the Reward Model

As we have discussed in Section 4.6.1, based on our satiation and reward models, the satiation influence $x_{k,t}$ of arm k forms a dynamical system where we only observe the value of the system when arm k is pulled. When arm k is pulled at time t and $n_{k,t} \neq 0$, we observe the satiation influence $\lambda_k s_{k,t}$ which becomes the next state $x_{k,t+1}$, i.e.,

$$\begin{aligned} x_{k,t+1} = \lambda_k s_{k,t} &= \lambda_k \gamma_k^{n_{k,t}} s_{k,t-n_{k,t}} + \lambda_k \gamma_k^{n_{k,t}} + \lambda_k \sum_{i=0}^{n_{k,t}-1} \gamma_k^i z_{k,t-1-i} \\ &= \gamma_k^{n_{k,t}} x_{k,t+1-n_{k,t}} + \lambda_k \gamma_k^{n_{k,t}} + \lambda_k \sum_{i=0}^{n_{k,t}-1} \gamma_k^i z_{k,t-1-i}. \end{aligned} \tag{C.2}$$

We note that the current state $x_{k,t}$ equals to $x_{k,t+1-n_{k,t}}$ since $x_{k,t+1-n_{k,t}}$ is the last observed satiation influence for arm k and $n_{k,t}$ is the number of steps since arm k is last pulled.

Exploration Settings Depending on the nature of the recommendation domain, we consider two types of exploration settings: one where the users only interact with the recommendation systems for a short time after they log in to the service (Appendix C.3.2) and the other where the users tend to interact with the system for a much longer time, e.g., automated music playlisting (Appendix C.3.2). In the first case, the learner collects multiple (n) short trajectories of user utilities, while in the second case, similar to Section 4.6.2, the learner obtains a single trajectory of user utilities that has length n. In both settings, we obtain that under some mild conditions, the estimation errors of our estimators for γ_k and λ_k are $O(1/\sqrt{n})$.

Exploration Strategies Generalizing from the case where arms are pulled repeatedly, we explore by pulling the same arm at a fixed interval m. In particular, when $m = 1$, the exploration strategy is the same as repeatedly pulling the same arm for multiple times, which is the

exploration strategy used in Section 4.6.1. When $m = K$, the exploration strategy is to pull the arms in a cyclic order. We present the estimator for γ_k, λ_k using the dataset collected by this exploration strategy in both the multiple trajectory and single trajectory settings.

Estimation using Multiple Trajectories

For each arm $k \in [K]$, we use $\mathcal{D}_k^{n,m}$ to denote a dataset containing n trajectories of evenly spaced observed satiation influences that are collected by our exploration phase. The time interval between two pulls of an arm is denoted by m. Each trajectory is of length at least $T_{\min} + 1$ for $T_{\min} > 1$. For trajectory $i \in [n]$, the observed satiation influences are denoted by $\tilde{x}_{k,1}^{(i)}, \ldots, \tilde{x}_{k,T_{\min}+1}^{(i)}, \ldots$, where $\tilde{x}_{k,1}^{(i)} = 0$ is the initial satiation influence and the rest of the satiation influences $\tilde{x}_{k,j}^{(i)}$ $(j > 1)$ is the difference between the first received reward, i.e., the base reward b_k, and the reward from the j-th pull of arm k. In other words, for $\tilde{x}_{k,j}^{(i)}, \tilde{x}_{k,j+1}^{(i)} \in \mathcal{D}_k^{n,m}$, it follows that

$$\tilde{x}_{k,j+1}^{(i)} = a_k \tilde{x}_{k,j}^{(i)} + d_k + \tilde{z}_{k,j}^{(i)}, \tag{C.3}$$

where $a_k = \gamma_k^m$, $d_k = \lambda_k \gamma_k^m$ and $\tilde{z}_{k,j}^{(i)}$ are the independent samples from $\mathcal{N}\left(0, \sigma_{z,k}^2\right)$ with $\sigma_{z,k}^2 = \lambda_k^2 \sigma_z^2 (1 - \gamma_k^{2m})/(1 - \gamma_k^2)$.

To estimate d_k, we use the estimator $\widehat{d}_k = \frac{1}{n} \sum_{i=1}^n \tilde{x}_{k,2}^{(i)} = d_k + \frac{1}{n} \sum_{i=1}^n \tilde{z}_{k,1}^{(i)}$. By the standard Gaussian tail bound, we obtain that for $\delta \in (0, 1)$, with probability $1 - \delta$,

$$|\widehat{d}_k - d_k| \leq \sqrt{\frac{2\sigma_{z,k}^2 \log(2/\delta)}{n}} =: \epsilon_d(n, \delta, k). \tag{C.4}$$

When estimating a_k, we first take the difference between the first $T_{\min} + 1$ entries of two trajectories i and $2i$ for $i \in \lfloor n/2 \rfloor$ and obtain a new trajectory $\tilde{y}_{k,1}^{(i)}, \ldots, \tilde{y}_{k,T_{\min}+1}^{(i)}$ where $\tilde{y}_{k,j}^{(i)} = \tilde{x}_{k,j}^{(i)} - \tilde{x}_{k,j}^{(2i)}$ for $j \in [T_{\min} + 1]$. We note that the new trajectory forms a linear dynamical system without the bias term d_k, i.e.,

$$\tilde{y}_{k,j+1}^{(i)} = a_k \tilde{y}_{k,j}^{(i)} + \tilde{w}_{k,j}^{(i)},$$

where $\tilde{w}_{k,j}^{(i)}$ are samples from $\mathcal{N}(0, 2\sigma_{z,k}^2)$. We use the ordinary least squares estimator to estimate a_k:

$$\begin{aligned} \widehat{a}_k &= \arg\min_a \sum_{i=1}^{\lfloor n/2 \rfloor} \left(\tilde{y}_{k,T_{\min}+1}^{(i)} - a \tilde{y}_{k,T_{\min}}^{(i)} \right)^2 \\ &= \frac{\sum_{i=1}^{\lfloor n/2 \rfloor} \tilde{y}_{k,T_{\min}}^{(i)} \tilde{y}_{k,T_{\min}+1}^{(i)}}{\sum_{i=1}^{\lfloor n/2 \rfloor} \left(\tilde{y}_{k,T_{\min}}^{(i)} \right)^2}. \end{aligned} \tag{C.5}$$

Theorem C.1

(Matni and Tu, 2019, Theorem II.4) Fix $\delta \in (0,1)$. Given $n \geq 64\log(2/\delta)$, with probability $1-\delta$, we have that

$$|\widehat{a}_k - a_k| \leq 4\sqrt{\frac{2\log(4/\delta)}{n\sum_{t=0}^{T_{\min}} a_k^{2t}}} =: \epsilon_a(n,\delta,k). \tag{C.6}$$

We notice that as the minimum length of the trajectory gets greater, the upper bound of the estimation error of a_k gets smaller. Using our estimators for a_k and d_k, we estimate γ_k and λ_k through $\widehat{\gamma}_k = |\widehat{a}_k|^{1/m}$ and $\widehat{\lambda}_k = |\widehat{d}_k/\widehat{a}_k|$.

Corollary C.1. *Fix $\delta \in (0,1)$. Suppose that for all $k \in [K]$, we are given $\mathcal{D}_k^{n,m}$ where $n \geq 64\log(2/\delta)$ and $\widehat{a}_k > 0$ where $\widehat{a}_k$ is defined in (C.5). Then, with probability $1-\delta$, we have that for all $k \in [K]$,*

$$|\widehat{\gamma}_k - \gamma_k| \leq \frac{\epsilon_a(n,\delta/K,k)}{\gamma_k^{m-1}} = O\left(\frac{1}{\sqrt{n}}\right) \quad \textit{and} \quad |\widehat{\lambda}_k - \lambda_k| \leq O\left(\frac{1}{\sqrt{n}}\right).$$

The proof of Corollary C.1 can be found in Appendix C.6.1. In the case where we have collected n trajectories of evenly spaced user utilities for each arm, when the sample size n is sufficient large, the estimation errors of $\widehat{\gamma}_k$ and $\widehat{\lambda}_k$ are $O(1/\sqrt{n})$.

Estimation using a Single Trajectory

In the case where the learner gets to interact with the user for a long period of time (which is the setting considered in Section 4.5 and Section 4.6), we collect a single trajectory of evenly spaced arm pulls for each arm: for each arm $k \in [K]$, we use $\mathcal{P}_k^{n,m}$ to denote a dataset containing a single trajectory of $n+1$ observed satiation influences $\tilde{x}_{k,1}, \ldots, \tilde{x}_{k,n+1}$, where similar to the multiple trajectories case, $\tilde{x}_{k,1} = 0$, $\tilde{x}_{k,j}$ $(j > 1)$ is the difference between the first received reward and the j-th received reward and the time interval between two consecutive pulls is m. Thus, for $\tilde{x}_{k,j}, \tilde{x}_{k,j+1} \in \mathcal{P}_k^{n,m}$, it follows that

$$\tilde{x}_{k,j+1} = a_k\tilde{x}_{k,j} + d_k + \tilde{z}_{k,j}, \tag{C.7}$$

where a_k, d_k and $\tilde{z}_{k,j}$ are defined the same as the ones in (C.3). For all $k \in [K]$, given $\mathcal{P}_k^{n,m}$, we use the following estimators to estimate $A_k = (a_k, d_k)^\top$,

$$\widehat{A}_k = \begin{pmatrix} \widehat{a}_k \\ \widehat{d}_k \end{pmatrix} = (\overline{\mathbf{X}}_\mathbf{k}{}^\top \overline{\mathbf{X}}_\mathbf{k})^{-1}\overline{\mathbf{X}}_\mathbf{k}{}^\top \mathbf{Y}_\mathbf{k}, \tag{C.8}$$

where $\mathbf{Y}_\mathbf{k} \in \mathbb{R}^n$ is an n-dimensional vector whose j-th entry is $\tilde{x}_{k,j+1}$ and $\overline{\mathbf{X}}_\mathbf{k} \in \mathbb{R}^{n\times 2}$ has its j-th row to be the vector $\overline{x}_{k,j} = (\tilde{x}_{k,j}, 1)^\top$. Finally, we take $\widehat{\gamma}_k = |\widehat{a}_k|^{1/m}$ and $\widehat{\lambda}_k = |\widehat{d}_k/\widehat{a}_k|$. We note that $\widehat{A}_k = \arg\min_{A_k \in \mathbb{R}^2} \|\mathbf{Y}_\mathbf{k} - \overline{\mathbf{X}}_\mathbf{k} A_k\|_2^2$, i.e., it is the ordinary least squares estimator for A_k given the dataset that treats $\tilde{x}_{k,j+1}$ to be the response of the covariates $\overline{x}_{k,j}$.

As we have noted earlier (Section 4.6.2), unlike the multiple trajectories setting, in the single trajectory case, the difficulty in analyzing the ordinary least squares estimator (C.8) comes from the fact that the samples are not independent. Asymptotic guarantees of the ordinary least squares estimators in this case have been studied previously in control theory and time series community (Hamilton, 1994; Ljung, 1999). The recent work on system identifications for linear dynamical systems focuses on studying the sample complexity of the problem (Simchowitz et al., 2018; Sarkar and Rakhlin, 2019). Adapting the proof of (Simchowitz et al., 2018, Theorem 2.4), we derive the following theorem for identifying our affine dynamical system (C.7).

Theorem C.2

Fix $\delta \in (0,1)$. For all $k \in [K]$, there exists a constant $n_0(\delta, k)$ such that if the dataset $\mathcal{P}_k^{n,m}$ satisfies $n \geq n_0(\delta, k)$, then

$$\mathbb{P}\left(\|\widehat{A}_k - A_k\|_2 \gtrsim \sqrt{1/(\psi n)}\right) \leq \delta,$$

where $\psi = \sqrt{\min\left\{\frac{\sigma_{z,k}^2(1-a_k)^2}{16d_k^2(1-a_k^2)+(1-a_k)^2\sigma_{z,k}^2}, \frac{\sigma_{z,k}^2}{4(1-a_k^2)}\right\}}$.

As shown in Theorem C.2, when $d_k = \lambda_k \gamma_k^m$ gets larger, the rates of convergence for $\widehat{A}_k$ gets slower. Given that we have a single trajectory of sufficient length, $|\widehat{a}_k - a_k| \leq O(1/\sqrt{n})$ and $|\widehat{d}_k - d_k| \leq O(1/\sqrt{n})$. Similar to the multiple trajectories case, as shown in Corollary C.2, the estimators of γ_k and λ_k also achieve $O(1/\sqrt{n})$ estimation error.

Corollary C.2. *Fix $\delta \in (0,1)$. Suppose that for all $k \in [K]$, we have $\mathbb{P}(\|\widehat{A}_k - A_k\|_2 \gtrsim 1/\sqrt{n}) \leq \delta$ and $\widehat{a}_k > 0$ where $\widehat{A}_k$ and $\widehat{a}_k$ are defined in (C.8). Then, with probability $1-\delta$, we have that for all $k \in [K]$,*

$$|\widehat{\gamma}_k - \gamma_k| \leq O\left(\frac{1}{\sqrt{n}}\right) \quad \textit{and} \quad |\widehat{\lambda}_k - \lambda_k| \leq O\left(\frac{1}{\sqrt{n}}\right).$$

In the next section, we assume that the satiation and reward models are estimated using the dataset collected by the proposed exploration strategies and estimators for multiple trajectories or a single trajectory of user utilities. We will show that performing planning based on these estimated models will give us policies that perform well for the true MDP.

C.3.3 Planning

For a continuous-state MDP, planning can be done through either dynamic programming with a discretized state space or approximate dynamic programming that uses function approximations. In Appendix C.3.3, we consider the case where we are given a continuous-state MDP planning oracle and provide guarantees of the optimal state-dependent policy planned under the estimated satiation dynamics and reward model. Within the state-dependent policies, we also consider a set of policies that only depend on time (Appendix C.3.3), i.e., the time-dependent competitor class defined in Section 4.5.2. In addition to not requiring discretization of the state space to

solve the planning problem, such policies can be deployed to settings where user utilities are hard to attain after the exploration stage. We will show that using the dataset (collected by our exploration strategy in Appendix C.3.2) with sufficient trajectories (or a sufficient long trajectory) to estimate $\{\gamma_k, \lambda_k\}_{k=1}^K$, the optimal policy $\pi^*_{\widehat{\mathcal{M}}}$ for $\widehat{\mathcal{M}} = \langle x_1, [K], \{\widehat{\gamma}_k, \widehat{\lambda}_k, b_k\}_{k=1}^K, T\rangle$ also performs well in the original MDP $\mathcal{M}$. We note that b_k is known exactly since it is the same as the first observed reward for arm k, as discussed in Appendix C.3.2.

Time-dependent Policy

We first show that finding the optimal time-dependent policy is equivalent to solving the bilinear program (4.4).

Lemma C.2

Consider a policy π that depends only on the time step t but not the state x_t, i.e., π satisfies $\pi_t = \pi_t(x_t) = \pi_t(x'_t)$ for all $t \in [T]$ and $x_t, x'_t \in \mathcal{X}$. Then, we have

$$V^{\pi}_{1,\mathcal{M}}(x_{\text{init}}) = \sum_{t=1}^{T} \mu_{\pi_t,t}(u_{\pi_t,0:t-1}),$$

where $u_{\pi_t,0:t-1}$ is the corresponding pull sequence of arm π_t under policy π and $\mu_{k,t}$ is defined in (4.3).

Remark C.1. *We denote the policy obtained by solving (4.4) using model parameters in $\mathcal{M}$ by $\pi^T_{\mathcal{M}}$. Because solving (4.4) is equivalent to maximizing $\sum_{t=1}^T \mu_{\pi_t,t}(u_{\pi_t,0:t-1})$, Lemma C.2 suggests that, for MDP $\mathcal{M}$, the best policy π that depends only on the time step t but not the exact state x_t (which we refer as time-dependent policies), is $\pi^T_{\mathcal{M}}$.*

Proposition C.1

Fix $\delta \in (0,1)$. Suppose that for all $k \in [K]$, we are given $\mathcal{D}_k^{n,m}$ such that $n \geq 64\log(2/\delta)$ and $\widehat{a}_k \in (\underline{a}, \overline{a})$ for some $0 < \underline{a} < \overline{a} < 1$ almost surely where $\widehat{a}_k$ is defined in (C.5). Consider a policy π that depends on only the time step t but not the state x_t. Then, with probability $1-\delta$, we have that

$$|V^{\pi}_{1,\mathcal{M}}(x_{\text{init}}) - V^{\pi}_{1,\widehat{\mathcal{M}}}(x_{\text{init}})| \leq O\left(\frac{T}{\sqrt{n}}\right).$$

Remark C.2. *Proposition C.1 applies to time-dependent policies. Such policies can be constructed from an optimal solution to (4.4) or the w-lookahead policy (4.5). From these results, we deduce that when the historical trajectory is of size $n = O(T)$, the $\sqrt{T}$-lookahead policy $\pi^w_{\widehat{\mathcal{M}}}$ obtained from solving (4.5) with the parameters from the estimated MDP $\widehat{\mathcal{M}}$ will be $O(\sqrt{T})$-separated from the optimal time-dependent policy $\pi^T_{\mathcal{M}}$ obtained by solving (4.4) with the true parameters of $\mathcal{M}$. That is,*

$$0 \leq V^{\pi^T_{\mathcal{M}}}_{1,\mathcal{M}}(x_{init}) - V^{\pi^w_{\widehat{\mathcal{M}}}}_{1,\mathcal{M}}(x_{init}) = V^{\pi^T_{\mathcal{M}}}_{1,\mathcal{M}}(x_{init}) - V^{\pi^T_{\mathcal{M}}}_{1,\widehat{\mathcal{M}}}(x_{init}) + V^{\pi^T_{\mathcal{M}}}_{1,\widehat{\mathcal{M}}}(x_{init}) - V^{\pi^T_{\widehat{\mathcal{M}}}}_{1,\widehat{\mathcal{M}}}(x_{init})$$

$$
\begin{aligned}
&\quad + V_{1,\widehat{\mathcal{M}}}^{\pi_{\widehat{\mathcal{M}}}^T}(x_{init}) - V_{1,\widehat{\mathcal{M}}}^{\pi_{\widehat{\mathcal{M}}}^w}(x_{init}) + V_{1,\widehat{\mathcal{M}}}^{\pi_{\widehat{\mathcal{M}}}^w}(x_{init}) - V_{1,\mathcal{M}}^{\pi_{\widehat{\mathcal{M}}}^w}(x_{init}) \\
&\leq |V_{1,\mathcal{M}}^{\pi_{\mathcal{M}}^T}(x_{init}) - V_{1,\widehat{\mathcal{M}}}^{\pi_{\mathcal{M}}^T}(x_{init})| + |V_{1,\widehat{\mathcal{M}}}^{\pi_{\widehat{\mathcal{M}}}^T}(x_{init}) - V_{1,\widehat{\mathcal{M}}}^{\pi_{\widehat{\mathcal{M}}}^w}(x_{init})| + |V_{1,\widehat{\mathcal{M}}}^{\pi_{\widehat{\mathcal{M}}}^w}(x_{init}) - V_{1,\mathcal{M}}^{\pi_{\widehat{\mathcal{M}}}^w}(x_{init})| \\
&\leq O(\sqrt{T}),
\end{aligned}
$$

where the second inequality follows from the fact that $V_{1,\widehat{\mathcal{M}}}^{\pi_{\mathcal{M}}^T}(x_{init}) - V_{1,\widehat{\mathcal{M}}}^{\pi_{\widehat{\mathcal{M}}}^T}(x_{init}) \leq 0$ *(since for the MDP* $\widehat{\mathcal{M}}$*,* $\pi_{\widehat{\mathcal{M}}}^T$ *is the optimal time-dependent policy), and the third (last) inequality is derived by applying Proposition C.1 twice and using Remark 4.4.*

State-dependent Policy

In Proposition C.2, we show that the difference between the value of the optimal state-dependent policy $\pi_{\mathcal{M}}^*$, and the value of the optimal state-dependent policy $\pi_{\widehat{\mathcal{M}}}^*$ planned under the estimated $\widehat{\mathcal{M}}$ is of order $O(T^2/\sqrt{n})$ where n is the number of historical trajectories if we use multiple trajectories to estimate γ_k and λ_k.

Proposition C.2

Fix $\delta \in (0,1)$. Suppose that for all $k \in [K]$, we are given $\mathcal{D}_k^{n,m}$ such that $n \geq 64\log(2/\delta)$ and $\widehat{a}_k \in (\underline{a}, \overline{a})$ for some $0 < \underline{a} < \overline{a} < 1$ almost surely where $\widehat{a}_k$ is defined in (C.5). Then, with probability $1-\delta$,

$$
|V_{1,\mathcal{M}}^*(x_{\text{init}}) - V_{1,\mathcal{M}}^{\pi_{\widehat{\mathcal{M}}}^*}(x_{\text{init}})| \leq O\left(\frac{T^2}{\sqrt{n}}\right).
$$

Remark C.3. *The assumptions in Proposition C.1 and C.2 correspond to the case where we use multiple trajectories to estimate the satiation dynamics and reward model. They can be replaced by conditions on single trajectory datasets when one uses a single trajectory to estimate the parameters.*

In summary, as Proposition C.2 suggests, when given a continuous-state MDP planning oracle, our algorithm obtain a policy $\pi_{\widehat{\mathcal{M}}}^*$ that is $O(T^2/\sqrt{n})$ away from the optimal policy $\pi_{\mathcal{M}}^*$ under the true MDP $\mathcal{M}$ where the size of the exploration stage for our algorithm (EEP) is $O(Kn)$ and the horizon of the exploitation/planning stage is T. We also note that the optimal state-dependent policy $\pi_{\mathcal{M}}^*$ is the optimal competitor policy when the competitor class (Section 4.5.2) contains all measurable functions from $\mathcal{X}$ to $[K]$.

C.4 Proofs of Section 4.6.2 and Appendix C.3.2

C.4.1 Proof of Theorem 4.3 and Theorem C.2

We notice that Theorem 4.3 is a consequence of Theorem C.2 when $m = 1$. More specifically, the dataset $\mathcal{P}_k^n$ and the parameter $A_k = (\gamma_k, \lambda_k\gamma_k)^\top$ in Theorem 4.3 is a special case of the dataset $\mathcal{P}_k^{n,m}$ and parameter $A_k = (\gamma_k^m, \lambda_k\gamma_k^m)^\top$ considered in Theorem C.2 by taking $m = 1$. Thus, below we directly present the proof of Theorem C.2 where we use the notation from Theorem C.2 (and Appendix C.3.2), i.e., $a_k = \gamma_k^m$ and $d_k = \lambda_k\gamma_k^m$.

We begin with presenting some key results from (Simchowitz et al., 2018); we utilize these results in establishing the sample complexity of our estimator for identifying an affine dynamical system in Appendix C.3.2.

Definition C.1

(Simchowitz et al., 2018, Definition 2.1) Let $\{\phi_t\}_{t\geq 1}$ be an $\{\mathcal{F}_t\}_{t\geq 1}$-adapted random process taking values in $\mathbb{R}$. We say $(\phi_t)_{t\geq 1}$ satisfies the (k, ν, p)-block martingale small-ball (BMSB) condition if, for any $j \geq 0$, one has $\frac{1}{k}\sum_{i=1}^k \mathbb{P}(|\phi_{j+i}| \geq \nu|\mathcal{F}_j) \geq p$ almost surely. Given a process $(X_t)_{t\geq 1}$ taking values in $\mathbb{R}^d$, we say that it satisfies the $(k, \Gamma_{\text{sb}}, p)$-BMSB condition for $\Gamma_{\text{sb}} \succ 0$ if for any fixed w in the unit sphere of $\mathbb{R}^d$, the process $\phi_t := \langle w, X_t\rangle$ satisfies $(k, \sqrt{w^\top\Gamma_{\text{sb}}w}, p)$-BMSB.

Proposition C.3

(Simchowitz et al., 2018, Proposition 2.5) Fix a unit vector $w \in \mathbb{R}^d$, define $\phi_t = w^\top X_t$. If the scalar process $\{\phi_t\}_{t\geq 1}$ satisfies the $(l, \sqrt{w^\top\Gamma_{\text{sb}}w}, p)$-BMSB condition for some $\Gamma_{\text{sb}} \in \mathbb{R}^{d\times d}$, then

$$\mathbb{P}\left(\sum_{t=1}^n \phi_t^2 \leq \frac{w^\top\Gamma_{\text{sb}}wp^2}{8} l\lfloor T/l\rfloor\right) \leq \exp\left(-\frac{\lfloor T/l\rfloor p^2}{8}\right).$$

Theorem C.3

(Simchowitz et al., 2018, Theorem 2.4) Fix $\delta \in (0,1)$, $T \in \mathbb{N}$ and $0 \prec \Gamma_{\text{sb}} \preceq \overline{\Gamma}$. Then if $(X_t, Y_t)_{t\geq 1} \in (\mathbb{R}^d \times \mathbb{R}^n)^n$ is a random sequence such that (a) $Y_t = AX_t + \eta_t$, where $\mathcal{F}_t = \sigma(\eta_1, \ldots, \eta_t)$ and $\eta_t|\mathcal{F}_{t-1}$ is σ^2-sub-Gaussian and mean zero, (b) $X_1, \ldots, X_T$ satisfies the $(l, \Gamma_{\text{sb}}, p)$-BMSB condition, and (c) $\mathbb{P}(\sum_{t=1}^n X_tX_t^\top \not\preceq T\overline{\Gamma}) \geq \delta$. Then if

$$T \geq \frac{10l}{p^2}\left(\log(1/\delta) + 2d\log(10/p) + \log\det(\overline{\Gamma}\Gamma_{\text{sb}}^{-1})\right),$$

we have that for $\widehat{A} = \arg\min_{A \in \mathbb{R}^{n \times d}} \sum_{t=1}^{T} \|Y_t - AX_t\|_2^2$,

$$\mathbb{P}\left(\|\widehat{A} - A\|_{\text{op}} > \frac{90\sigma}{p} \sqrt{\frac{n + d\log(10/p) + \log\det\left(\overline{\Gamma}\Gamma_{\text{sb}}^{-1}\right) + \log(1/\delta)}{T\lambda_{\min}(\Gamma_{\text{sb}})}} \right) \leq 3\delta.$$

We note that in the proof of Theorem C.3 in (Simchowitz et al., 2018), condition (b) is used through applying Proposition C.3 to ensure that for any unit vector $w \in \mathbb{R}^d$,

$$\mathbb{P}\left(\sum_{t=1}^{T} \langle w, X_t \rangle^2 \leq \frac{(w^\top \Gamma_{\text{sb}} w)p^2}{8} l\lfloor T/l \rfloor \right) \leq \exp\left(-\frac{\lfloor T/l \rfloor p^2}{8} \right). \tag{C.9}$$

To apply Theorem C.3 in our setting to obtain Theorem C.2, we verify condition (a) and (c). For condition (b), we show a result similar to (C.9). The below technical lemmas are used in our proof of Theorem C.2.

Lemma C.3

Let a, b be scalars with $b > 0$. Suppose that $X \sim N(a, b)$. Then for any $\theta \in [0, 1]$,

$$\mathbb{P}(|X| \geq \sqrt{\theta(a^2 + b)}) \geq \frac{(1 - \theta)^2}{9}.$$

Proof. By the Paley-Zygmund inequality,

$$\mathbb{P}(|X| \geq \sqrt{\theta \mathbb{E}[X^2]}) = \mathbb{P}(X^2 \geq \theta \mathbb{E}[X^2]) \geq (1 - \theta)^2 \frac{\mathbb{E}[X^2]^2}{\mathbb{E}[X^4]}.$$

Using the mean and variance of non-central chi-squared distributions, we obtain that

$$\begin{aligned} \mathbb{E}[X^2] &= a^2 + b, \\ \mathbb{E}[X^4] &= a^4 + 6a^2 b + 3b^2 = (a^2 + 3b)^2 - 6b^2. \end{aligned}$$

Plugging them back to the Paley-Zygmund inequality, we have that

$$\mathbb{P}(|X| \geq \sqrt{\theta(a^2 + b)} \geq \frac{(1 - \theta)^2}{9},$$

where the last inequality uses the fact that $\mathbb{E}[X^4] \leq (a^2 + 3b)^2 \leq 9(a^2 + b)^2 = 9\mathbb{E}[X^2]^2$. □

Lemma C.4

Let $\{\phi_t\}_{t \geq 1}$ be a scalar process satisfying that

$$\frac{1}{l} \sum_{i=1}^{l} \mathbb{P}(|\phi_{t+i}| \geq \nu_t | \mathcal{F}_t) \geq p,$$

for ν_t depending on $\mathcal{F}_t$. If $\mathbb{P}(\min_t \nu_t \geq \nu) \geq 1-\delta$ for $\nu > 0$ that depends on δ, then

$$\mathbb{P}\left(\sum_{t=1}^{T} \phi_t^2 \leq \frac{\nu^2 p^2}{8} l\lfloor T/l \rfloor\right) \leq \exp\left(-\frac{3\lfloor T/l \rfloor p}{4}\right) + \delta.$$

Proof. We begin with partitioning $Z_1, \ldots, Z_T$ into $S := \lfloor T/l \rfloor$ blocks of size l. Consider the random variables

$$B_j = \mathbf{1}\left(\sum_{i=1}^{l} \phi_{jl+i}^2 \geq \frac{\nu_{jl}^2 pk}{2}\right), \text{ for } 0 \leq j \leq S-1.$$

We observe that

$$\begin{aligned}
\mathbb{P}\left(\sum_{t=1}^{T} \phi_t^2 \leq \frac{\nu^2 p^2}{8} l\lfloor T/l \rfloor\right) &= \mathbb{P}\left(\left\{\sum_{t=1}^{T} \phi_t^2 \leq \frac{\nu^2 p^2}{8} l\lfloor T/l \rfloor\right\} \cap \{\min_t \nu_t \geq \nu\}\right) \\
&\quad + \mathbb{P}\left(\left\{\sum_{t=1}^{T} \phi_t^2 \leq \frac{\nu^2 p^2}{8} l\lfloor T/l \rfloor\right\} \cap \{\min_t \nu_t < \nu\}\right) \\
&\leq \mathbb{P}\left(\left\{\sum_{t=1}^{T} \phi_t^2 \leq \frac{\nu_{\lfloor t/l \rfloor l}^2 p^2}{8} lS\right\} \cap \{\min_t \nu_t \geq \nu\}\right) + \mathbb{P}(\min_t \nu_t < \nu) \\
&\leq \mathbb{P}\left(\sum_{t=1}^{T} \phi_t^2 \leq \frac{\nu_{\lfloor t/l \rfloor l}^2 p^2}{8} kS\right) + \delta.
\end{aligned}$$

Using Chernoff bound, we obtain that

$$\begin{aligned}
\mathbb{P}\left(\sum_{t=1}^{T} \phi_t^2 \leq \frac{\nu_{\lfloor t/l \rfloor l}^2 p^2}{8} kS\right) &\leq \mathbb{P}\left(\sum_{j=0}^{S-1}\sum_{i=1}^{l} \phi_{jl+i}^2 \leq \frac{\nu_{jl}^2 p^2}{8} lS\right) = \mathbb{P}\left(\sum_{j=0}^{S-1}\sum_{i=1}^{l} \phi_{jl+i}^2 \leq \frac{\nu_{jl}^2 p^2}{8} lS\right) \\
&\leq \mathbb{P}\left(\sum_{j=0}^{S-1} B_j \leq \frac{p}{4} S\right) \leq \inf_{\lambda \leq 0} e^{-\frac{pS}{4}} \mathbb{E}[e^{\lambda \sum_{j=0}^{S-1} B_j}],
\end{aligned}$$

where the second to the last inequality uses the fact that $\frac{\nu_{jl}^2 pl}{2} B_j \leq \sum_{i=1}^{l} \phi_{jl+i}^2$ Further, we have that

$$\begin{aligned}
\mathbb{E}[B_j | \mathcal{F}_{jl}] = \mathbb{P}\left(\sum_{i=1}^{l} \phi_{jl+i}^2 \geq \frac{\nu_{jl}^2 pl}{2} \Big| \mathcal{F}_{jl}\right) &\geq \mathbb{P}\left(\frac{1}{l}\sum_{i=1}^{l} \mathbf{1}\left\{|\phi_{jl+i}| \geq \nu_{jl}\right\} \geq \frac{p}{2} \Big| \mathcal{F}_{jl}\right) \\
&\geq \frac{p}{2},
\end{aligned}$$

where the first inequality uses the fact that $\frac{1}{\nu_{jl}^2}\phi_{jl+i}^2 \geq \mathbf{1}\{\phi_{jl+i}| \geq \nu_{jl}\}$ and the last inequality uses the fact that for a random variable X supported on $[0,1]$ almost surely such that $\mathbb{E}[X] \geq p$

for some $p \in (0,1)$, then for all $t \in [0,p]$, $\mathbb{P}(X \geq t) \geq \frac{p-t}{1-t}$. This is true because

$$\mathbb{P}(X \geq t) = \int_t^1 d\mathbb{P}(x) \geq \int_t^1 x d\mathbb{P}(x) = \int_0^1 x d\mathbb{P}(x) - \int_0^t x d\mathbb{P}(x) = p - t(1 - \mathbb{P}(X \geq t)).$$

In our case, $\mathbb{E}\left[\frac{1}{l}\sum_{i=1}^{l} \mathbf{1}\{|\phi_{jl+i}| \geq \nu_{jl}\} \middle| \mathcal{F}_{jl}\right] = \frac{1}{l}\sum_{i=1}^{l} \mathbb{P}\left(|\phi_{jl+i}| \geq \nu_{jl} \middle| \mathcal{F}_{jl}\right) \geq p$. Thus, we obtain that for $\lambda \leq 0$, i.e., $e^{\lambda} \leq 1$,

$$\mathbb{E}[e^{\lambda B_j}|\mathcal{F}_{jl}] = e^{\lambda}\mathbb{P}\left(B_j = 1 \middle| \mathcal{F}_{jl}\right) + \mathbb{P}(B_j = 0) = (e^{\lambda} - 1)\mathbb{E}[B_j|\mathcal{F}_{jl}] + 1 \leq (e^{\lambda} - 1)\frac{p}{2} + 1.$$

By law of iterated expectation, we obtain that

$$\begin{aligned}\mathbb{E}[e^{\lambda \sum_{j=0}^{S-1} B_j}] &= \mathbb{E}\left[e^{\lambda \sum_{j=0}^{S-2} B_j}\mathbb{E}[e^{\lambda B_j}|\mathcal{F}_{(S-1)k}]\right] \leq \left((e^{\lambda} - 1)\frac{p}{2} + 1\right)\mathbb{E}\left[e^{\lambda \sum_{j=0}^{S-2} B_j}\right] \\ &\leq \left((e^{\lambda} - 1)\frac{p}{2} + 1\right)^S.\end{aligned}$$

Finally, we need to find

$$\inf_{\lambda \leq 0} e^{-pS/4}\left((e^{\lambda} - 1)\frac{p}{2} + 1\right)^S.$$

We can see that $\lambda^* = -\infty$, which gives that

$$\inf_{\lambda \leq 0} e^{-pS/4}\left((e^{\lambda} - 1)\frac{p}{2} + 1\right)^S = e^{-pS/4}\left(1 - \frac{p}{2}\right)^S \leq e^{-pS/4}e^{-pS/2} = e^{-3pS/4},$$

where we have used the fact that $1 + x \leq e^x$ for all real-valued x. □

To apply Theorem C.3, we first recall that the affine dynamical system we aim to identify is as follows:

$$\tilde{x}_{k,j+1} = a_k\tilde{x}_{k,j} + d_k + \tilde{z}_{k,j},$$

where $\tilde{x}_{k,1} = 0$, $a_k \in (0,1)$ and $\tilde{z}_{k,j} \sim \mathcal{N}(0, \sigma_{z,k}^2)$. We define the following quantities

$$\Gamma_{k,j} := \sigma_{z,k}^2 \sum_{i=0}^{j-1} a_k^{2i}, \qquad d_{k,j} := \sum_{i=0}^{j-1} a_k^j d_k,$$

and $\Gamma_{k,\infty} = \sigma_{z,k}^2 \sum_{i=0}^{\infty} a_k^{2i} = \frac{\sigma_{z,k}^2}{1-a_k^2}$. We notice that for all $t \in [T]$, $j \geq 1$,

$$\tilde{x}_{k,t+j}|\tilde{x}_{k,t} \sim \mathcal{N}\left(a_k^j\tilde{x}_{k,t} + d_{k,j}, \Gamma_{k,j}\right).$$

Lemma C.5

Fix $t \geq 0$ and $j \geq 1$. Recall that $\overline{x}_{k,t} := (\tilde{x}_{k,t}, 1) \in \mathbb{R}^2$. Fix a unit vector $w \in \mathbb{R}^2$. For any $\epsilon \in (0,1)$, we have

$$\mathbb{P}\left(|\langle w, \overline{x}_{k,t+j}\rangle| \geq \frac{1}{\sqrt{2}}\sqrt{\min\left\{1-\epsilon, \Gamma_{k,j} - \left(\frac{1}{\epsilon}-1\right)(a_k^j \tilde{x}_{k,t} + d_{k,j})^2\right\}}\right) \geq \frac{1}{36}$$

Proof. By Lemma C.3, we have that for any unit vector $w \in \mathbb{R}^2$,

$$\mathbb{P}\left\{|\langle w, \overline{x}_{k,t+j}\rangle| \geq \frac{1}{\sqrt{2}}\sqrt{\left(w_1\left(a_k^j \tilde{x}_{k,t} + d_{k,j}\right) + w_2\right)^2 + w_1^2 \Gamma_{k,j}} \,\middle|\, \overline{x}_{k,t}\right\} \geq \frac{1}{36}.$$

For all $\epsilon \in (0,1)$, we have

$$\begin{aligned}
((w_1(a_k^j \tilde{x}_{k,t} + d_{k,j}) + w_2)^2 + w_1^2\Gamma_{k,j} &= \left(w_1\left(a_k^j \tilde{x}_{k,t} + d_{k,j}\right)\right)^2 + w_2^2 + 2w_2 w_1\left(a_k^j \tilde{x}_{k,t} + d_{k,j}\right) + w_1^2\Gamma_{k,j} \\
&\geq (1-\epsilon)w_2^2 - \left(\frac{1}{\epsilon}-1\right)\left(w_1\left(a_k^j \tilde{x}_{k,t} + d_{k,j}\right)\right)^2 + w_1^2\Gamma_{k,j} \\
&\geq \min\left\{1-\epsilon, \Gamma_{k,j} - \left(\frac{1}{\epsilon}-1\right)(a_k^j \tilde{x}_{k,t} + d_{k,j})^2\right\}.
\end{aligned}$$

□

Lemma C.6

Fix $\delta \in (0,1)$. $\{\overline{x}_{k,t}\}_{t=1}^n$ satisfy that for any unit vector $w \in \mathbb{R}^2$,

$$\mathbb{P}\left(\sum_{t=1}^n \langle w, \overline{x}_{k,t}\rangle^2 \leq \frac{\psi^2 p^2}{16} j_\star \lfloor n/j_\star \rfloor\right) \leq \exp\left(-\frac{3\lfloor n/j_\star \rfloor p}{4}\right) + \delta$$

with $p = 1/72$,

$$j_\star := \left\lceil \max\left\{-\log_{a_k}\left(1 + (1-a_k)\frac{\sqrt{2\Gamma_{k,\infty}\log(n/\delta)}}{d_k}\right), -\log_{a_k}\sqrt{2}\right\}\right\rceil,$$

$$\psi := \sqrt{\min\left\{\frac{\Gamma_{k,\infty}}{\frac{16 d_k^2}{(1-a_k)^2} + \Gamma_{k,\infty}}, \frac{\Gamma_{k,\infty}}{4}\right\}}.$$

Proof. Fix $\delta \in (0,1)$. Recall that from Lemma C.4, we have shown that for all $t \geq 0$ and $k \geq 1$,

given a unit vector $w \in \mathbb{R}^2$, for any $\epsilon \in (0,1)$, we have

$$\mathbb{P}\left\{|\langle w, \overline{x}_{k,t+j}\rangle| \geq \frac{1}{\sqrt{2}}\sqrt{\min\left\{1-\epsilon, \Gamma_{k,j} - \left(\frac{1}{\epsilon}-1\right)(a_k^j \tilde{x}_{k,t} + d_{k,j})^2\right\}}\right\} \geq \frac{1}{36}.$$

Denote $q_{t,j} = a_k^j \tilde{x}_{k,t} + d_{k,j}$ where $\tilde{x}_{k,t} \sim \mathcal{N}(d_{k,t}, \Gamma_{k,t})$. Fix $\delta \in (0,1)$. Using the standard Gaussian tail bound and the union bound, we have that with probability $1-\delta$,

$$\max_{t\in[T]} q_{t,j} \leq a_k^j \left(\frac{d_k}{1-a_k} + \sqrt{2\Gamma_\infty \log(n/\delta)}\right) + \frac{d_k}{1-a_k}.$$

When $j \geq j_\star$, $\Gamma_{k,j} \geq \Gamma_{k,\infty}/2$, and with probability $1-\delta$, $\max_{t\in[T]} q_{t,j} \leq \frac{2d_k}{1-a_k}$. Thus, for $j \geq j_\star$, and

$$\epsilon = \frac{\frac{4d_k^2}{(1-a_k)^2}}{\frac{4d_k^2}{(1-a_k)^2} + \Gamma_\infty/4},$$

we have

$$\begin{aligned}
\nu_{t,j}^2 &:= \min\left\{1-\varepsilon, \Gamma_{k,j} - \left(\frac{1}{\varepsilon}-1\right) q_{t,j}^2\right\} \\
&\geq \min\left\{1-\varepsilon, \Gamma_{k,\infty}/2 - \left(\frac{1}{\varepsilon}-1\right)\frac{4d_k^2}{(1-a_k)^2}\right\} \\
&\geq \min\left\{\frac{\Gamma_{k,\infty}}{\frac{16d_k^2}{(1-a_k)^2} + \Gamma_{k,\infty}}, \frac{\Gamma_{k,\infty}}{4}\right\} = \psi^2.
\end{aligned}$$

Putting it altogether, we have

$$\frac{1}{2j_\star}\sum_{j=1}^{2j_\star} \mathbb{P}\left(|\langle w, \overline{x}_{k,t+j}\rangle| \geq \nu_{t,j}/\sqrt{2} | \mathcal{F}_t\right) \geq \frac{1}{2j_\star}\sum_{j=j_\star}^{2j_\star} \mathbb{P}(|\langle w, \overline{x}_{k,t+j}\rangle| \geq \nu_{t,j_\star}/\sqrt{2} | \mathcal{F}_t) \geq \frac{1}{72}.$$

Further, we have

$$\mathbb{P}\left(\min_{t\in[T]} \nu_{t,j_\star}^2 \geq \psi^2\right) \geq 1-\delta.$$

Applying Lemma C.4, we have that for $p = \frac{1}{72}$,

$$\mathbb{P}\left(\sum_{t=1}^n \langle w, \overline{x}_{k,t}\rangle^2 \leq \frac{\psi^2 p^2}{16} j_\star \lfloor n/j_\star \rfloor\right) \leq \exp\left(-\frac{3\lfloor n/j_\star \rfloor p}{4}\right) + \delta.$$

□

Proof of Theorem C.2. Based on our setup, condition (a) of Theorem C.3 is satisfied. For any n, using Lemma C.6 with $\delta = \exp(-n)$, we have that

$$\forall w \in \mathbb{R}^2, \quad \mathbb{P}\left(\sum_{t=1}^{n} \langle w, \overline{x}_{k,t}\rangle^2 \leq \frac{\psi^2 p^2}{16} j_\star \lfloor n/j_\star \rfloor\right) \leq \exp\left(-\frac{3\lfloor n/j_\star \rfloor p}{4}\right) + \delta \leq 2\exp\left(-\frac{3\lfloor n/j_\star \rfloor p}{4}\right),$$

with $p = 1/72$,

$$j_\star := \left\lceil \max\left\{ -\log_{a_k}\left(1 + (1-a_k)\frac{\sqrt{2\Gamma_{k,\infty}(\log(n)+n)}}{d_k}\right), -\log_{a_k}\sqrt{2}\right\}\right\rceil,$$

$$\psi := \sqrt{\min\left\{\frac{\Gamma_{k,\infty}}{\frac{16 d_k^2}{(1-a_k)^2} + \Gamma_{k,\infty}}, \frac{\Gamma_{k,\infty}}{4}\right\}}.$$

Thus, we have provided a similar result to (C.9), which is what condition (b) of Theorem C.3 is used for. In this case, we have $\Gamma_{\text{sb}} = \psi I$ where I is a 2×2 identity matrix. Finally, to verify condition (c), we notice that we have

$$\overline{\Gamma}_{k,j} := \mathbb{E}[\overline{x}_{k,j}\overline{x}_{k,j}^\top] = \begin{pmatrix} \frac{b_k^2(1-a_k^{j-1})^2}{(1-a_k)^2} + \frac{\sigma_{z,k}^2(1-a_k^{2j-2})}{1-a_k^2} & \frac{(1-a_k^{j-1})b_k}{1-a_k} \\ \frac{(1-a_k^{j-1})b_k}{1-a_k} & 1 \end{pmatrix}.$$

and we denote

$$\overline{\Gamma} := \overline{\Gamma}_{k,n} + \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix} + \Gamma_{\text{sb}},$$

which gives that $0 \prec \Gamma_{\text{sb}} \prec \overline{\Gamma}$ and for all $j \geq 1$, $0 \preceq \overline{\Gamma}_{k,j} \prec \overline{\Gamma}$. Then, we have that

$$\begin{aligned}
\mathbb{P}\left(\overline{\mathbf{X}}_k^\top \overline{\mathbf{X}}_k \not\preceq \frac{2n}{\delta}\overline{\Gamma}\right) &= \mathbb{P}\left(\lambda_{\max}\left((n\overline{\Gamma})^{-1/2}\overline{\mathbf{X}}_k^\top \overline{\mathbf{X}}_k (n\overline{\Gamma})^{-1/2}\right) \geq \frac{2}{\delta}\right) \\
&\leq \frac{\delta}{2}\mathbb{E}\left[\lambda_{\max}\left((n\overline{\Gamma})^{-1/2}\overline{\mathbf{X}}_k^\top \overline{\mathbf{X}}_k (n\overline{\Gamma})^{-1/2}\right)\right] \\
&\leq \frac{\delta}{2}\mathbb{E}\left[\text{tr}\left((n\overline{\Gamma})^{-1/2}\overline{\mathbf{X}}_k^\top \overline{\mathbf{X}}_k (n\overline{\Gamma})^{-1/2}\right)\right] \leq \delta,
\end{aligned}$$

where the last inequality is true since $\mathbb{E}\left[\overline{\mathbf{X}}_k^\top \overline{\mathbf{X}}_k\right] = \sum_{j=1}^{n} \Gamma_{k,j} \preceq n\overline{\Gamma}$ (for all $j \in [n]$, $\text{trace}(\overline{\Gamma} - \overline{\Gamma}_{k,j}) > 0$ and $\det(\overline{\Gamma} - \overline{\Gamma}_{k,j}) > 0$). Following Theorem C.3, for $\delta \in (0,1)$, when the number of samples satisfy that

$$\frac{n}{j_\star} \geq \frac{10}{p^2}\left(\log(1/\delta) + 4\log(10/p) + \log\det(\overline{\Gamma}\Gamma_{\text{sb}}^{-1})\right),$$

we have that

$$\mathbb{P}\left(\|\widehat{A}_k - A_k\|_2 > \frac{90\sigma_{z,k}}{p}\sqrt{\frac{1 + 2\log(10/p) + \log\det\left(\overline{\Gamma}\Gamma_{\text{sb}}^{-1}\right) + \log(1/\delta)}{n\psi}}\right) \leq 3\delta.$$

□

C.4.2 Proof of Corollary 4.1 and Corollary C.2

Similar to Appendix C.4.1, Corollary 4.1 is a special case of Corollary C.2 when $m = 1$. Hence, we directly present the proof of Corollary C.2 below.

Proof of Corollary C.2. Fix $\delta \in (0, 1)$. We have that with probability $1 - \delta$, $\epsilon(n, \delta, k) := \|\widehat{A}_k - A_k\|_2 \leq O(1/\sqrt{n})$. With probability at least $1 - \frac{\delta}{K}$, $\epsilon_{a_k} := |\widehat{a}_k - a_k| \leq \|\widehat{A}_k - A_k\|_2 = \epsilon(n, \delta/K, k) = O(1/\sqrt{n})$ and $\epsilon_{d_k} := |\widehat{d}_k - b_k| \leq \|\widehat{A}_k - A_k\|_2 = \epsilon(n, \delta/K, k) = O(1/\sqrt{n})$. When $m = 1$, then $|\widehat{\gamma}_k - \gamma_k| = ||\widehat{a}_k| - a_k| \leq |\widehat{a}_k - a_k| = \epsilon_{a_k} \leq \epsilon(n, \delta/K, k)$. When $m \geq 2$, since $\gamma_k \neq 0$, we have that

$$|\widehat{\gamma}_k - \gamma_k| = \left| \frac{|\widehat{a}_k| - a_k}{|\widehat{a}_k|^{(m-1)/m} + |\widehat{a}_k|^{(m-2)/m}\gamma_k + \ldots + \gamma_k^{m-1}} \right| \leq \frac{|\widehat{a}_k - a_k|}{\gamma_k^{m-1}}.$$

On the other hand, we obtain that

$$|\widehat{\lambda}_k - \lambda_k| = \left| \left| \frac{\widehat{d}_k}{\widehat{a}_k} \right| - \frac{d_k}{a_k} \right| \leq \left| \frac{\widehat{d}_k}{\widehat{a}_k} - \frac{d_k}{\widehat{a}_k} + \frac{d_k}{\widehat{a}_k} - \frac{d_k}{a_k} \right| \leq \frac{\epsilon_{d_k}}{\widehat{a}_k} + \frac{\lambda_k \epsilon_{a_k}}{\widehat{a}_k} \leq O\left(\frac{1}{\sqrt{n}}\right).$$

The proof completes as follows:

$$\mathbb{P}\left(\forall k \in [K], |\widehat{\gamma}_k - \gamma_k| \leq O(1/\sqrt{n}), |\widehat{\lambda}_k - \lambda_k| \leq O(1/\sqrt{n})\right) \geq \prod_{k=1}^{K} \left(1 - \frac{\delta}{K}\right) \geq 1 - \delta,$$

where the last inequality follows from Bernoulli's inequality. □

C.5 Additional Proofs and Discussion of Section 4.6

C.5.1 Proof of Theorem 4.4

Lemma C.7

Consider any episode $i+1$ (from time t_i+1 to t_{i+1}) where the initial state $x^i = (\mu_{1,t_i+1}(u_{1,0:t_i}), n_{1,t_i+1}, \ldots, \mu_{K,t_i+1}(u_{K,0:t_i}), n_{K,t_i})$ and $\{u_{k,0:t_i}\}_{k=1}^K$ are the past pull sequences of the proposed policy $\pi_{1:t_i}$. For all $\tilde{\pi}_{t_i+1:t_{i+1}}$ such that $\tilde{\pi}_t = \tilde{\pi}_t(x_t) = \tilde{\pi}_t(x'_t), \tilde{\pi}_t \in [K], \forall t \in [t_i+1, t_{i+1}], x_t, x'_t \in \mathcal{X}$, we have that

$$\sum_{t=t_i+1}^{t_{i+1}} \mathbb{E}_{x_{t_i+2},\ldots,x_{t_i}}\left[r(x_t, \tilde{\pi}_t(x_t))|x_{t_i+1} = x^i\right] = \sum_{t=t_i+1}^{t_{i+1}} \mu_{k,t}(u_{k,0:t-1}),$$

where $\{u_{k,t_i+1:t_{i+1}}\}_{k=1}^K$ is the arm pull sequence of $\tilde{\pi}_{t_i+1:t_{i+1}}$.

Proof. Let k denote $\tilde{\pi}_t$ where $t \in \{t_i+1, \ldots, t_{i+1}\}$. Recall that we use $u_{k,0:t-1}$ to denote the pull sequence of arm k under policy $\tilde{\pi}_{1:t_{i+1}} = (\pi_{1:t_i}, \tilde{\pi}_{t_i+1:t_{i+1}})$. If k has not been pulled before time t by $\tilde{\pi}_{1:t_{i+1}}$, then $\mathbb{E}_{x_{t_i+2},\ldots,x_{t_{i+1}}}[r(x_t, \tilde{\pi}_t)|x_{t_i+1} = x^i] = b_{\pi_t} = \mu_{\pi_t,t}(u_{\pi_t,0:t-1})$. If k has been pulled before, then let $q_1, \ldots, q_n$ denote the time steps that arm k has been pulled before time t by $\tilde{\pi}_{1:t_{i+1}}$, i.e., $u_{k,q_i} = 1$ for $i \in [n]$ and $u_{k,t'} = 0$ for $t' \notin \{q_1, \ldots, q_n\}$. We have that for $t \in \{t_i+1, \ldots, t_{i+1}\}$,

$$\begin{aligned}
&\mathbb{E}_{x_{t_i+2},\ldots,x_{t_{i+1}}}\left[r(x_t, \tilde{\pi}_t)|x_{t_i+1} = x^i\right] \\
=&b_k - \left(\mathbb{E}_{x_{t_i+2},\ldots,x_{t_{i+1}-1}}\left[\mathbb{E}_{x_{t_{i+1}}}\left[\gamma_k^{n_{k,t_{i+1}}} x_{k,t_{i+1}} + \lambda_k \gamma_k^{n_{k,t_{i+1}}}\right]|x_{t_i+1} = x^i\right]\right) \\
=&b_k - \left(\mathbb{E}_{x_{t_i+2},\ldots,x_{q_n}}\left[\mathbb{E}_{x_{q_n+1}}\left[\gamma_k^{n_{k,t_{i+1}}} x_{k,q_n+1} + \lambda_k \gamma_k^{n_{k,t_{i+1}}}\right]|x_{t_i+1} = x^i\right]\right) \\
=&b_k - \left(\mathbb{E}_{x_{t_i+2},\ldots,x_{q_n}}\left[\gamma_k^{n_{k,t_{i+1}}}\left(\gamma_k^{n_{k,q_n}} x_{k,q_n} + \lambda_k \gamma_k^{n_{k,q_n}}\right) + \lambda_k \gamma_k^{n_{k,t_{i+1}}}|x_{t_i+1} = x^i\right]\right) \\
=&\ldots = b_k - \lambda_k\left(\gamma_k^{n_{k,t_{i+1}}} + \gamma_k^{n_{k,t_{i+1}}+n_{k,q_n}} + \ldots + \gamma_k^{n_{k,t_{i+1}}+n_{k,q_n}+\ldots n_{k,q_1}}\right) \\
=&\mu_{k,t}(u_{k,0:t-1}),
\end{aligned}$$

where the second equality is true because when arm k is not pulled for example at time $t_{i+1}-1$, the state for arm k at time $t_{i+1}-1$ will satisfy that $x_{k,t_{i+1}} = x_{k,t_{i+1}-1}$ and $n_{k,t_{i+1}} = n_{k,t_{i+1}-1}+1$ with probability 1. In this case, we have that

$$\begin{aligned}
\mathbb{E}_{x_{t_{i+1}}}\left[\gamma_k^{n_{k,t_{i+1}}} x_{k,t_{i+1}} + \lambda_k \gamma_k^{n_{k,t_{i+1}}}\middle|x_{t_{i+1}-1}\right] &= \gamma_k^{n_{k,t_{i+1}-1}+1} x_{k,t_{i+1}-1} + \lambda_k \gamma_k^{n_{k,t_{i+1}-1}+1} \\
&= \gamma_k^{n_{k,t_{i+1}}} x_{k,t_{i+1}-1} + \lambda_k \gamma_k^{n_{k,t_{i+1}}}.
\end{aligned}$$

The third equality is true since when arm k is pulled for example at time q_n, then we have that

$$\mathbb{E}_{q_{n+1}\sim p_{\mathcal{M}}(\cdot|x_{q_n},k,q_n)}\left[\gamma_k^{n_{k,t_{i+1}}} x_{k,q_n+1} + \lambda_k \gamma_k^{n_{k,t_{i+1}}}\right]$$

$$= \gamma_k^{n_{k,t_{i+1}}} \left(\gamma_k^{n_{k,q_n}} x_{k,q_n} + \lambda_k \gamma_k^{n_{k,q_n}} \right) + \lambda_k \gamma_k^{n_{k,t_{i+1}}},$$

where $p_{\mathcal{M}}$ is given in Appendix C.3.1. The second to last last equality holds because $x_{k,t_i+1} = \mu_{k,t_i+1}(u_{k,0:t_i})$ where $\mu_{k,t}(\cdot)$ is defined in (4.3). □

Lemma C.8

For any episode $i+1$ (from time t_i+1 to t_{i+1}), given the past arm pull sequences $\{u_{k,0:t_i}\}_{k=1}^K$ of the proposed policy $\pi_{1:t_i}$, the optimal time-dependent competitor policy $\tilde{\pi}_{t_i+1:t_{i+1}}$, where $\tilde{\pi}_t = \tilde{\pi}_t(x_t) = \tilde{\pi}_t(x_t'), \tilde{\pi}_t \in [K], \forall t \in [t_i+1, t_{i+1}], x_t, x_t' \in \mathcal{X}$, for this episode is given by $\textsc{Lookahead}(\{\lambda_k, \gamma_k, b_k\}_{k=1}^K, \{u_{k,0:t_i}\}_{k=1}^K, t_i, t_{i+1})$ where $\{\lambda_k, \gamma_k, b_k\}_{k=1}^K$ are the true reward parameters for the rebounding bandits instance.

Proof. By Lemma C.7, we have that the optimal time-dependent competitor policy $\tilde{\pi}_{t_i+1:t_{i+1}}$ maximizes $\sum_{t=t_i+1}^{t_{i+1}} \mu_{k,t}(u_{k,0:t-1})$, by choosing $u_{k,t_i+1:t_{i+1}}$. Thus, by the definition of $\textsc{Lookahead}$ (4.5), given our proposed policy $\pi_{1:t_i}$, the optimal time-dependent competitor policy is given by $\textsc{Lookahead}(\{\lambda_k, \gamma_k, b_k\}_{k=1}^K, \{u_{k,0:t_i}\}_{k=1}^K, t_i, t_{i+1})$. □

Proof of Theorem 4.4. Using Lemma C.8, we have that given our policy $\pi_{1:T}$ and its corresponding pull sequence $u_{k,0:t-1}$ for $k \in [K], t \in [T]$, the optimal competitor policy for episode $i+1$ where $i \in \{0, \ldots, \lfloor T/w \rfloor\}$ (episode $i+1$ ranges from time $t_i + 1 = iw + 1$ to $t_{i+1} = \min\{iw + w, T\}$) is given by $\textsc{Lookahead}(\{\lambda_k, \gamma_k, b_k\}_{k=1}^K, \{u_{k,0:t_i}\}_{k=1}^K, t_i, t_{i+1})$. We use $\mathbf{M}(\{\lambda_k, \gamma_k, b_k\}_{k=1}^K, \{u_{k,0:t_i}\}_{k=1}^K, t_i, t_{i+1})$ to denote the (optimal) objective value of (4.5) given by $\textsc{Lookahead}(\{\lambda_k, \gamma_k, b_k\}_{k=1}^K, \{u_{k,0:t_i}\}_{k=1}^K, t_i, t_{i+1})$. Denote $\overline{b} = \max_k b_k$ and $\underline{b} = \min_k b_k$.

Exploration Stage Recall that in Algorithm 4.1, we have defined $\widetilde{T} = T^{2/3} + w - (T^{2/3} \bmod w)$ which is a multiple of w. For the first $\widetilde{T}$ time steps, as defined in Algorithm 4.1, our policy $\pi_{1:\widetilde{T}}$ is a time-dependent policy, i.e., it satisfies that $\pi_t = \pi_t(x_t) = \pi_t(x_t'), \pi_t \in [K], \forall t \in [1, \widetilde{T}], x_t, x_t' \in \mathcal{X}$. Using C.7, we obtain that the regret for the first $\widetilde{T}/w$ episodes is given by

$$\begin{aligned}
&\sum_{i=0}^{\tilde{T}/w-1} \max_{\tilde{\pi}_{1:w} \in \mathcal{C}^w} \mathbb{E}\left[\sum_{j=1}^{w} r(x_{iw+j}, \tilde{\pi}_j(x_{iw+j})) \middle| x_{iw+1} = x^i \right] \\
&\qquad - \sum_{i=0}^{\tilde{T}/w-1} \mathbb{E}\left[r(x_{iw+j}, \pi_{iw+j}(x_{iw+j})) \middle| x_{iw+1} = x^i \right] \\
\leq &\sum_{i=0}^{\tilde{T}/w-1} \mathbf{M}(\{\lambda_k, \gamma_k, b_k\}_{k=1}^K, \{u_{k,0:iw}\}_{k=1}^K, iw, iw+w) - \widetilde{T}\left(\underline{b} - \frac{\overline{\lambda}\overline{\gamma}}{1-\overline{\gamma}} \right) \\
\leq &\widetilde{T}\left(\overline{b} - \underline{b} + \frac{\overline{\lambda}\overline{\gamma}}{1-\overline{\gamma}} \right) \lesssim \widetilde{T} \lesssim T^{2/3}.
\end{aligned}$$

since $\widetilde{T} \leq T^{2/3} + w$ and by assumption, $w \leq T^{2/3}$.

Estimation Stage By Theorem 4.3 and Corollary 4.1, we have that for any $\delta \in (0,1)$ and $n \geq n_0(\delta, k)$ where $n_0(\delta, k)$ depends on δ logarithmically, with probability $1-\delta$, for all $k \in [K]$ $|\widehat{\gamma}_k - \gamma_k| \leq \frac{C_{\gamma_k} \log(1/\delta)}{\sqrt{n}}$ and $|\widehat{\lambda}_k - \lambda_k| \leq \frac{C_{\lambda_k} \log(1/\delta)}{\sqrt{n}}$ when $\widehat{\gamma}_k > 0$.

We define two numbers $T_0' := \min_T\{T : (\sum_{k=1}^K n_0(k, T^{-1/3}))^{3/2} = C_1 K (\log T)^{3/2} < T\}$ and $T_0'' := \min_T \left\{ T : \max_k \gamma_k + \frac{C_{\gamma_k}}{\sqrt{T^{2/3}/K}} < 1 \right\}$. These two numbers exist as T can be chosen to be arbitrarily large. Take $T_0 = \max\{T_0', T_0''\}$. Then for all $T \geq T_0$, with probability $1-\delta$ where $\delta = T^{-1/3}$, we have that $\forall k \in [K], |\widehat{\gamma}_k - \gamma_k| \leq \epsilon_\gamma = O(\sqrt{K} T^{-1/3} \log T)$, $|\widehat{\lambda}_k - \lambda_k| \leq \epsilon_\lambda = O(\sqrt{K} T^{-1/3} \log T)$ and $\left(\epsilon_\lambda \left|\frac{\widehat{\gamma}_k}{1-\widehat{\gamma}_k}\right| + \epsilon_\gamma \left|\frac{\overline{\lambda}}{(1-\widehat{\gamma}_k)(1-\gamma_k)}\right|\right) \leq O(\sqrt{K} T^{-1/3} \log T)$ since $\widehat{\gamma}_k \leq \gamma_k + \frac{C_{\gamma_k}}{\sqrt{T_0^{2/3}/K}} < 1$ and $\gamma_k \leq \overline{\gamma} < 1$.

For any pull sequence $u_{k,0:t-1}$, using our obtained estimated parameters $\{\widehat{\gamma}_k, \widehat{\lambda}_k, \widehat{b}_k\}_{k=1}^K$, we define the estimated reward function: for $t \geq 2$, $\widehat{\mu}_{k,t}(u_{k,0:t-1}) = b_k - \widehat{\lambda}_k \left(\sum_{i=1}^{t-1} \widehat{\gamma}_k^{t-i} u_{k,i}\right)$, and for $t = 1$, $\widehat{\mu}_{k,1}(u_{k,0:1}) = b_k = \mu_{k,1}(u_{k,0:1})$, where we note that $\widehat{b}_k = b_k$ since it is the reward of the first pull of arm k. Given $t \geq 2$, we have that

$$
\begin{aligned}
&|\mu_{k,t}(u_{k,0:t-1}) - \widehat{\mu}_{k,t}(u_{k,0:t-1})| \\
=&\left|\widehat{\lambda}_k \left(\sum_{i=1}^{t-1} \widehat{\gamma}_k^{t-i} u_{k,i}\right) - \lambda_k \left(\sum_{i=1}^{t-1} \gamma_k^{t-i} u_{k,i}\right)\right| \\
=&\left|\widehat{\lambda}_k \left(\sum_{i=1}^{t-1} \widehat{\gamma}_k^{t-i} u_{k,i}\right) - \lambda_k \left(\sum_{i=1}^{t-1} \widehat{\gamma}_k^{t-i} u_{k,i}\right) + \lambda_k \left(\sum_{i=1}^{t-1} \widehat{\gamma}_k^{t-i} u_{k,i}\right) - \lambda_k \left(\sum_{i=1}^{t-1} \gamma_k^{t-i} u_{k,i}\right)\right| \\
\leq&|\widehat{\lambda}_k - \lambda_k| \left|\frac{\widehat{\gamma}_k}{1-\widehat{\gamma}_k}\right| + \overline{\lambda} \left|\frac{\widehat{\gamma}_k}{1-\widehat{\gamma}_k} - \frac{\gamma_k}{1-\gamma_k}\right| \\
\leq&\epsilon_\lambda \left|\frac{\widehat{\gamma}_k}{1-\widehat{\gamma}_k}\right| + \epsilon_\gamma \left|\frac{\overline{\lambda}}{(1-\widehat{\gamma}_k)(1-\gamma_k)}\right|.
\end{aligned} \tag{C.10}
$$

Planning Stage Given our policy $\pi_{1:T}$ (along with its pull sequence $\{u_{k,0:T}\}_{k=1}^K$), starting from time $\widetilde{T}+1$, for any episode $i+1 \geq \widetilde{T}/w$, we denote the optimal competitor policy to be $\pi^*_{t_i+1:t_{i+1}} = \texttt{Lookahead}(\{\lambda_k, \gamma_k, b_k\}_{k=1}^K, \{u_{k,0:t_{i-1}}\}_{k=1}^K, t_i, t_{i+1})$ where $t_i = iw$ and $t_{i+1} = \min\{iw+w, T\}$. The cumulative expected reward collected by $\pi^*_{t_i+1:t_{i+1}}$ and $\pi_{t_i+1:t_{i+1}}$ has the difference

$$
\begin{aligned}
&\mathbf{M}(\{\lambda_k, \gamma_k, b_k\}_{k=1}^K, \{u_{k,0:t_{i-1}}\}_{k=1}^K, t_i, t_{i+1}) - \mathbf{M}(\{\widehat{\lambda}_k, \widehat{\gamma}_k, b_k\}_{k=1}^K, \{u_{k,0:t_{i-1}}\}_{k=1}^K, t_i, t_{i+1}) \\
=&\sum_{t=t_i+1}^{t_{i+1}} \mu_{\pi^*_t,t}(u^*_{\pi^*_t,0:t-1}) - \sum_{t=t_i+1}^{t_{i+1}} \mu_{\pi_t,t}(u_{\pi_t,0:t-1}) \\
=&\sum_{t=t_i+1}^{t_{i+1}} \mu_{\pi^*_t,t}(u^*_{\pi^*_t,0:t-1}) - \sum_{t=t_i+1}^{t_{i+1}} \widehat{\mu}_{\pi^*_t,t}(u^*_{\pi^*_t,0:t-1}) \\
&+ \sum_{t=t_i+1}^{t_{i+1}} \widehat{\mu}_{\pi^*_t,t}(u^*_{\pi^*_t,0:t-1}) - \sum_{t=t_i+1}^{t_{i+1}} \widehat{\mu}_{\pi_t,t}(u_{\pi_t,0:t-1})
\end{aligned}
$$

$$
\begin{aligned}
&+ \sum_{t=t_i+1}^{t_{i+1}} \widehat{\mu}_{\pi_t,t}(u_{\pi_t,0:t-1}) - \sum_{t=t_i+1}^{t_{i+1}} \mu_{\pi_t,t}(u_{\pi_t,0:t-1}) \\
\leq &\sum_{t=t_i+1}^{t_{i+1}} \mu_{\pi_t^*,t}(u^*_{\pi_t^*,0:t-1}) - \sum_{t=t_i+1}^{t_{i+1}} \widehat{\mu}_{\pi_t^*,t}(u^*_{\pi_t^*,0:t-1}) \\
&+ \sum_{t=t_i+1}^{t_{i+1}} \widehat{\mu}_{\pi_t,t}(u_{\pi_t,0:t-1}) - \sum_{t=t_i+1}^{t_{i+1}} \mu_{\pi_t,t}(u_{\pi_t,0:t-1}).
\end{aligned}
$$

where $u^*_{\pi_t^*,0:t-1}$ is the corresponding pull sequence of arm π_t^* under policy $\pi^*_{1:t} = (\pi_{1:t_i}, \pi^*_{t_i+1:t})$, and the last inequality holds because $\pi_{t_i+1:t_{i+1}} = \texttt{Lookahead}(\{\widehat{\lambda}_k, \widehat{\gamma}_k, \widehat{b}_k\}_{k=1}^K, \{u_{k,0:t_i}\}_{k=1}^K, t_i, t_{i+1})$ is the optimal solution under the estimated parameters $\{\widehat{\lambda}_k, \widehat{\gamma}_k, \widehat{b}_k\}_{k=1}^K$ and π's previous past pull sequence $\{u_{k,0:t_i}\}_{k=1}^K$. Further, using (C.10) and the fact that $t_i - t_{i-1} \leq w$, we obtain that

$$
\begin{aligned}
&\sum_{t=t_i+1}^{t_{i+1}} \mu_{\pi_t^*,t}(u^*_{\pi_t^*,0:t-1}) - \sum_{t=t_i+1}^{t_{i+1}} \widehat{\mu}_{\pi_t^*,t}(u^*_{\pi_t^*,0:t-1}) \\
&\quad + \sum_{t=t_i+1}^{t_{i+1}} \widehat{\mu}_{\pi_t,t}(u_{\pi_t,0:t-1}) - \sum_{t=t_i+1}^{t_{i+1}} \mu_{\pi_t,t}(u_{\pi_t,0:t-1}) \\
\leq &2w \max_k \left(\epsilon_\lambda \left| \frac{\widehat{\gamma}_k}{1-\widehat{\gamma}_k} \right| + \epsilon_\gamma \left| \frac{\overline{\lambda}}{(1-\widehat{\gamma}_k)(1-\gamma_k)} \right| \right).
\end{aligned}
$$

Finally, putting it altogether, we have obtained that for all $T \geq T_0$,

$$
\begin{aligned}
&\mathrm{Reg}^w(T) = \textstyle\sum_{i=0}^{\lceil T/w \rceil - 1} \max_{\tilde{\pi}_{1:w} \in \mathcal{C}^w} \mathbb{E}\left[\sum_{j=1}^{\min\{w, T-iw\}} r(x_{iw+j}, \tilde{\pi}_j(x_{iw+j})) \middle| x_{iw+1} = x^i \right] \\
&\qquad - \mathbb{E}\left[\sum_{j=1}^{\min\{w, T-iw\}} r(x_{iw+j}, \pi_{iw+j}(x_{iw+j})) \middle| x_{iw+1} = x^i \right] \\
\leq &O(T^{2/3}) + (1 - T^{-1/3}) \left(\sum_{i=T/w}^{\lceil T/w \rceil - 1} 2wO(\sqrt{K} T^{-1/3} \log T) \right) + T^{-1/3} \left(T \left(\overline{b} - \underline{b} + \frac{\overline{\lambda}\overline{\gamma}}{1-\overline{\gamma}} \right) \right) \\
\leq &O(T^{2/3}) + (T - T^{2/3}) O(\sqrt{K} T^{-1/3} \log T) + O(T^{2/3}) \\
\leq &O(\sqrt{K} T^{2/3} \log T),
\end{aligned}
$$

which we notice that with probability $\delta = T^{-1/3}$, the cumulative expected reward from time $\widetilde{T}$ to T between the optimal competitor policy and our policy π is at most $T\left(\overline{b} - \underline{b} + \frac{\overline{\lambda}\overline{\gamma}}{1-\overline{\gamma}}\right)$. This completes the proof. □

C.5.2 Exploration Strategies

In the exploration phase of Algorithm 4.1 (from time 1 to $\widetilde{T}$), in addition to playing each arm repeatedly for $\widetilde{T}/K$ times, in general, we could explore by playing each arm at a fixed interval,

i.e., the time interval between two consecutive pulls of arm k should be a constant m_k. For example, this includes playing the arms cyclically with the cylce being $1, 2, \ldots, K$ or playing the first two arms in an alternating fashion from time 1 to $2\widetilde{T}/K$, then the next two arms, etc. As shown in Theorem C.2 and Corollary C.2, using the datasets (of size n) collected by these exploration strategies, we can obtain estimators $\widehat{\gamma}_k$ and $\widehat{\lambda}_k$ with the estimation error being on the order of $O(1/\sqrt{n})$. Using these results (in replacement of Theorem 4.3 and Corollary 4.1 in the estimation stage of the proof of Theorem 4.4), we can obtain that there exists T_0 such that for all $T \geq T_0$, the regret upper bound of EEP under these exploration strategies are of order $O(\sqrt{K}T^{2/3}\log T)$.

C.6 Additional Proofs of Appendix C.3

C.6.1 Proof of Corollary C.1

Proof. Fix $\delta \in (0,1)$. By Theorem C.1, for all $k \in [K]$, with probability $1 - \frac{\delta}{2K}$, we have the following: When $m = 1$, then $|\widehat{\gamma}_k - \gamma_k| = ||\widehat{a}_k| - a_k| \le |\widehat{a}_k - a_k| \le \epsilon_a(n, \frac{\delta}{2K}, k)$. When $m \ge 2$, we have that

$$|\widehat{\gamma}_k - \gamma_k| = \left| \frac{|\widehat{a}_k| - a_k}{|\widehat{a}_k|^{(m-1)/m} + |\widehat{a}_k|^{(m-2)/m}\gamma_k + \ldots + \gamma_k^{m-1}} \right| \le \frac{|\widehat{a}_k - a_k|}{\gamma_k^{m-1}}.$$

On the other hand, given that $|\widehat{a}_k - a_k| \le \epsilon_a(n, \frac{\delta}{2K}, k)$, we have that with probability $1 - \frac{\delta}{2K}$,

$$|\widehat{\lambda}_k - \lambda_k| = \left| \left| \frac{\widehat{d}_k}{\widehat{a}_k} \right| - \frac{d_k}{a_k} \right| \le \left| \frac{\widehat{d}_k}{\widehat{a}_k} - \frac{d_k}{\widehat{a}_k} + \frac{d_k}{\widehat{a}_k} - \frac{d_k}{a_k} \right| \le \frac{\epsilon_d(n, \frac{\delta}{2K}, k)}{\widehat{a}_k} + \frac{\lambda_k \epsilon_a(n, \frac{\delta}{2K}, k)}{\widehat{a}_k} \le O\left(\frac{1}{\sqrt{n}}\right).$$

The proof completes as follows:

$$\begin{aligned} \mathbb{P}\left(\forall k, |\widehat{\gamma}_k - \gamma_k| \le \frac{|\widehat{a}_k - a_k|}{\gamma_k^{m-1}}, |\widehat{\lambda}_k - \lambda_k| \le \frac{\epsilon_d(n, \frac{\delta}{2K}, k)}{\widehat{a}_k} + \frac{\lambda_k \epsilon_a(n, \frac{\delta}{2K}, k)}{\widehat{a}_k} \right) &\ge \prod_{k=1}^{K} \left(1 - \frac{\delta}{2K}\right)^2 \\ &\ge 1 - \delta, \end{aligned}$$

where the last inequality follows from Bernoulli's inequality. □

C.6.2 Proof of Lemma C.2

Proof. Let $\pi_{1:T}$ denote the sequence that policy π will take from time 1 to T. By the definition of the value function, we have that

$$V^{\pi}_{1,\mathcal{M}}(x_{\text{init}}) = b_{\pi_1} + \sum_{t=2}^{T} \mathbb{E}_{x_2,\ldots,x_t}\left[r(x_t, \pi_t)\right],$$

where $x_t \sim p_{\mathcal{M}}(\cdot | x_{t-1}, \pi_{t-1}, t-1)$ is a state vector drawn from the transition distribution defined in Section C.3.1. Let k denote π_t and $u_{k,0:t-1}$ denote the past pull sequence for arm k under policy π. If k has not been pulled before time t, then $\mathbb{E}_{x_2,\ldots,x_t}[r(x_t, \pi_t)] = b_{\pi_t} = \mu_{\pi_t,t}(u_{\pi_t,0:t-1})$. If k has been pulled before, then let $t_1, \ldots, t_n$ denote the time steps that arm k has been pulled before time t. We have that

$$\begin{aligned} \mathbb{E}_{x_2,\ldots,x_t}\left[r(x_t, k)\right] &= b_k - \left(\mathbb{E}_{x_2,\ldots,x_{t-1}}\left[\mathbb{E}_{x_t \sim p_{\mathcal{M}}(\cdot|x_{t-1},k,t-1)}\left[\gamma_k^{n_{k,t}} x_{k,t} + \lambda_k \gamma_k^{n_{k,t}}\right]\right]\right) \\ &= b_k - \left(\mathbb{E}_{x_2,\ldots,x_{t_n}}\left[\mathbb{E}_{x_{t_n+1} \sim p_{\mathcal{M}}(\cdot|x_{t_n},k,t_n)}\left[\gamma_k^{n_{k,t}} x_{k,t_n+1} + \lambda_k \gamma_k^{n_{k,t}}\right]\right]\right) \\ &= b_k - \left(\mathbb{E}_{x_2,\ldots,x_{t_n}}\left[\gamma_k^{n_{k,t}}\left(\gamma_k^{n_{k,t_n}} x_{k,t_n} + \lambda_k \gamma_k^{n_{k,t_n}}\right) + \lambda_k \gamma_k^{n_{k,t}}\right]\right) \\ &= \ldots = b_k - \lambda_k \left(\gamma_k^{n_{k,t}} + \gamma_k^{n_{k,t}+n_{k,t_n}} + \ldots + \gamma_k^{n_{k,t}+n_{k,t_n}+\ldots n_{k,t_1}}\right) \end{aligned}$$

$$= \mu_{k,t}(u_{k,0:t-1}),$$

where we note that the second equality is true because when arm k is not pulled for example at time $t-1$, the state for arm k at time $t-1$ will satisfy that $x_{k,t} = x_{k,t-1}$ and $n_{k,t} = n_{k,t-1}+1$ with probability 1. In this case, we have that $\mathbb{E}_{x_t \sim p_{\mathcal{M}}(\cdot|x_{t-1},k,t-1)}\left[\gamma_k^{n_{k,t}} x_{k,t} + \lambda_k \gamma_k^{n_{k,t}}\right] = \gamma_k^{n_{k,t-1}+1} x_{k,t-1} + \lambda_k \gamma_k^{n_{k,t-1}+1} = \gamma_k^{n_{k,t}} x_{k,t-1} + \lambda_k \gamma_k^{n_{k,t}}$. The third equality is true since when arm k is pulled for example at time $t-1$, then we have that $\mathbb{E}_{x_t \sim p_{\mathcal{M}}(\cdot|x_{t-1},k,t-1)}\left[\gamma_k^{n_{k,t}} x_{k,t} + \lambda_k \gamma_k^{n_{k,t}}\right] = \gamma_k^{n_{k,t}}\left(\gamma_k^{n_{k,t-1}} x_{k,t-1} + \lambda_k \gamma_k^{n_{k,t-1}}\right) + \lambda_k \gamma_k^{n_{k,t}}$. The proof completes by summing over $\mathbb{E}_{x_2,\ldots,x_t}\left[r(x_t,\pi_t)\right]$ for all $t \geq 2$. □

C.6.3 Proof of Proposition C.1

Proof. Fix $\delta \in (0,1)$. Let E_1 be the event that

$$\forall k \in [K],\ |\widehat{\gamma}_k - \gamma_k| = \epsilon_{\gamma_k} \leq O\left(\frac{1}{\sqrt{n}}\right),\ |\widehat{\lambda}_k - \lambda_k| = \epsilon_{\lambda_k} \leq O\left(1/\sqrt{n}\right).$$

From Corollary C.1, we have that $\mathbb{P}(E_1) \geq 1-\delta$. Let $\pi_{1:T}$ denote the sequence that policy π will take from time 1 to T. From Lemma C.2, we have that

$$|V^{\pi}_{1,\mathcal{M}}(x_{\text{init}}) - V^{\pi}_{1,\widehat{\mathcal{M}}}(x_{\text{init}})| = \left|\sum_{t=1}^{T} \mu_{\pi_t,t}(u_{\pi_t,0:t-1}) - \widehat{\mu}_{\pi_t,t}(u_{\pi_t,0:t-1})\right|,$$

where $u_{\pi_t,0:t-1}$ is the past pull sequence for arm π_t under policy π before time t and $\widehat{\mu}_{k,t}(u_{k,0:t-1}) = b_k - \widehat{\lambda}_k\left(\sum_{i=1}^{t-1} \widehat{\gamma}_k^{t-i} u_{k,i}\right)$ for $t \geq 2$ and $\widehat{\mu}_{k,1}(u_{k,0:1}) = b_k = \mu_{k,1}(u_{k,0:1})$. Given $t \geq 2$, let k denote π_t, we have that

$$\begin{aligned}
&|\mu_{k,t}(u_{k,0:t-1}) - \widehat{\mu}_{k,t}(u_{k,0:t-1})| \\
=&\left|\widehat{\lambda}_k\left(\sum_{i=1}^{t-1}\widehat{\gamma}_k^{t-i}u_{k,i}\right) - \lambda_k\left(\sum_{i=1}^{t-1}\gamma_k^{t-i}u_{k,i}\right)\right| \\
=&\left|\widehat{\lambda}_k\left(\sum_{i=1}^{t-1}\widehat{\gamma}_k^{t-i}u_{k,i}\right) - \lambda_k\left(\sum_{i=1}^{t-1}\widehat{\gamma}_k^{t-i}u_{k,i}\right) + \lambda_k\left(\sum_{i=1}^{t-1}\widehat{\gamma}_k^{t-i}u_{k,i}\right) - \lambda_k\left(\sum_{i=1}^{t-1}\gamma_k^{t-i}u_{k,i}\right)\right| \\
\leq&|\widehat{\lambda}_k - \lambda_k|\left|\frac{\widehat{\gamma}_k}{1-\widehat{\gamma}_k}\right| + \overline{\lambda}\left|\frac{\widehat{\gamma}_k}{1-\widehat{\gamma}_k} - \frac{\gamma_k}{1-\gamma_k}\right| \\
\leq&\frac{\widehat{\gamma}_k \epsilon_{\lambda_k}}{1-\widehat{\gamma}_k} + \frac{\overline{\lambda}\epsilon_{\gamma_k}}{(1-\widehat{\gamma}_k)(1-\gamma_k)}
\end{aligned}$$

Since $\widehat{\gamma}_k < 1$ ($\widehat{a}_k \in (\underline{a}, \overline{a})$) almost surely and with probability $1-\delta$, for all $k \in [K]$, $\epsilon_{\gamma_k} \leq O\left(1/\sqrt{n}\right)$ and $\epsilon_{\lambda_k} \leq O\left(1/\sqrt{n}\right)$. We have that with probability $1-\delta$,

$$\left|\sum_{t=1}^{T} \mu_{\pi_t,t}(u_{\pi_t,0:t-1}) - \widehat{\mu}_{\pi_t,t}(u_{\pi_t,0:t-1})\right| \leq \sum_{t=1}^{T} |\mu_{\pi_t,t}(u_{\pi_t,0:t-1}) - \widehat{\mu}_{\pi_t,t}(u_{\pi_t,0:t-1})| \leq \left(\frac{T}{\sqrt{n}}\right).$$

□

C.6.4 Proof of Proposition C.2

Proof. Fix $\delta \in (0,1)$. Let E_1 be the event that

$$\forall k \in [K],\ |\widehat{\gamma}_k - \gamma_k| = \epsilon_{\gamma_k} \le O\left(\frac{1}{\sqrt{n}}\right),\ |\widehat{\lambda}_k - \lambda_k| = \epsilon_{\lambda_k} \le O\left(1/\sqrt{n}\right).$$

From Corollary C.2, we have that $\mathbb{P}(E_1) \ge 1 - \delta/2$. Let $\epsilon_\lambda := \max_k \epsilon_{\lambda_k}$. Let E_2 denote the event that $\forall t \in [T], k \in [K], |x_{k,t}| \le B(\delta/2)$ (C.1). We know that $\mathbb{P}(E_2) \ge 1 - \delta/2$. When E_1 and E_2 happen, we first observe that for all positive integer n and $k \in [K]$,

$$|\widehat{\gamma}_k^n - \gamma_k^n| \le |\widehat{\gamma}_k - \gamma_k| \left(n \max(\gamma_k^{n-1}, \widehat{\gamma}_k^{n-1})\right) \le \frac{|\widehat{\gamma}_k - \gamma_k|}{\max(\gamma_k, \widehat{\gamma}_k) \ln\left(1/\max(\gamma_k, \widehat{\gamma}_k)\right)} = O(1/\sqrt{n}),$$

whereand the second inequality uses the assumption that $\widehat{a}_k, \gamma_k$ are bounded away from 0 and 1.

To continue, we first bound the distance between the transition function in $\widehat{\mathcal{M}}$ and $\mathcal{M}$. At any any time t and state $x_t = (x_{1,t}, n_{1,t}, \ldots, x_{K,t}, n_{K,t})$, when we pull arm $\pi_t = k$, the next state x_{t+1} is updated by: (i) for arm k, $n_{k,t+1} = 1$ and (ii) for all other arms $k' \ne k$, $n_{k,t+1} = n_{k,t} + 1$ if $n_k \ne 0$, $n_{k,t+1} = 0$ if $n_{k,t} = 0$, and $x_{k',t+1} = x_{k',t}$. Then, by (Devroye et al., 2018, Theorem 1.3), we have that when $n_{\pi_t,t} \ne 0$,

$$\begin{aligned}
&\|p_{\widehat{\mathcal{M}}}(x_{t+1}|x_t, \pi_t, t) - p_{\mathcal{M}}(x_{t+1}|x_t, \pi_t, t)\|_1 \\
\overset{(*)}{\le}& \frac{3|\widehat{\lambda}_k^2 \sum_{i=0}^{n_{k,t}-1} \widehat{\gamma}_k^{2i} - \lambda_k^2 \sum_{i=0}^{n_{k,t}-1} \gamma_k^{2i}|}{\lambda_k^2 \sum_{i=0}^{n_k-1} \gamma_k^{2i}} + \frac{|\gamma_k^{n_{k,t}} x_{k,t} + \lambda_k \gamma_k^{n_{k,t}} - \widehat{\gamma}_k^{n_{k,t}} x_{k,t} - \widehat{\lambda}_k \widehat{\gamma}_k^{n_{k,t}}|}{\lambda_k \sqrt{\sum_{i=0}^{n_{k,t}-1} \gamma_k^{2i}}} \\
=& \frac{3|\widehat{\lambda}_k^2 \left(\sum_{i=0}^{n_{k,t}-1} \widehat{\gamma}_k^{2i} - \sum_{i=0}^{n_{k,t}-1} \gamma_k^{2i}\right) + (\widehat{\lambda}_k^2 - \lambda_k^2) \sum_{i=0}^{n_{k,t}-1} \gamma_k^{2i}|}{\lambda_k^2 \sum_{i=0}^{n_k-1} \gamma_k^{2i}} \\
&+ \frac{|\widehat{\gamma}_k^{n_{k,t}} - \gamma_k^{n_{k,t}}| B(\delta/2) + |\lambda_k \gamma_k^{n_{k,t}} - \widehat{\lambda}_k \widehat{\gamma}_k^{n_{k,t}}|}{\lambda_k \sqrt{\sum_{i=0}^{n_{k,t}-1} \gamma_k^{2i}}} \\
\overset{(**)}{\le}& 3\left|\widehat{\lambda}_k^2 \left(\sum_{i=0}^{n_{k,t}-1} \widehat{\gamma}_k^{2i} - \sum_{i=0}^{n_{k,t}-1} \gamma_k^{2i}\right) + (\widehat{\lambda}_k^2 - \lambda_k^2) \sum_{i=0}^{n_{k,t}-1} \gamma_k^{2i}\right| + |\widehat{\gamma}_k^{n_{k,t}} - \gamma_k^{n_{k,t}}| \left(B(\delta/2) + \lambda_k\right) \\
&+ |\lambda_k \widehat{\gamma}_k^{n_{k,t}} - \widehat{\lambda}_k \widehat{\gamma}_k^{n_{k,t}}| \\
\le& 3\left((\lambda_k + \epsilon_\lambda)^2 \left|\frac{1}{1 - \widehat{\gamma}_k^2} - \frac{1}{1 - \gamma_k^2}\right| + \frac{|\widehat{\lambda}_k - \lambda_k|(2\lambda_k + \epsilon_\lambda)}{1 - \gamma_k^2}\right) \\
&+ |\widehat{\gamma}_k^{n_{k,t}} - \gamma_k^{n_{k,t}}| \left(B(\delta/2) + \lambda_k\right) + |\lambda_k - \widehat{\lambda}_k| \\
=& 3\left(\frac{(\lambda_k + \epsilon_\lambda)^2 |\widehat{\gamma}_k^2 - \gamma_k^2|}{(1 - \widehat{\gamma}_k^2)(1 - \gamma_k^2)} + \frac{|\widehat{\lambda}_k - \lambda_k|(2\lambda_k + \epsilon_\lambda)}{1 - \gamma_k^2}\right) + |\widehat{\gamma}_k^{n_{k,t}} - \gamma_k^{n_{k,t}}| \left(B(\delta/2) + \lambda_k\right) + |\lambda_k - \widehat{\lambda}_k| \\
=:& \epsilon_P = O\left(\frac{1}{\sqrt{n}}\right),
\end{aligned}$$

where $(*)$ holds since $p_{\mathcal{M}}\left(x_{t+1}|x_t,\pi_t,t\right)$ is a Gaussian density with mean $\gamma_k^{n_{k,t}}x_{k,t}+\lambda_k\gamma_k^{n_{k,t}}$ and variance $\lambda_k^2\sum_{i=0}^{n_k-1}\gamma_k^{2i}$ and $(**)$ uses the fact that $\lambda_k^2\sum_{i=0}^{n_k-1}\gamma_k^{2i}\geq\lambda_k^2\geq 1$. When $n_{\pi_t,t}=0$ and condition (i) and (ii) are fulfilled, we have that $\|p_{\widehat{\mathcal{M}}}\left(x_{t+1}|x_t,\pi_t,t\right)-p_{\mathcal{M}}\left(x_{t+1}|x_t,\pi_t,t\right)\|_1=0$. Otherwise, that is, if condition (i) or (ii) is not satisfied, we also have that $\|p_{\widehat{\mathcal{M}}}\left(x_{t+1}|x_t,\pi_t,t\right)-p_{\mathcal{M}}\left(x_{t+1}|x_t,\pi_t,t\right)\|_1=0$ since $p_{\widehat{\mathcal{M}}}\left(x_{t+1}|x_t,\pi_t,t\right)=p_{\mathcal{M}}\left(x_{t+1}|x_t,\pi_t,t\right)=0$. Next, we examine the difference of the expected reward obtained by pulling arm k at state x_t at time t in MDP $\mathcal{M}$ and $\widehat{\mathcal{M}}$; when $n_{k,t}\neq 0$, this is given by

$$\begin{aligned}
|\widehat{r}(x_t,k)]-r(x_t,k)]| &= |\gamma_k^{n_{k,t}}x_{k,t}+\lambda_k\gamma_k^{n_{k,t}}-\widehat{\gamma}_k^{n_{k,t}}x_{k,t}-\widehat{\lambda}_k\widehat{\gamma}_k^{n_{k,t}}|\\
&\leq |x_{k,t}|\cdot|\gamma^{n_{k,t}}-\widehat{\gamma}^{n_{k,t}}|+|\lambda_k\gamma_k^{n_{k,t}}-\lambda_k\widehat{\gamma}_k^{n_{k,t}}+\lambda_k\widehat{\gamma}_k^{n_{k,t}}-\widehat{\lambda}_k\widehat{\gamma}_k^{n_{k,t}}|\\
&\leq \left(B(\delta/2)+\lambda_k\right)|\widehat{\gamma}_k^{n_{k,t}}-\gamma_k^{n_{k,t}}|+|\widehat{\lambda}_k-\lambda_k|=:\epsilon_R=O\left(\frac{1}{\sqrt{n}}\right),
\end{aligned}$$

where $\widehat{r}(x_t,k)$ is the expected reward of pulling arm k at state x_t in MDP $\widehat{\mathcal{M}}$. Putting it altogether, we have that for any deterministic policy π,

$$\begin{aligned}
V_{1,\mathcal{M}}^{\pi}(x_{\text{init}})-V_{1,\widehat{\mathcal{M}}}^{\pi}(x_{\text{init}}) &= r(x_{\text{init}},\pi_1(x_{\text{init}}))-\widehat{r}(x_{\text{init}},\pi_1(x_{\text{init}}))+\mathbb{E}_{x_2\sim p_{\mathcal{M}}(\cdot|x_1,\pi,1)}[V_{2,\mathcal{M}}^{\pi}(x_2)]\\
&\quad-\mathbb{E}_{x_2\sim p_{\widehat{\mathcal{M}}}(\cdot|x_1,\pi,1)}[V_{2,\widehat{\mathcal{M}}}^{\pi}(x_2)]\\
&\leq \epsilon_R+\mathbb{E}_{x_2\sim p_{\mathcal{M}}(\cdot|x_1,\pi,1)}[V_{2,\mathcal{M}}^{\pi}(x_2)]-\mathbb{E}_{x_2\sim p_{\widehat{\mathcal{M}}}(\cdot|x_1,\pi,1)}[V_{2,\mathcal{M}}^{\pi}(x_2)]\\
&\quad+\mathbb{E}_{x_2\sim p_{\widehat{\mathcal{M}}}(\cdot|x_1,\pi,1)}[V_{2,\mathcal{M}}^{\pi}(x_2)]-\mathbb{E}_{x_2\sim p_{\widehat{\mathcal{M}}}(\cdot|x_1,\pi,1)}[V_{2,\widehat{\mathcal{M}}}^{\pi}(x_2)]\\
&\leq T\epsilon_R+\sum_{t=1}^{T}\mathbb{E}_{\widehat{\mathcal{M}},\pi}\Big\{\mathbb{E}_{x_{t+1}\sim p_{\mathcal{M}}(\cdot|x_t,\pi,t)}[V_{t+1,\mathcal{M}}^{\pi}(x_{t+1})]\\
&\quad-\mathbb{E}_{x_{t+1}\sim p_{\widehat{\mathcal{M}}}(\cdot|x_t,\pi,t)}[V_{t+1,\mathcal{M}}^{\pi}(x_{t+1})]\Big\}\\
&\leq T\epsilon_R+T^2\epsilon_P\max_k b_k,
\end{aligned}$$

where $p_{\mathcal{M}}(\cdot|x_t,\pi,t)$ denotes $p(\cdot|x_t,\pi_t(x_t),t)$ in MDP $\mathcal{M}$ and the last inequality uses the fact that $\langle p_{\mathcal{M}}(\cdot|x_t,\pi,t)-p_{\widehat{\mathcal{M}}}(\cdot|x_t,\pi,t),\ V_{t+1,\mathcal{M}}^{\pi}\rangle\leq\|p_{\mathcal{M}}(\cdot|x_t,\pi,t)-p_{\widehat{\mathcal{M}}}(\cdot|x_t,\pi,t)\|_1\|V_{t+1,\mathcal{M}}^{\pi}\|_\infty\leq\epsilon_P T\max_k b_k$. Finally, we have that

$$\begin{aligned}
V_{1,\mathcal{M}}^{*}(x_{\text{init}})-V_{1,\mathcal{M}}^{\pi^*_{\widehat{\mathcal{M}}}}(x_{\text{init}}) &= V_{1,\mathcal{M}}^{\pi^*_{\mathcal{M}}}(x_{\text{init}})-V_{1,\widehat{\mathcal{M}}}^{\pi^*_{\mathcal{M}}}(x_{\text{init}})+V_{1,\widehat{\mathcal{M}}}^{\pi^*_{\mathcal{M}}}(x_{\text{init}})-V_{1,\widehat{\mathcal{M}}}^{\pi^*_{\widehat{\mathcal{M}}}}(x_{\text{init}})\\
&\quad+V_{1,\widehat{\mathcal{M}}}^{\pi^*_{\widehat{\mathcal{M}}}}(x_{\text{init}})-V_{1,\mathcal{M}}^{\pi^*_{\widehat{\mathcal{M}}}}(x_{\text{init}})\leq 2T\epsilon_R+2T^2\epsilon_P\max_k b_k,
\end{aligned}$$

where the equation follows from the fact that $V_{1,\mathcal{M}}^{*}(x_{\text{init}})=V_{1,\mathcal{M}}^{\pi^*_{\mathcal{M}}}(x_{\text{init}})$ and rearranging the terms, and the inequality follows from applying the bound of $V_{1,\mathcal{M}}^{\pi}(x_{\text{init}})-V_{1,\widehat{\mathcal{M}}}^{\pi}(x_{\text{init}})\leq T\epsilon_R+T^2\epsilon_P\max_k b_k$ that was derived above for $\pi=\pi^*_{\mathcal{M}}$ and $\pi=\pi^*_{\widehat{\mathcal{M}}}$ and using the fact that the policy $\pi^*_{\widehat{\mathcal{M}}}$ is optimal for MDP $\widehat{\mathcal{M}}$. Let E_3 denote the event that $V_{1,\mathcal{M}}^{*}(x_{\text{init}})-V_{1,\mathcal{M}}^{\pi^*_{\widehat{\mathcal{M}}}}(x_{\text{init}})\leq O(T^2/\sqrt{n})$. Putting it altogether, we have that $\mathbb{P}(E_3)\geq\mathbb{P}(E_2,E_1)=1-\mathbb{P}(E_2^c\cup E_1^c)\geq 1-\delta$. □

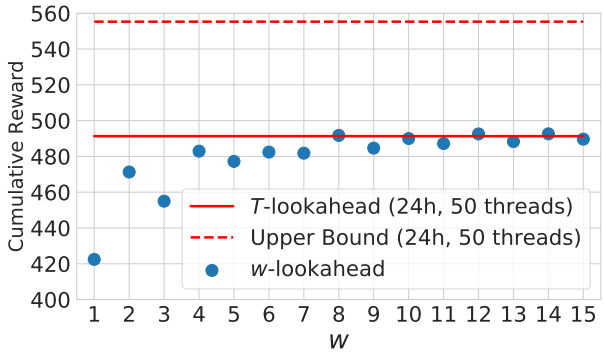
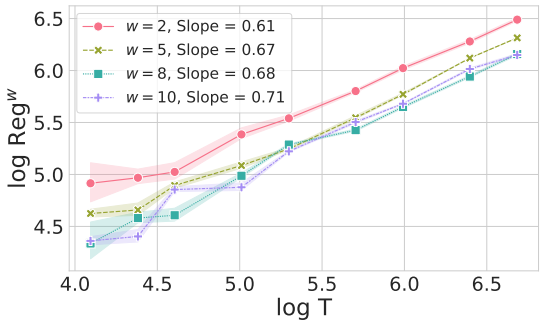
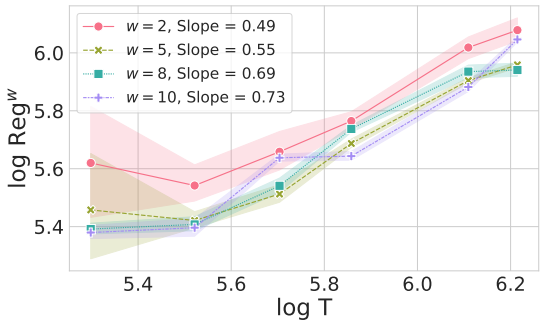


Figure C.1: The expected cumulative reward collected by w-lookahead policies (blue dots) when $T = 100$. When solving for the T-lookahead policy (solving (4.4) with $T = 100$), after 24 hours, Gurobi 9.1 obtains an objective value of 491.3 (red solid line) with an upper bound 555.3 (red dotted line) and an absolute optimality gap 64.0 (13.0%). The true expected cumulative reward for T-lookahead policy for this problem lies in between the solid and dotted red lines.



(a) $\log \mathrm{Reg}^w$ v.s. $\log T$



(b) $\log \mathrm{Reg}^w$ v.s. $\log T$

Figure C.2: Figure C.2a shows the log-log plot of the w-step lookahead regret of w-lookahead EEP (averaged over 20 random runs) under different T when there are 5 arms. Figure C.2b shows the log-log plot of the w-step lookahead regret of w-lookahead EEP (averaged over 20 random runs) under different T when there are 10 arms.

C.7 Additional Experimental Details and Results

We present additional experimental details and results.

w-lookahead Performance When evaluating the performance of w-lookahead policies, in addition to the case where $T = 30$ (Figure 4.3a), we have also run the experiments with $T = 100$ (Figure C.1). When solving for the 100-lookahead policy, we have increased the number of threads to 50 to solve for (4.4) and stopped the program at a time limit of 24 hours. In such settings, we obtain an upper bound on the absolute optimality gap of 64.0 (percentage optimality gap of 13.0%). When solved for w-lookahead policies with w in between 1 and 15 using 10 threads, Gurobi ends up solving (4.5) within 40s for all different w values. Thus, despite using significantly lower computational time, w-lookahead policies achieve a similar expected

cumulative reward to the T-lookahead policies (see Figures 4.3a and C.1).

EEP Performance Figure 4.3b is the log-log plot of the w-step lookahead regret of w-lookahead EEP against the horizon T when $T = 60, 80, 100, 150, 200, 300, 400$ (averaged over 20 random runs) and Figure C.2a is the log-log plot when $T = 60, 80, 100, 150, 200, 300, 400, 600, 800$ (averaged over 20 random runs), under the experimental setup provided in Section 4.7.

Finally, we present the result when we include 5 additional arms to the existing problem. The 5 new arms have parameters $\gamma_6 = .4, \gamma_7 = .5, \gamma_8 = .6, \gamma_9 = .8, \gamma_{10} = .7$, $\lambda_6 = 2, \lambda_7 = 3, \lambda_8 = 2, \lambda_9 = 3, \lambda_{10} = 1$, and $b_6 = 10, b_7 = 5, b_8 = 6, b_9 = 7, b_{10} = 8$. Figure C.2b is the log-log plot of the w-step lookahead regret of w-lookahead EEP against the horizon T when $T = 200, 250, 300, 350, 450, 500$ (averaged over 20 random runs).

Appendix D

Departing Bandits: Additional Details

D.1 Extension: Planning Beyond Two User Types

In this section, we treat the planning task with two categories ($K = 2$) but potentially many types (i.e., $M \geq 2$). For convenience, we formalize the results in this section in terms of $M = 2$, but the results are readily extendable for the more general $2 \times M$ case. We derive an almost-optimal planning policy via dynamic programming, and then explain why it cannot be used for learning as we did in the previous section.

For reasons that will become apparent later on, we define by V_H^π as the return of a policy π until the H's iteration. Using Theorem 5.3, we have that

$$\mathbb{E}[V_H^\pi(b)] = \sum_{i=1}^{H} b \cdot \mathbf{P}_{1,x}^{m_{1,i}} \cdot \mathbf{P}_{2,x}^{m_{2,i}} + (1-b)\mathbf{P}_{1,y}^{m_{1,i}} \cdot \mathbf{P}_{2,y}^{m_{2,i}},$$

where $m_{1,i} := |\{a_j = 1, j \leq i\}|$ and $m_{2,i} := |\{a_j = 2, j \leq i\}|$ are calculated based on the belief-category walk $b_1, a_1, b_2, a_2, \ldots$ induced by π. Further, let $\tilde{\pi}^*$ denote the policy maximizing V_H.

Notice that there is a bijection from $H-$iterations policies to $(m_{1,i}, m_{2,i})_{i=1}^H$; hence, we can find $\tilde{\pi}^*$ by finding the arg max of the expression on the right-hand-side of the above equation, in terms of $(m_{1,i}, m_{2,i})_{i=1}^H$. Formally, we want to solve the integer linear programming (ILP),

$$\begin{aligned} &\text{maximize } \sum_{i=1}^{H} b \cdot \mathbf{P}_{1,x}^{m_{1,i}} \cdot \mathbf{P}_{2,x}^{m_{2,i}} + (1-b)\mathbf{P}_{1,y}^{m_{1,i}} \cdot \mathbf{P}_{2,y}^{m_{2,i}} \\ &\text{subject to } m_{a,i} = \sum_{l=1}^{i} z_{a,l} \text{ for } a \in \{1,2\}, i \in [H], \\ &\qquad z_{a,i} \in \{0,1\} \text{ for } a \in \{1,2\}, i \in [H], \\ &\qquad z_{1,i} + z_{2,i} = 1 \text{ for } i \in [H]. \end{aligned} \tag{D.1}$$

Despite that this problem involves integer programming, we can solve it using dynamic programming in $O(H^2)$ runtime. Notice that the optimization is over a subset of binary variables

$(z_{1,i}, z_{2,i})_{i=1}^{H}$. Let Z^H be the set of feasible solutions of the ILP, and similarly let Z^h denote set of prefixes of length $h \leq H$ of Z^H.

For any $h \in [H]$ and $\mathbf{z} \in Z^h$, define

$$D^h(\mathbf{z}) \stackrel{\text{def}}{=} \sum_{i=1}^{h} b \cdot \mathbf{P}_{1,x}^{m_{1,i}} \cdot \mathbf{P}_{2,x}^{m_{2,i}} + (1-b)\mathbf{P}_{1,y}^{m_{1,i}} \cdot \mathbf{P}_{2,y}^{m_{2,i}},$$

where $m_{a,i} = \sum_{l=1}^{i} z_{a,l}$ for $j \in \{1,2\}, i \in [h]$ as in the ILP.

Consequently, solving the ILP is equivalent to maximizing D^H over the domain Z^H.

Next, for any $h \in [H]$ and two integers c_1, c_2 such that $c_1 + c_2 = h$, define

$$\tilde{D}^h(c_1, c_2) \stackrel{\text{def}}{=} \max_{\substack{\mathbf{z} \in Z^h, \\ m_{1,h}(\mathbf{z}) = c_1 \\ m_{2,h}(\mathbf{z}) = c_2}} D^h(\mathbf{z}). \tag{D.2}$$

Under this construction, $\max_{c_1,c_2} \tilde{D}^H(c_1, c_2)$ over c_1, c_2 such that $c_1 + c_2 = H$ is precisely the value of the ILP.

Reformulating Equation (D.2) for $h > 1$,

$$\tilde{D}^h(c_1,c_2) = \max_{\substack{z_1,z_2 \in \{0,1\} \\ z_1+z_2=1}} \left\{ \tilde{D}^{h-1}(c_1 - z_1, c_2 - z_2) + \alpha(c_1,c_2) \right\},$$

where $\alpha(m_1, m_2) \stackrel{\text{def}}{=} b \cdot x_1^{m_1} \cdot x_2^{m_2} + (1-b) y_1^{m_1} \cdot y_2^{m_2}$. For every h, there are only $h+1$ possible values $\tilde{D}^h$ can take: All the ways of dividing h into non-negative integers c_1 and c_2; therefore, having computed $\tilde{D}^{h-1}$ for all h feasible inputs, we can compute $\tilde{D}^h(c_1, c_2)$ in $O(h)$. Consequently, computing $\max_{c_1,c_2} \tilde{D}^H(c_1, c_2)$, which is precisely the value of the ILP in (D.1), takes $O(H^2)$ run-time. Moreover, the policy $\tilde{\pi}^*$ can be found using backtracking. We remark that an argument similar to Lemma 5.4 implies that $\mathbb{E}[V^{\pi^*} - V^{\tilde{\pi}^*}] \leq \frac{1}{2^{O(H)}}$; hence, $\tilde{\pi}^*$ is almost optimal.

To finalize this section, we remark that this approach could also work for $K > 2$ categories. Naively, for a finite horizon H, there are K^H possible policies. The dynamic programming procedure explain above makes the search operate in run-time of $O(H^K)$. The run-time, exponential in the number of categories but polynomial in the horizon, is feasible when the number of categories is small.

D.2 Experimental Evaluation

For general real-world datasets, we propose a scheme to construct semi-synthetic problem instances with many arms and many user types, using rating data sets with multiple ratings per user. We exemplify our scheme on the MovieLens Dataset Harper and Konstan (2015). As a pre-processing step, we set movie genres to be the categories of interest, select a subset of categories $|A|$ of size k (e.g., sci-fi, drama, and comedy), and select the number of user types, m.

Remove any user who has not provided a rating for at least one movie from each category $a \in A$. When running the algorithm, randomly draw users from the data, and given a recommended category a, suggest them a random movie which they have rated, and set their click probability to $1 - r$, where $r \in [0, 1]$ is their normalized rating of the suggested movie.

D.3 UCB Policy for Sub-exponential Returns

An important tool for analyzing sub-exponential random variables is Bernstein's Inequality, which is a concentration inequality for sub-exponential random variables (see, e.g., Jia et al. (2021)). Being a major component of the regret analysis for Algorithm 5.2, we state it here for convenience.

Lemma D.1

(Bernstein's Inequality) Let a random variable X be sub-exponential with parameters (τ^2, b). Then for every $v \geq 0$:

$$\mathbb{P}[|X - \mathbb{E}[X]| \geq v] \leq \begin{cases} 2\exp(-\frac{v^2}{2\tau^2}) & v \leq \frac{\tau^2}{b} \\ 2\exp(-\frac{v}{2b}) & else \end{cases}.$$

D.4 Single User Type: Proofs from Section 5.4

To simplify the proofs, we use the following notation: For a fixed-arm policy π^a, we use $V_j^{\pi^a}$ to denote its return from iteration j until the user departs. Namely,

$$V_j^{\pi^a} = \sum_{i=j}^{N^{\pi^a}} \mathbf{P}_a$$

Throughout this section, we will use the following Observation.

Observation D.1. *For every policy π and iteration j,*

$$\mathbb{E}[V_j^\pi] = \mathbf{P}_{\pi_j}(1 + \mathbb{E}[V_{j+1}^\pi]) + (1 - \mathcal{L}_{\pi_j})(1 - \mathbf{P}_{\pi_j})\mathbb{E}[V_{j+1}^\pi] = \mathbb{E}[V_{j+1}^\pi](1 - \mathcal{L}_{\pi_j}(1 - \mathbf{P}_{\pi_j})) + \mathbf{P}_{\pi_j}.$$

Lemma 5.1

A policy π^{a^*} is optimal if

$$a^* \in \arg\max_{a \in [K]} \frac{\mathbf{P}_a}{\mathcal{L}_a(1 - \mathbf{P}_a)}.$$

Proof. First, recall that every POMDP has an optimal Markovian policy which is deterministic (we refer the reader to Section 5.5.1 for full formulation of the problem as POMDP). Having

independent rewards and a single state implies that there exists $\mu^* \in \mathbb{N}$ such that $\mathbb{E}[V_j^*] = \mu^*$ for every $j \in \mathbb{N}$ (similarly to standard MAB problems, there exists a fixed-arm policy which is optimal).
Assume by contradiction that the optimal policy π^{a^*} holds

$$a^* \notin \arg\max_{a \in [k]} \frac{\mathbf{P}_a}{\mathcal{L}_a(1-\mathbf{P}_a)}.$$

Now, notice that

$$\mathbb{E}[V^{\pi^{a'}}] = \mathbb{E}[V_1^{\pi^{a'}}] = \mathbb{E}[V_2^{\pi^{a'}}](1 - \mathcal{L}_{a'}(1-\mathbf{P}_{a'})) + \mathbf{P}_{a'}$$

Solving the recurrence relation and summing the geometric series we get

$$\mathbb{E}[V^{\pi^{a'}}] = \mathbf{P}_{a'} \sum_{j=0}^{\infty} (1 - \mathcal{L}_{a'}(1-\mathbf{P}_{a'}))^j = \frac{\mathbf{P}_{a'}}{\mathcal{L}_{a'}(1-\mathbf{P}_{a'})}.$$

Finally,

$$a^* \notin \arg\max_{a \in [k]} \frac{\mathbf{P}_a}{\mathcal{L}_a(1-\mathbf{P}_a)},$$

yields that any fixed-armed policy, $\pi^{a'}$ such that

$$a' \in \arg\max_{a \in [k]} \frac{\mathbf{P}_a}{\mathcal{L}_a(1-\mathbf{P}_a)}$$

holds $\mathbb{E}[V^{\pi^{a'}}] > \mathbb{E}[V^{\pi^{a^*}}]$, a contradiction to the optimality of π^{a^*} . □

Lemma D.2

For each $a \in [k]$, the centered random return $V^{\pi^a} - \mathbb{E}[V^{\pi^a}]$ is sub-exponential with parameter $C_2 = -4/\ln(1 - \mathcal{L}_a(1-\mathbf{P}_a))$.

In order to show that returns of fixed-arm policies are sub-exponential random variables, we first show that the number of iterations of users recommended by fixed-arm policies is also a sub-exponential. For this purpose, we state here a lemma that implies that every geometric r.v. is a sub-exponential r.v.. The proof of the next lemma appears, e.g., in Hillar and Wibisono (2018) (Lemma 4.3).

Lemma D.3

Let X be a geometric random variable with parameter $r \in (0,1)$, so that:

$$\mathbb{P}[X = x] = (1-r)^{x-1} r, \quad x \in \mathbb{N}.$$

Then X satisfies Property (2) from Definition 5.1. Namely, X is sub-exponential with

parameter $C_2 = -2/\ln(1-r)$. Formally,

$$\forall p \geq 0 : (\mathbb{E}[|X|^p])^{1/p} \leq -\frac{2}{\ln(1-r)}p.$$

The lemma above and Observation 5.1 allow us to deduce that the variables N_a are sub-exponential in the first part of the following Corollary (the case in which $\mathcal{L}_a = 0$ follows immediately from definition.). The second part of the lemma follows directly from the equivalences between Properties (2) and (1) in Definition 5.1.

Corollary D.1. *For each $a \in [K]$, the number of iterations a user recommended by π^a stays within the system, N_a, is sub-exponential with parameter $C_2^a = -2/\ln(1-\mathcal{L}_a(1-\mathbf{P}_a))$. In addition, there exist constants $C_1^a > 0$ for every $a \in [K]$ such that*

$$\forall a \in [K],\ v \geq 0 : \mathbb{P}[|N_a| > v] \leq \exp(1 - \frac{v}{C_1^a}).$$

The next Proposition D.1 is used for the proof of Lemma D.2.

Proposition D.1

For every $a \in [K]$,

$$|\mathbb{E}[V^{\pi^a}]| \leq \frac{-2}{\ln(1-\mathcal{L}_a(1-\mathbf{P}_a))}$$

Proof. First, notice that

$$\begin{aligned}(1-\mathcal{L}_a(1-\mathbf{P}_a))\ln(1-\mathcal{L}_a(1-\mathbf{P}_a)) &> (1-\mathcal{L}_a(1-\mathbf{P}_a))\frac{-\mathcal{L}_a(1-\mathbf{P}_a)}{1-\mathcal{L}_a(1-\mathbf{P}_a)} = -\mathcal{L}_a(1-\mathbf{P}_a) \\ &> -2\mathcal{L}_a(1-\mathbf{P}_a),\end{aligned}$$

where the first inequality is due to $\frac{x}{1+x} \leq \ln(1+x)$ for every $x \geq -1$. Rearranging,

$$\frac{1-\mathcal{L}_a(1-\mathbf{P}_a)}{\mathcal{L}_a(1-\mathbf{P}_a)} < \frac{-2}{\ln(1-\mathcal{L}_a(1-\mathbf{P}_a))}. \tag{D.3}$$

For each user, the realization of V^{π^a} is less or equal to the realization of $N_a - 1$ for the same user (as users provide negative feedback in their last iteration); hence,

$$|\mathbb{E}[V^{\pi^a}]| = \mathbb{E}[V^{\pi^a}] \leq \mathbb{E}[N_a] - 1 = \frac{1}{\mathcal{L}_a(1-\mathbf{P}_a)} - 1 = \frac{1-\mathcal{L}_a(1-\mathbf{P}_a)}{\mathcal{L}_a(1-\mathbf{P}_a)} < \frac{-2}{\ln(1-\mathcal{L}_a(1-\mathbf{P}_a))}.$$

□

We proceed by showing that returns of fix-armed policies satisfy Property (1) from Definition 5.1.

Lemma D.2

For each $a \in [k]$, the centered random return $V^{\pi^a} - \mathbb{E}[V^{\pi^a}]$ is sub-exponential with parameter $C_2 = -4/\ln(1 - \mathcal{L}_a(1 - \mathbf{P}_a))$.

Proof. We use Property (1) from Definition 5.1 to derive that V^{π^a} is also sub-exponential. This is true since the tails of V^{π^a} satisfy that for all $v \geq 0$,

$$\mathbb{P}[|V^{\pi^a}| > v] \leq \mathbb{P}[|N_a| > v + 1] \leq \mathbb{P}[|N_a| > v] \leq_{(1)} \exp(1 - \frac{v}{C_1}),$$

where the first inequality follows since $|N_a| > v + 1$ is a necessary condition for $|V^{\pi^a}| > v$, and the last inequality follows from Corollary D.1. Along with Definition 5.1, we conclude that

$$\mathbb{E}[|V^{\pi^a}|^p]^{1/p} \leq -2/\ln(1 - \mathcal{L}_a(1 - \mathbf{P}_a))p. \tag{D.4}$$

Now, applying Minkowski's inequality and then Jensen's inequality (as $f(z) = z^p, g(z) = |z|$ are convex for every $p \geq 1$) we get

$$(\mathbb{E}[|V^{\pi^a} - \mathbb{E}[V^{\pi^a}]|^p])^{1/p} \leq \mathbb{E}[|V^{\pi^a}|^p]^{1/p} + \mathbb{E}[\mathbb{E}[|V^{\pi^a}|]^p]^{1/p} \leq \mathbb{E}[|V^{\pi^a}|^p]^{1/p} + |\mathbb{E}[V^{\pi^a}]|.$$

Using Proposition D.1 and Inequality (D.4), we get

$$\mathbb{E}[|V^{\pi^a}|^p]^{1/p} + |\mathbb{E}[V^{\pi^a}]| \leq \frac{-2}{\ln(1 - \mathcal{L}_a(1 - \mathbf{P}_a))} + \frac{1}{\mathcal{L}_a(1 - \mathbf{P}_a)} - 1 \leq \frac{-4}{\ln(1 - \mathcal{L}_a(1 - \mathbf{P}_a))}$$

Hence $V^{\pi^a} - \mathbb{E}[V^{\pi^a}]$ is sub-exponential with parameter $C_2 = -4/\ln(1 - \mathcal{L}_a(1 - \mathbf{P}_a))$. □

Lemma 5.2

For each category $a \in [K]$, the centred random variable $V^{\pi^a} - \mathbb{E}[V^{\pi^a}]$ is sub-exponential with parameters (τ_a^2, b_a), such that

$$\tau_a = b_a = -\frac{8e}{\ln(1 - \mathcal{L}_a(1 - \mathbf{P}_a))}.$$

Proof. Throughout this proof, we will use the sub-exponential norm, $||\cdot||_{\psi_1}$, which is defined as

$$||Z||_{\psi_1} = \sup_{p \geq 1} \frac{(\mathbb{E}[|Z|^p])^{1/p}}{p}.$$

Let

$$X = \frac{V^{\pi^a} - \mathbb{E}[V^{\pi^a}]}{||V^{\pi^a} - \mathbb{E}[V^{\pi^a}]||_{\psi_1}}, \quad y = \gamma \cdot ||V^{\pi^a} - \mathbb{E}[V^{\pi^a}]||_{\psi_1}.$$

We have that

$$||X||_{\psi_1} = ||\frac{V^{\pi^a} - \mathbb{E}[V^{\pi^a}]}{||V^{\pi^a} - \mathbb{E}[V^{\pi^a}]||_{\psi_1}}||_{\psi_1} = 1. \tag{D.5}$$

Let γ be such that $|\gamma| < 1/b_a = -\frac{\ln(1-\mathcal{L}_a(1-\mathbf{P}_a))}{8e}$. From Lemma D.2 we conclude that

$$|\gamma| = |\frac{y}{||V^{\pi^a} - \mathbb{E}[V^{\pi^a}]||_{\psi_1}}| \leq -\frac{\ln(1-\mathcal{L}_a(1-\mathbf{P}_a))}{8e} = \frac{1}{2e} \cdot \frac{1}{||V^{\pi^a} - \mathbb{E}[V^{\pi^a}]||_{\psi_1}};$$

hence, $|y| < \frac{1}{2e}$.
Summing the geometric series, we get

$$\sum_{p=2}^{\infty}(e|y|)^p = \frac{e^2y^2}{1-e|y|} < 2e^2y^2 \tag{D.6}$$

In addition, notice that $yX = \gamma(V^{\pi^a} - \mathbb{E}[V^{\pi^a}])$.
Next, from the Taylor series of $\exp(\cdot)$ we have

$$\mathbb{E}[\exp(\gamma(V^{\pi^a} - \mathbb{E}[V^{\pi^a}]))] = \mathbb{E}[\exp(yX)] = 1 + y\mathbb{E}[x] + \sum_{p=2}^{\infty}\frac{y^p\mathbb{E}[X^p]}{p!}.$$

Combining the fact that $\mathbb{E}[X] = 0$ and (D.5) to the above,

$$1 + y\mathbb{E}[x] + \sum_{p=2}^{\infty}\frac{y^p\mathbb{E}[X^p]}{p!} \leq 1 + \sum_{p=2}^{\infty}\frac{y^pp^p}{p!}.$$

By applying $p! \geq (\frac{p}{e})^p$ and (D.6), we get

$$1 + \sum_{p=2}^{\infty}\frac{y^pp^p}{p!} \leq 1 + \sum_{p=2}^{\infty}(e|y|)^p \leq 1 + 2e^2y^2 \leq \exp(2e^2y^2) = \exp(2e^2(\gamma \cdot ||V^{\pi^a} - \mathbb{E}[V^{\pi^a}]||_{\psi_1})^2),$$

where the last inequality is due to $1 + x \leq e^x$.

Note that $||V^{\pi^a} - \mathbb{E}[V^{\pi^a}]||_{\psi_1} \leq -\frac{4}{\ln(1-\mathcal{L}_a(1-\mathbf{P}_a))})^2)$. Ultimately,

$$\begin{aligned}\mathbb{E}[\exp(\gamma(V^{\pi^a} - \mathbb{E}[V^{\pi^a}]))] &\leq \exp\left(2e^2\gamma^2(-\frac{4}{\ln(1-\mathcal{L}_a(1-\mathbf{P}_a))})^2\right) \\ &= \exp\left(\frac{1}{2}\gamma^2(-\frac{8e}{\ln(1-\mathcal{L}_a(1-\mathbf{P}_a))})^2\right).\end{aligned}$$

This concludes the proof of the lemma. □

D.5 Two User Types and Two Categories: Proofs from Section 5.5

D.5.1 Planning when K = 2

Theorem 5.3

For every policy π and an initial belief $b \in [0,1]$, the expected return is given by

$$\mathbb{E}[V^{\pi}(b)] = \sum_{i=1}^{\infty} b \cdot \mathbf{P}_{1,x}^{m_{1,i}} \cdot \mathbf{P}_{2,x}^{m_{2,i}} + (1-b)\mathbf{P}_{1,y}^{m_{1,i}} \cdot \mathbf{P}_{2,y}^{m_{2,i}},$$

where $m_{1,i} := |\{a_j = 1, j \leq i\}|$ and $m_{2,i} := |\{a_j = 2, j \leq i\}|$ are calculated based on the belief-category walk $b_1, a_1, b_2, a_2, \ldots$ induced by π.

Proof. Let $\beta_i^{\pi}(b) := b \cdot \mathbf{P}_{1,x}{}^{m_{1,i}} \cdot \mathbf{P}_{2,x}{}^{m_{2,i}} + (1-b)\mathbf{P}_{1,y}{}^{m_{1,i}} \cdot \mathbf{P}_{2,y}{}^{m_{2,i}}$. We will prove that for every policy π and every belief b, we have that $\mathbb{E}[V_H^{\pi_a}(b)] = \sum_{i=1}^{H} \beta_i^{\pi}(b)$ by a backward induction over H.

For the base case, consider $H = 1$. We have that

$$\mathbb{E}[V_1^{\pi}(b_1)] = b_1 \cdot \mathbf{P}_{a_1,x} + (1-b)\mathbf{P}_{a_1,y} = b \cdot \mathbf{P}_{1,x}{}^{m_{1,1}} \cdot \mathbf{P}_{2,x}{}^{m_{2,1}} + (1-b)\mathbf{P}_{1,y}{}^{m_{1,1}} \cdot \mathbf{P}_{2,y}{}^{m_{2,1}} = \beta_1^{\pi}(b)$$

as $m_{a,1} = \mathbb{I}[a_1 = a]$.

For the inductive step, assume that $\mathbb{E}[V_{H-1}^{\pi}(b)] = \sum_{i=1}^{H-1} \beta_i^{\pi}(b)$ for every $b \in [0,1]$. We need to show that $\mathbb{E}[V_H^{\pi}(b)] = \sum_{i=1}^{H} \beta_i^{\pi}(b)$ for every $b \in [0,1]$.

Indeed,

$$\begin{aligned}
\mathbb{E}[V_H^{\pi}(b_1)] &= \beta_1^{\pi}(b_1)(1 + \mathbb{E}[V_{H-1}^{\pi}(b'(b_1, a_1, liked))]) \\
&= \beta_1^{\pi}(b_1)(1 + \mathbb{E}[V_{H-1}^{\pi}(b_2)]) \\
&= \beta_1^{\pi}(b_1)(1 + \sum_{i=2}^{H-1} \beta_i^{\pi}(b_2)) \\
&= \sum_{i=1}^{H} \beta_i^{\pi}(b_1),
\end{aligned}$$

where the second to last equality is due to the induction hypothesis and the assumption that π is a deterministic stationary policy. The proof completes by realizing that $\mathbb{E}[V^{\pi}(b)] = \lim_{H\to\infty} \mathbb{E}[V_H^{\pi}(b)] = \lim_{H\to\infty} \sum_{i=1}^{H} \beta_i^{\pi}(b) = \sum_{i=1}^{\infty} \beta_i^{\pi}(b)$, since the sum is finite and has positive summands. □

D.5.2 Dominant Row (DR)

Lemma 5.3

For any instance such that $\mathbf{P}$ has a dominant row a, the fixed policy π^a is an optimal policy.

Proof. We will show that for every iteration j, no matter what were the previous topic recommendations were, selecting topic 1 rather than topic 2 can only increase the value.

Let π be a stationary policy such that $\pi(b_j) = 2$. Changing it into a policy π^j that is equivalent to π for all iterations but iteration $j+1$ in which it recommends topic 1 can only improve the value.

Since $\mathbf{P}_{1,x}, \mathbf{P}_{2,x}, \mathbf{P}_{1,y}, \mathbf{P}_{2,y} \geq 0$, $\mathbf{P}_{1,x} - \mathbf{P}_{2,x} \geq 0$, $b, 1-b \geq 0$ and this structure satisfies $\mathbf{P}_{2,y} - \mathbf{P}_{1,y} \leq 0$, we get that for every $\bar{m}_{1,j}, \bar{m}_{2,j}, n_{1,j}, n_{2,j} \in \mathbb{N}$ and for every b,

$$b \cdot \mathbf{P}_{1,x}^{\bar{m}_{1,j}+n_{1,j}} \cdot \mathbf{P}_{2,x}^{\bar{m}_{2,j}+n_{2,j}}(\mathbf{P}_{1,x} - \mathbf{P}_{2,x}) \geq (1-b)\mathbf{P}_{1,y}^{\bar{m}_{1,j}+n_{1,j}} \cdot \mathbf{P}_{2,y}^{\bar{m}_{2,j}+n_{2,j}}(\mathbf{P}_{2,y} - \mathbf{P}_{1,y});$$

thus,

$$b \cdot \mathbf{P}_{1,x}^{\bar{m}_{1,j}+1+n_{1,j}} \cdot \mathbf{P}_{2,x}^{\bar{m}_{2,j}+n_{2,j}} + (1-b)\mathbf{P}_{1,y}^{\bar{m}_{1,j}+1+n_{1,j}} \cdot \mathbf{P}_{2,y}^{m_{2,j}+n_{2,j}} \geq$$
$$b \cdot \mathbf{P}_{1,x}^{\bar{m}_{1,j}+n_{1,j}} \cdot \mathbf{P}_{2,x}^{\bar{m}_{2,j}+1+n_{2,j}} + (1-b)\mathbf{P}_{1,y}^{\bar{m}_{1,j}+n_{1,j}} \cdot \mathbf{P}_{2,y}^{\bar{m}_{2,j}+1+n_{2,j}}.$$

Hence for every time step $j+1$, choosing topic 1 instead of topic 2 leads to increased value of each of the summation element $b \cdot \mathbf{P}_{1,x}^{m_{1,i}} \cdot \mathbf{P}_{2,x}^{m_{2,i}} + (1-b)\mathbf{P}_{1,y}^{m_{1,i}} \cdot \mathbf{P}_{2,y}^{m_{2,i}}$ such that $m_{1,i} = \bar{m}_{1,j} + n_{1,j} \geq \bar{m}_{1,j}$ and $m_{2,i} = \bar{m}_{2,j} + n_{2,j} \geq \bar{m}_{2,j}$. We deduce that

$$\mathbb{E}[V^{\pi^j}(b)] \geq \mathbb{E}[V^{\pi}(b)].$$

□

D.5.3 Dominant Column (DC)

Before proving the main theorem (Theorem 5.4), we prove two auxiliary lemmas.

Lemma D.4

For $\mathbf{P}$ with a DC structure, if a policy π is optimal then it recommends topic 1 for all iteration $j' \geq j+1$ such that

$$\sum_{i=j+1}^{\infty} \mathbf{P}_{1,x}^{m_{1,i}} \mathbf{P}_{2,x}^{m_{2,i}} > \sum_{i=j+1}^{\infty} \frac{1-b}{b} \cdot \frac{\mathbf{P}_{2,y} - \mathbf{P}_{1,y}}{\mathbf{P}_{1,x} - \mathbf{P}_{2,x}} \cdot \frac{\mathbf{P}_{2,x}}{\mathbf{P}_{2,y}} \mathbf{P}_{1,y}^{m_{1,i}} \mathbf{P}_{2,y}^{m_{2,i}}. \tag{D.7}$$

Proof. First, assume by contradiction that there exists an optimal policy π that recommends topic 2 in iteration $j+1$ such that (D.7) holds.

Let π^j be the policy that is equivalent to π but recommend topic 1 instead of topic 2 in iteration $j+1$. Since π and π^j recommends the same topic until iteration j, along with the optimality of π, we have

$$\mathbb{E}[V^{\pi^j}(b)] - \mathbb{E}[V^{\pi}(b)] = \mathbb{E}[V_{j+1}^{\pi^j}(b)] - \mathbb{E}[V_{j+1}^{\pi}(b)] \leq 0.$$

Expending the above equation,

$$\sum_{i=j+1}^{\infty} b\cdot\mathbf{P}_{1,x}^{m_{1,i}^{\pi}+1}\cdot\mathbf{P}_{2,x}^{m_{2,i}^{\pi}-1}+(1-b)\mathbf{P}_{1,y}^{m_{1,i}^{\pi}+1}\cdot\mathbf{P}_{2,y}^{m_{2,i}^{\pi}-1}-\Big(\sum_{i=j+1}^{\infty} b\cdot\mathbf{P}_{1,x}^{m_{1,i}^{\pi}}\cdot\mathbf{P}_{2,x}^{m_{2,i}^{\pi}}+(1-b)\mathbf{P}_{1,y}^{m_{1,i}^{\pi}}\cdot\mathbf{P}_{2,y}^{m_{2,i}^{\pi}}\Big)\leq 0$$

$$\sum_{i=j+1}^{\infty} b\cdot\mathbf{P}_{1,x}^{m_{1,i}^{\pi}}\cdot\mathbf{P}_{2,x}^{m_{2,i}^{\pi}}(\frac{\mathbf{P}_{1,x}}{\mathbf{P}_{2,x}}-1)\leq\sum_{i=j+1}^{\infty}(1-b)\mathbf{P}_{1,y}^{m_{1,i}^{\pi}}\cdot\mathbf{P}_{2,y}^{m_{2,i}^{\pi}}(1-\frac{\mathbf{P}_{1,y}}{\mathbf{P}_{2,y}})$$

$$\frac{b(\mathbf{P}_{1,x}-\mathbf{P}_{2,x})}{\mathbf{P}_{2,x}}\sum_{i=j+1}^{\infty}\cdot\mathbf{P}_{1,x}^{m_{1,i}^{\pi}}\cdot\mathbf{P}_{2,x}^{m_{2,i}^{\pi}}\leq\frac{(1-b)(\mathbf{P}_{2,y}-\mathbf{P}_{1,y})}{\mathbf{P}_{2,y}}\sum_{i=j+1}^{\infty}\mathbf{P}_{1,y}^{m_{1,i}^{\pi}}\cdot\mathbf{P}_{2,y}^{m_{2,i}^{\pi}}$$

$$\sum_{i=j+1}^{\infty}\mathbf{P}_{1,x}^{m_{1,i}^{\pi}}\cdot\mathbf{P}_{2,x}^{m_{2,i}^{\pi}}\leq\frac{1-b}{b}\cdot\frac{\mathbf{P}_{2,x}}{\mathbf{P}_{2,y}}\cdot\frac{\mathbf{P}_{2,y}-\mathbf{P}_{1,y}}{\mathbf{P}_{1,x}-\mathbf{P}_{2,x}}\sum_{i=j+1}^{\infty}\mathbf{P}_{1,y}^{m_{1,i}^{\pi}}\cdot\mathbf{P}_{2,y}^{m_{2,i}^{\pi}},$$

which is a contradiction to (D.7).

For the second part of the lemma, assume that condition (D.7) holds for some iteration $j+1\in\mathbb{N}$ and some optimal policy π; hence, $\pi(b,m_{1,j}^{\pi},m_{2,j}^{\pi})=1$ and we have $m_{1,j+1}^{\pi}=m_{1,j}^{\pi}+1$ and $m_{2,j+1}^{\pi}=m_{2,j}^{\pi}$. Exploiting this fact, we have that

$$\sum_{i=j+2}^{\infty}\mathbf{P}_{1,x}^{m_{1,i}^{\pi}}\mathbf{P}_{2,x}^{m_{2,i}^{\pi}}=\sum_{i=j+1}^{\infty}\mathbf{P}_{1,x}^{m_{1,i}^{\pi}+1}\mathbf{P}_{2,x}^{m_{2,i}^{\pi}}=\mathbf{P}_{1,x}\sum_{i=j+1}^{\infty}\mathbf{P}_{1,x}^{m_{1,i}^{\pi}}\mathbf{P}_{2,x}^{m_{2,i}^{\pi}}>(D.7),$$

implying

$$\begin{aligned}
&\mathbf{P}_{1,x}\sum_{i=j+1}^{\infty}\frac{1-b}{b}\cdot\frac{\mathbf{P}_{2,y}-\mathbf{P}_{1,y}}{\mathbf{P}_{1,x}-\mathbf{P}_{2,x}}\cdot\frac{\mathbf{P}_{2,x}}{\mathbf{P}_{2,y}}\mathbf{P}_{1,y}^{m_{1,i}^{\pi}}\mathbf{P}_{2,y}^{m_{2,i}^{\pi}}\\
>&(\mathbf{P}_{1,x}\geq\mathbf{P}_{1,y})\mathbf{P}_{1,y}\sum_{i=j+1}^{\infty}\frac{1-b}{b}\cdot\frac{\mathbf{P}_{2,y}-\mathbf{P}_{1,y}}{\mathbf{P}_{1,x}-\mathbf{P}_{2,x}}\cdot\frac{\mathbf{P}_{2,x}}{\mathbf{P}_{2,y}}\mathbf{P}_{1,y}^{m_{1,i}^{\pi}}\mathbf{P}_{2,y}^{m_{2,i}^{\pi}}\\
=&\sum_{i=j+1}^{\infty}\frac{1-b}{b}\cdot\frac{\mathbf{P}_{2,y}-\mathbf{P}_{1,y}}{\mathbf{P}_{1,x}-\mathbf{P}_{2,x}}\cdot\frac{\mathbf{P}_{2,x}}{\mathbf{P}_{2,y}}\mathbf{P}_{1,y}^{m_{1,i}^{\pi}+1}\mathbf{P}_{2,y}^{m_{2,i}^{\pi}}\\
=&\sum_{i=j+2}^{\infty}\frac{1-b}{b}\cdot\frac{\mathbf{P}_{2,y}-\mathbf{P}_{1,y}}{\mathbf{P}_{1,x}-\mathbf{P}_{2,x}}\cdot\frac{\mathbf{P}_{2,x}}{\mathbf{P}_{2,y}}\mathbf{P}_{1,y}^{m_{1,i}^{\pi}}\mathbf{P}_{2,y}^{m_{2,i}^{\pi}}.
\end{aligned}$$

□

An immediate consequence of Lemma D.4 is the following corollary.

Corollary D.2. *For any DC-structured* $\mathbf{P}$ *and every belief* $b\in[0,1]$*, the optimal policy* π *first recommends topic* 2 *for at most*

$$argmin_N\sum_{i=1}^{N}\mathbf{P}_{2,x}^{m_{2,i}^{\pi}}>\frac{1-b}{b}\cdot\frac{\mathbf{P}_{2,y}-\mathbf{P}_{1,y}}{\mathbf{P}_{1,x}-\mathbf{P}_{2,x}}\cdot\frac{\mathbf{P}_{2,x}}{\mathbf{P}_{2,y}}\sum_{i=1}^{N}\mathbf{P}_{2,y}^{m_{2,i}^{\pi}}$$

times, and then recommends topic 1 *permanently. In addition,* $N\in\mathbb{N}$ *since* $\mathbf{P}_{2,x}>\mathbf{P}_{2,y}$.

Theorem 5.4

For any instance such that $\mathbf{P}$ has a dominant column, one of the following four policies is optimal:

$$\pi^1, \pi^2, \pi^{2:\lfloor N^* \rfloor}, \pi^{2:\lceil N^* \rceil},$$

where $N^* = N^*(\mathbf{P}, \mathbf{q})$ is a constant, and $\pi^{2:\lfloor N^* \rfloor}$ ($\pi^{2:\lceil N^* \rceil}$) stands for recommending Category 2 until iteration $\lfloor N^* \rfloor$ ($\lceil N^* \rceil$) and then switching to Category 1.

Proof. Due to Theorem 5.3 and Corollary D.2, we can write the expected value of a policy as a function of N when $\mathbf{P}$ has a DC structure:

$$\begin{aligned}
&\mathbb{E}[V^{\pi_N}(b)] = \sum_{i=1}^{\infty} b \cdot \mathbf{P}_{1,x}^{m_{1,i}} \cdot \mathbf{P}_{2,x}^{m_{2,i}} + (1-b)\mathbf{P}_{1,y}^{m_{1,i}} \cdot \mathbf{P}_{2,y}^{m_{2,i}} \\
=&\sum_{i=1}^{N} b \cdot \mathbf{P}_{2,x}^{i} + (1-b)\mathbf{P}_{2,y}^{i} + \sum_{i=N+1}^{\infty} b \cdot \mathbf{P}_{2,x}^{N} \cdot \mathbf{P}_{1,x}^{i-N} + (1-b)\mathbf{P}_{2,y}^{N} \cdot \mathbf{P}_{1,y}^{i-N} \\
=&b \cdot \frac{\mathbf{P}_{2,x}(\mathbf{P}_{2,x}^{N} - 1)}{\mathbf{P}_{2,x} - 1} + (1-b) \cdot \frac{\mathbf{P}_{2,y}(\mathbf{P}_{2,y}^{N} - 1)}{\mathbf{P}_{2,y} - 1} + b \cdot \mathbf{P}_{2,x}^{N} \cdot \sum_{i=1}^{\infty} \mathbf{P}_{1,x}^{i} + (1-b)\mathbf{P}_{2,y}^{N} \cdot \sum_{i=1}^{\infty} \mathbf{P}_{1,y}^{i} \\
=&b \cdot \frac{\mathbf{P}_{2,x}(\mathbf{P}_{2,x}^{N} - 1)}{\mathbf{P}_{2,x} - 1} + (1-b) \cdot \frac{\mathbf{P}_{2,y}(\mathbf{P}_{2,y}^{N} - 1)}{\mathbf{P}_{2,y} - 1} + b \cdot \mathbf{P}_{2,x}^{N} \cdot \frac{\mathbf{P}_{1,x}}{1 - \mathbf{P}_{1,x}} + (1-b)\mathbf{P}_{2,y}^{N} \frac{\mathbf{P}_{1,y}}{1 - \mathbf{P}_{1,y}} \\
=&\mathbf{P}_{2,x}^{N} \cdot b\Big(\frac{\mathbf{P}_{2,x}}{\mathbf{P}_{2,x} - 1} + \frac{\mathbf{P}_{1,x}}{1 - \mathbf{P}_{1,x}}\Big) + \mathbf{P}_{2,y}^{N}(1-b)\Big(\frac{\mathbf{P}_{2,y}}{\mathbf{P}_{2,y} - 1} + \frac{\mathbf{P}_{1,y}}{1 - \mathbf{P}_{1,y}}\Big) + \frac{b\mathbf{P}_{2,x}}{1 - \mathbf{P}_{2,x}} + \frac{(1-b)\mathbf{P}_{2,y}}{1 - \mathbf{P}_{2,y}}.
\end{aligned} \tag{D.8}$$

Equation (D.8) could be cast as $c_1 \cdot \mathbf{P}_{2,x}^{N} + c_2\mathbf{P}_{2,y}^{N} + c_3(\mathbf{P}_{2,x}, \mathbf{P}_{2,y})$ for **positive** c_1, **negative** c_2 and positive c_3. Let $f : \mathbb{R} \leftarrow \mathbb{R}$ be the continuous function such that $f(N) = c_1 \cdot \mathbf{P}_{2,x}^{N} + c_2\mathbf{P}_{2,y}^{N} + c_3(\mathbf{P}_{2,x}, \mathbf{P}_{2,y})$.

We take the derivative w.r.t. N to find the saddle point of f:

$$\frac{d}{dN} f = c_1 \cdot \ln \mathbf{P}_{2,x} \cdot \mathbf{P}_{2,x}^{N} + c_2 \ln \mathbf{P}_{2,y} \cdot \mathbf{P}_{2,y}^{N} = 0,$$

which suggests the saddle point of f is

$$\tilde{N} = \frac{\ln\big(-\frac{c_2 \ln \mathbf{P}_{2,y}}{c_1 \ln \mathbf{P}_{2,x}}\big)}{\ln\big(\frac{\mathbf{P}_{2,x}}{\mathbf{P}_{2,y}}\big)}.$$

Next, set $N^* \stackrel{\text{def}}{=} \max\{0, \tilde{N}\}$. Since f has a single saddle point and for every $n \in \mathbb{N}$ it holds that $f(N) = \mathbb{E}[V^{\pi_N}(b)]$, to determine the optimal policy, one only needs to compare the value $\mathbb{E}[V^{\pi_N}(b)]$ at the boundary points ($N = 0, N = \infty$) and at the closest integers to the saddle point ($N = \lfloor N^* \rfloor, N = \lceil N^* \rceil$). □

D.5.4 Dominant Diagonal (SD)

Theorem 5.5

For any instance such that $\mathbf{P}$ has a dominant diagonal, either π^1 or π^2 is optimal.

Proof. We prove the following claim by a backward induction over the number of iterations remaining: For every $k = H-1, \dots 1$ it holds that for every policy π and belief b,

$$\mathbb{E}[V_k^{\pi}(b)] \leq \max\{\mathbb{E}[V_k^{\pi^1}(b)], \mathbb{E}[V_k^{\pi^2}(b)]\}.$$

First, we notice that when $k = H-1$, the only possible policies are π^1 and π^2. For $k = H-2$, we prove the statement by contradiction. There are only two ways to selects topics when $k = H-2$:

$$\pi'_{1:H} = (\pi_{1:H-2}, \underbrace{1}_{H-1}, \underbrace{2}_{H}) \quad \text{and} \quad \pi''_{1:H} = (\pi_{1:H-2}, \underbrace{2}_{H-1}, \underbrace{1}_{H}).$$

Let m_1 and m_2 denote the number of times π has played topic 1 and 2 till time $H-2$, inclusive. Assume that the policy π' is optimal. In particular, it holds that $\mathbb{E}[V_k^{\pi_1}] \leq \mathbb{E}[V_k^{\pi'}]$ and $\mathbb{E}[V_k^{\pi_2}] \leq \mathbb{E}[V_k^{\pi'}]$. We get

$$b\mathbf{P}_{1,x}^{m_1}\mathbf{P}_{2,x}^{m_2}\mathbf{P}_{1,x}(\mathbf{P}_{1,x} - \mathbf{P}_{2,x}) \leq \mathbf{P}_{1,y}^{m_1}\mathbf{P}_{2,y}^{m_2}(1-b)\mathbf{P}_{1,y}(\mathbf{P}_{2,y} - \mathbf{P}_{1,y}), \tag{D.9}$$

and

$$\mathbf{P}_{1,y}^{m_1}\mathbf{P}_{2,y}^{m_2}(1-b)(\mathbf{P}_{2,y} - \mathbf{P}_{1,y})(1+\mathbf{P}_{2,y}) \leq b\mathbf{P}_{1,x}^{m_1}\mathbf{P}_{2,x}^{m_2}(\mathbf{P}_{1,x} - \mathbf{P}_{2,x})(1+\mathbf{P}_{2,x}). \tag{D.10}$$

As both sides of (D.9) and (D.10) are positive, the multiplication of their left hand sides is smaller than the multiplication of their right hand sides, i.e.,

$$\begin{aligned} &b\mathbf{P}_{1,x}^{m_1}\mathbf{P}_{2,x}^{m_2}\mathbf{P}_{1,x}(\mathbf{P}_{1,x} - \mathbf{P}_{2,x})\mathbf{P}_{1,y}^{m_1}\mathbf{P}_{2,y}^{m_2}(1-b)(\mathbf{P}_{2,y} - \mathbf{P}_{1,y})(1+\mathbf{P}_{2,y}) \\ &\leq \mathbf{P}_{1,y}^{m_1}\mathbf{P}_{2,y}^{m_2}(1-b)\mathbf{P}_{1,y}(\mathbf{P}_{2,y} - \mathbf{P}_{1,y})b\mathbf{P}_{1,x}^{m_1}\mathbf{P}_{2,x}^{m_2}(\mathbf{P}_{1,x} - \mathbf{P}_{2,x})(1+\mathbf{P}_{2,x}) \end{aligned}$$

Dividing both sides by $b\mathbf{P}_{1,x}^{m_1}\mathbf{P}_{2,x}^{m_2}(\mathbf{P}_{1,x} - \mathbf{P}_{2,x})\mathbf{P}_{1,y}^{m_1}\mathbf{P}_{2,y}^{m_2}(1-b)(\mathbf{P}_{2,y} - \mathbf{P}_{1,y}) > 0$, we obtain

$$\mathbf{P}_{1,x}(1+\mathbf{P}_{2,y}) \leq \mathbf{P}_{1,y}(1+\mathbf{P}_{2,x}),$$

which is a contradiction as $\mathbf{P}_{1,x} > \mathbf{P}_{1,y}$ and $1+\mathbf{P}_{2,y} > 1+\mathbf{P}_{2,x}$.

Now, assume that the policy π'' is optimal. In particular, it holds that $\mathbb{E}[V_k^{\pi^1}] \leq \mathbb{E}[V_k^{\pi''}]$ and $\mathbb{E}[V_k^{\pi^2}] \leq \mathbb{E}[V_k^{\pi''}]$. We get

$$\mathbf{P}_{1,x}^{m_1}\mathbf{P}_{2,x}^{m_2}b(\mathbf{P}_{1,x} - \mathbf{P}_{2,x})(1+\mathbf{P}_{1,x}) \leq \mathbf{P}_{1,y}^{m_1}\mathbf{P}_{2,y}^{m_2}(1-b)(1+\mathbf{P}_{1,y})(\mathbf{P}_{2,y} - \mathbf{P}_{1,y}), \tag{D.11}$$

and

$$\mathbf{P}_{1,y}^{m_1}\mathbf{P}_{2,y}^{m_2}(1-b)\mathbf{P}_{2,y}(\mathbf{P}_{2,y} - \mathbf{P}_{1,y}) \leq \mathbf{P}_{1,x}^{m_1}\mathbf{P}_{2,x}^{m_2}b\mathbf{P}_{2,x}(\mathbf{P}_{1,x} - \mathbf{P}_{2,x}). \tag{D.12}$$

As both sides of (D.11) and (D.12) are positive, the multiplication of their left hand sides is smaller than the multiplication of their right hand sides,

$$\begin{aligned}&\mathbf{P}_{1,x}^{m_1}\mathbf{P}_{2,x}^{m_2}b(\mathbf{P}_{1,x}-\mathbf{P}_{2,x})(1+\mathbf{P}_{1,x})\mathbf{P}_{1,y}^{m_1}\mathbf{P}_{2,y}^{m_2}(1-b)\mathbf{P}_{2,y}(\mathbf{P}_{2,y}-\mathbf{P}_{1,y})\\ \leq\ &\mathbf{P}_{1,y}^{m_1}\mathbf{P}_{2,y}^{m_2}(1-b)(1+\mathbf{P}_{1,y})(\mathbf{P}_{2,y}-\mathbf{P}_{1,y})\mathbf{P}_{1,x}^{m_1}\mathbf{P}_{2,x}^{m_2}b\mathbf{P}_{2,x}(\mathbf{P}_{1,x}-\mathbf{P}_{2,x}).\end{aligned}$$

Dividing both sides by $\mathbf{P}_{1,x}^{m_1}\mathbf{P}_{2,x}^{m_2}b(\mathbf{P}_{1,x}-\mathbf{P}_{2,x})\mathbf{P}_{1,y}^{m_1}\mathbf{P}_{2,y}^{m_2}(1-b)(\mathbf{P}_{2,y}-\mathbf{P}_{1,y})>0$, we obtain

$$\mathbf{P}_{2,y}(1+\mathbf{P}_{1,x})\leq\mathbf{P}_{2,x}(1+\mathbf{P}_{1,y}),$$

which is again, a contradiction as $\mathbf{P}_{2,x}<\mathbf{P}_{2,y}$ and $1+\mathbf{P}_{1,y}<1+\mathbf{P}_{1,x}$.

For $H\geq 3$, we prove the statement by contradiction. Suppose not, i.e., the optimal policy π switch recommended topic at least once. Let t denote the time step where π switch for the last time. We first consider the case where π has switched from topic 2 to topic 1 at time t. More specifically, we have

$$\pi_{1:H}=(\pi_{1:t-2},\underbrace{2}_{\pi_{t-1}},\underbrace{1}_{\pi_t},\underbrace{1,\ldots,1}_{\pi_{t+1:H-1}},\underbrace{1}_{\pi_H}).$$

Consider another policy $\tilde{\pi}$ (that behaves the same as π except at time step $t-1$) defined as

$$\tilde{\pi}_{1:H}=(\pi_{1:t-2},\underbrace{2}_{\pi_{t-1}},\underbrace{2}_{\pi_t},\underbrace{1,\ldots,1}_{\pi_{t+1:H-1}},\underbrace{1}_{\pi_H}).$$

Let m_1 and m_2 denote the number of times π has recommended topic 1 and 2 till (and include) time $t-1$. Since π is optimal, we have the difference between the value of π and $\tilde{\pi}$ to be non-negative, i.e.,

$$\mathbb{E}[V_H^{\pi}]-\mathbb{E}[V_H^{\tilde{\pi}}]=\sum_{i=1}^{H-t+1}b\mathbf{P}_{1,x}^{m_1+i-1}\mathbf{P}_{2,x}^{m_2+1}(\mathbf{P}_{1,x}-\mathbf{P}_{2,x})+(1-b)\mathbf{P}_{1,y}^{m_1+i-1}\mathbf{P}_{2,y}^{m_2+1}(\mathbf{P}_{1,y}-\mathbf{P}_{2,y})\geq 0, \tag{D.13}$$

where the difference is induced by the discrepancy of the two policies from time step t to H. Consider another policy π' (that behaves the same as π except at time step H) defined as

$$\pi'_{1:H}=(\pi_{1:t-2},\underbrace{2}_{\pi_{t-1}},\underbrace{1}_{\pi_t},\underbrace{1,\ldots,1}_{\pi_{t+1:H-1}},\underbrace{2}_{\pi_H}).$$

Since π is optimal, we have the difference between the value of π and π' to be non-negative, i.e.,

$$\mathbb{E}[V_H^{\pi}]>\mathbb{E}[V_H^{\pi'}]\Rightarrow b\mathbf{P}_{1,x}^{m_1+H-t}\mathbf{P}_{2,x}^{m_2}(\mathbf{P}_{1,x}-\mathbf{P}_{2,x})>(1-b)\mathbf{P}_{1,y}^{m_1+H-t}\mathbf{P}_{2,y}^{m_2}(\mathbf{P}_{2,y}-\mathbf{P}_{1,y}),$$

where the difference is induced by the discrepancy of the two policies from time step H. Multiplying both sides by $\mathbf{P}_{1,y}>0$, we get

$$\mathbf{P}_{1,y}b\mathbf{P}_{1,x}^{m_1+H-t}\mathbf{P}_{2,x}^{m_2}(\mathbf{P}_{1,x}-\mathbf{P}_{2,x})>(1-b)\mathbf{P}_{1,y}^{m_1+H-t+1}\mathbf{P}_{2,y}^{m_2}(\mathbf{P}_{2,y}-\mathbf{P}_{1,y}).$$

Using $\frac{\mathbf{P}_{1,x}}{\mathbf{P}_{1,y}} > 1$, and $\mathbf{P}_{1,y} b \mathbf{P}_{1,x}^{m_1+H-t} \mathbf{P}_{2,x}^{m_2} (\mathbf{P}_{1,x} - \mathbf{P}_{2,x}) > 0$,

$$b\mathbf{P}_{1,x}^{m_1+H-t+1} \mathbf{P}_{2,x}^{m_2} (\mathbf{P}_{1,x} - \mathbf{P}_{2,x}) > (1-b)\mathbf{P}_{1,y}^{m_1+H-t+1} \mathbf{P}_{2,y}^{m_2} (\mathbf{P}_{2,y} - \mathbf{P}_{1,y});$$

hence,

$$b\mathbf{P}_{1,x}^{m_1+H-t+1} \mathbf{P}_{2,x}^{m_2} (\mathbf{P}_{1,x} - \mathbf{P}_{2,x}) + (1-b)\mathbf{P}_{1,y}^{m_1+H-t+1} \mathbf{P}_{2,y}^{m_2} (\mathbf{P}_{1,y} - \mathbf{P}_{2,y}) \geq 0. \tag{D.14}$$

Next, we construct a new policy π_{new} that outperforms π. We let π^{new} to be the policy defined as below

$$\pi_{1:H}^{\text{new}} = (\pi_{1:t-2}, \underbrace{1}_{\pi_{t-1}}, \underbrace{1}_{\pi_t}, \underbrace{1, \ldots, 1}_{\pi_{t+1:H-1}}, \underbrace{1}_{\pi_H}).$$

The value difference between π^{new} and π (caused by the discrepancy of the two policies from time $t-1$ to H) is

$$\begin{aligned}
\mathbb{E}[V_H^{\pi^{\text{new}}}] - \mathbb{E}[V_H^{\pi}] = & \sum_{i=1}^{H-t+1} b\mathbf{P}_{1,x}^{m_1+i-1} \mathbf{P}_{2,x}^{m_2} (\mathbf{P}_{1,x} - \mathbf{P}_{2,x}) + (1-b)\mathbf{P}_{1,y}^{m_1+i-1} \mathbf{P}_{2,y}^{m_2} (\mathbf{P}_{1,y} - \mathbf{P}_{2,y}) \\
& + b\mathbf{P}_{1,x}^{m_1+H-t+1} \mathbf{P}_{2,x}^{m_2} (\mathbf{P}_{1,x} - \mathbf{P}_{2,x}) + (1-b)\mathbf{P}_{1,y}^{m_1+H-t+1} \mathbf{P}_{2,y}^{m_2} (\mathbf{P}_{1,y} - \mathbf{P}_{2,y}) \\
& > \sum_{i=1}^{H-t+1} b\mathbf{P}_{1,x}^{m_1+i-1} \mathbf{P}_{2,x}^{m_2+1} (\mathbf{P}_{1,x} - \mathbf{P}_{2,x}) + (1-b)\mathbf{P}_{1,y}^{m_1+i-1} \mathbf{P}_{2,y}^{m_2+1} (\mathbf{P}_{1,y} - \mathbf{P}_{2,y}) \\
& \geq 0,
\end{aligned}$$

where the first inequality is true because $\mathbf{P}_{2,x} < \mathbf{P}_{2,y}$, $\mathbf{P}_{1,x} - \mathbf{P}_{2,x} > 0$ and $\mathbf{P}_{1,y} - \mathbf{P}_{2,y} < 0$, therefore for every $1 \leq i \leq H - t + 1$

$$b\mathbf{P}_{1,x}^{m_1+i-1} \mathbf{P}_{2,x}^{m_2} (\mathbf{P}_{1,x} - \mathbf{P}_{2,x})(1 - \mathbf{P}_{2,x}) > 0 > (1-b)\mathbf{P}_{1,y}^{m_1+i-1} \mathbf{P}_{2,y}^{m_2} (\mathbf{P}_{2,y} - \mathbf{P}_{1,y})(\mathbf{P}_{2,y} - 1)$$

along with (D.14). The second inequality follows from (D.13). Thus, we have successfully find another policy $\pi_{1:H}^{\text{new}}$ that differs from π and achieves a higher value, which is a contradiction.

next, we consider the case where π has switched from topic 1 to topic 2 at time t, i.e.,

$$\pi_{1:H} = (\pi_{1:t-2}, \underbrace{1}_{\pi_{t-1}}, \underbrace{2}_{\pi_t}, \underbrace{2, \ldots, 2}_{\pi_{t+1:H-1}}, \underbrace{2}_{\pi_H}).$$

Consider another policy $\tilde{\pi}$ (that behaves the same as π except at time step t) defined as

$$\tilde{\pi}_{1:H} = (\pi_{1:t-2}, \underbrace{1}_{\pi_{t-1}}, \underbrace{1}_{\pi_t}, \underbrace{2, \ldots, 2}_{\pi_{t+1:H-1}}, \underbrace{2}_{\pi_H}).$$

Since π is optimal, we have the difference between the value of π and $\tilde{\pi}$ to be non-negative, i.e.,

$$\mathbb{E}[V_H^{\pi}] - \mathbb{E}[V_H^{\tilde{\pi}}] = \sum_{i=1}^{H-t+1} b\mathbf{P}_{1,x}^{m_1+1} \mathbf{P}_{2,x}^{m_2+i-1} (\mathbf{P}_{2,x} - \mathbf{P}_{1,x}) + (1-b)\mathbf{P}_{1,y}^{m_1+1} \mathbf{P}_{2,y}^{m_2+i-1} (\mathbf{P}_{2,y} - \mathbf{P}_{1,y}) \tag{D.15}$$

$$\geq 0, \tag{D.16}$$

where the difference follows from the discrepancy between the two policies from time step t to H.

Consider another policy π' (that behaves the same as π except at time step H) defined as

$$\pi'_{1:H} = (\pi_{1:t-2}, \underbrace{1}_{\pi_{t-1}}, \underbrace{2}_{\pi_t}, \underbrace{2, \ldots, 2}_{\pi_{t+1:H-1}}, \underbrace{1}_{\pi_H}).$$

Since π is optimal, we have the difference between the value of π and π' to be non-negative, i.e.,

$$\mathbb{E}[V_H^{\pi}] > \mathbb{E}[V_H^{\pi'}] \Rightarrow (1-b)\mathbf{P}_{1,y}^{m_1}\mathbf{P}_{2,y}^{m_2+H-t}(\mathbf{P}_{2,y} - \mathbf{P}_{1,y}) \geq b\mathbf{P}_{1,x}^{m_1}\mathbf{P}_{2,x}^{m_2+H-t}(\mathbf{P}_{1,x} - \mathbf{P}_{2,x}),$$

where the difference is induced by the discrepancy of the two policies from time step H. Multiplying both sides by $\mathbf{P}_{2,x} > 0$,

$$\mathbf{P}_{2,x}(1-b)\mathbf{P}_{1,y}^{m_1}\mathbf{P}_{2,y}^{m_2+H-t}(\mathbf{P}_{2,y} - \mathbf{P}_{1,y}) \geq b\mathbf{P}_{1,x}^{m_1}\mathbf{P}_{2,x}^{m_2+H-t+1}(\mathbf{P}_{1,x} - \mathbf{P}_{2,x}).$$

Using $\mathbf{P}_{2,x}(1-b)\mathbf{P}_{1,y}^{m_1}\mathbf{P}_{2,y}^{m_2+H-t}(\mathbf{P}_{2,y} - \mathbf{P}_{1,y}) > 0$ and $\frac{\mathbf{P}_{2,y}}{\mathbf{P}_{2,x}} \geq 1$, we get

$$(1-b)\mathbf{P}_{1,y}^{m_1}\mathbf{P}_{2,y}^{m_2+H-t+1}(\mathbf{P}_{2,y} - \mathbf{P}_{1,y}) \geq b\mathbf{P}_{1,x}^{m_1}\mathbf{P}_{2,x}^{m_2+H-t+1}(\mathbf{P}_{1,x} - \mathbf{P}_{2,x});$$

hence,

$$b\mathbf{P}_{1,x}^{m_1}\mathbf{P}_{2,x}^{m_2+H-t+1}(\mathbf{P}_{2,x} - \mathbf{P}_{1,x}) + (1-b)\mathbf{P}_{1,y}^{m_1}\mathbf{P}_{2,y}^{m_2+H-t+1}(\mathbf{P}_{2,y} - \mathbf{P}_{1,y}) \geq 0. \tag{D.17}$$

Again, we will construct a new policy π_{new} that outperforms π. We let π^{new} to be the policy defined as below

$$\pi^{\text{new}}_{1:H} = (\pi_{1:t-2}, \underbrace{2}_{\pi_{t-1}}, \underbrace{2}_{\pi_t}, \underbrace{2, \ldots, 2}_{\pi_{t+1:H-1}}, \underbrace{1}_{\pi_H}).$$

Now, the value difference between π^{new} and π (caused by the discrepancy of the two policies from time $t-1$ to H) is

$$\begin{aligned}
\mathbb{E}[V_H^{\pi^{\text{new}}}] - \mathbb{E}[V_H^{\pi}] = &\sum_{i=1}^{H-t+1} \left(b\mathbf{P}_{1,x}^{m_1}\mathbf{P}_{2,x}^{m_2+i-1}(\mathbf{P}_{2,x} - \mathbf{P}_{1,x}) + (1-b)\mathbf{P}_{1,y}^{m_1}\mathbf{P}_{2,y}^{m_2+i-1}(\mathbf{P}_{2,y} - \mathbf{P}_{1,y})\right) \\
&+ b\mathbf{P}_{1,x}^{m_1}\mathbf{P}_{2,x}^{m_2+H-t+1}(\mathbf{P}_{2,x} - \mathbf{P}_{1,x}) + (1-b)\mathbf{P}_{1,y}^{m_1}\mathbf{P}_{2,y}^{m_2+i-1}(\mathbf{P}_{2,y} - \mathbf{P}_{1,y}) \\
&> \sum_{i=1}^{H-t+1} b\mathbf{P}_{1,x}^{m_1+1}\mathbf{P}_{2,x}^{m_2+i-1}(\mathbf{P}_{2,x} - \mathbf{P}_{1,x}) + (1-b)\mathbf{P}_{1,y}^{m_1+1}\mathbf{P}_{2,y}^{m_2+i-1}(\mathbf{P}_{2,y} - \mathbf{P}_{1,y}) \\
&\geq 0,
\end{aligned}$$

where the first inequality is true because $\mathbf{P}_{1,y} < \mathbf{P}_{1,x}$, $\mathbf{P}_{2,x} - \mathbf{P}_{1,x} < 0$ and $\mathbf{P}_{2,y} - \mathbf{P}_{1,y} > 0$ and (D.17), and the second from (D.15). Similarly, we have successfully find another policy $\pi^{\text{new}}_{1:H}$ that differs from π and achieves a higher value, which is a contradiction.

We have covered all cases, so the inductive argument holds. This concludes the proof of the theorem. □

D.5.5 UCB-based regret bound

Lemma 5.5

Let $\tilde{\tau} = \frac{8e}{\ln(\frac{1}{1-\epsilon})}$ and $\eta = 1$. For every threshold policy $\pi \in \Pi_H$, the centred random variable $V^\pi - \mathbb{E}[V^\pi]$ is (τ^2, b)-sub-exponential with (τ^2, b) satisfying $\tilde{\tau} \geq \tau$ and $\eta \geq b^2/\tau^2$.

Proof. Let γ be such that

$$|\gamma| < -\frac{\ln(1-\epsilon)}{8e} \leq \min_{a\in\{1,2\},i\in\{x,y\}}\{-\frac{\ln(1-\mathcal{L}_{a,i}(1-\mathbf{P}_{a,i}))}{8e}\} = \min_{a\in\{1,2\},i\in\{x,y\}}\{-\frac{\ln(\mathbf{P}_{a,i})}{8e}\}.$$

Next, we have that

$$\begin{aligned}
&\mathbb{E}[\exp(\gamma(V^\pi - \mathbb{E}[V^\pi]))] \\
&\leq \sum_{a\in\{1,2\}} \mathbb{E}[\exp(\gamma(V^{\pi_a} - \mathbb{E}[V^{\pi_a}]))] \Big| type(t) \in \arg\max_{i\in[1,2]} \mathbf{P}_{a,i}] \cdot \mathbb{P}[type(t) \in \arg\max_{i\in[1,2]} \mathbf{P}_{a,i}] \\
&\leq \max_{a\in\{1,2\}} \{\mathbb{E}[\exp(\gamma(\bar{V}^{\pi^a} - \mathbb{E}[\bar{V}^{\pi^a}]))]\}
\end{aligned}$$

Where $\bar{V}^{\pi_a}$ is the return for the instance $\langle [1],[2],\mathbf{q},\bar{\mathbf{P}},\bar{\mathcal{L}}\rangle$ such that for every $a \in \{1,2\}$: $\bar{\mathbf{P}}_{a,1} = \max_{i\in\{x,y\}} \mathbf{P}_{a,i}$ and $\mathcal{L}_{a,1} = 1$.

Finally, from Lemma 5.2 we get

$$\begin{aligned}
&\max_{a\in\{1,2\}} \{\mathbb{E}[\exp(\gamma(\bar{V}^{\pi^a} - \mathbb{E}[\bar{V}^{\pi^a}]))]\} \\
&\leq \max_{a\in\{1,2\}} \exp((-\frac{8e}{\ln(\bar{\mathbf{P}}_{a,1})})^2 \frac{\gamma^2}{2}) \\
&= \max_{a\in\{1,2\},i\in\{x,y\}} \exp((-\frac{8e}{\ln(\mathbf{P}_{a,i})})^2 \frac{\gamma^2}{2}).
\end{aligned}$$

Choosing

$$\tau = b = \max_{a\in\{1,2\},i\in\{x,y\}} -\frac{8e}{\ln(\mathbf{P}_{a,i})}$$

completes the proof as

$$\max_{a\in\{1,2\},i\in\{x,y\}} -\frac{8e}{\ln(\mathbf{P}_{a,i})} \leq -\frac{8e}{\ln(1-\epsilon)} = \tilde{\tau} \quad \text{and} \quad \frac{\tau^2}{b^2} = 1 = \eta.$$

□

Lemma 5.4

For every $H \in \mathbb{N}$, it holds that

$$\mathbb{E}\left[V^{\pi^*} - \max_{\pi \in \Pi_H} V^{\pi}\right] \leq \frac{1}{2^{O(H)}}.$$

Proof. Recall that $V^{\pi} = \sum_{j=1}^{N^{\pi}} r_j(\pi_j)$, where we drop the dependence on the user index for readability. Formulating differently, for any $H \in \mathbb{N}$ it holds that

$$V^{\pi} = \sum_{j=1}^{H} \mathbf{1}_{j \leq N^{\pi}} \cdot r_j(\pi_j) + \sum_{j=H+1}^{\infty} \mathbf{1}_{j \leq N^{\pi}} \cdot r_j(\pi_j).$$

Using the same representation for $V^{\pi'}$ and taking expectation, we get that

$$\begin{aligned}
\mathbb{E}\left[V^{\pi} - V^{\pi'}\right] &\leq \mathbb{E}\left[\sum_{j=1}^{H} \mathbf{1}_{j \leq N^{\pi}} \cdot r_j(\pi_j) - \sum_{j=1}^{H} \mathbf{1}_{j \leq N^{\pi'}} \cdot r_j(\pi'_j)\right] + \mathbb{E}\left[\sum_{j=H+1}^{\infty} \mathbf{1}_{j \leq N^{\pi}} \cdot r_j(\pi_j)\right] \\
&\leq 0 + \mathbb{E}\left[\sum_{j=H+1}^{\infty} \mathbf{1}_{j \leq N^{\pi}} \cdot r_j(\pi_j)\right] = \sum_{j=H+1}^{\infty} \mathbb{P}\left(j \leq N^{\pi}\right) r_j(\pi_j) \\
&\leq \sum_{j=H+1}^{\infty} (1-\epsilon)^j (1-\epsilon) = (1-\epsilon)^{H+2} \sum_{j=0}^{\infty} (1-\epsilon)^j \\
&\leq (1-\epsilon)^H \frac{1}{\epsilon} \leq \frac{e^{-\epsilon H}}{\epsilon} = \frac{1}{2^{O(H)}}.
\end{aligned}$$

□

Appendix E

Risk-sensitive Supervised Learning: Additional Details

E.1 Hölder Risk Functionals

In the definition of Hölder risk functionals, we require d to be a quasi-metric, which we provide the definition here.

Definition E.1

A function $d : \mathcal{L}_\infty(\mathbb{R}, \mathbb{B}(\mathbb{R})) \times \mathcal{L}_\infty(\mathbb{R}, \mathbb{B}(\mathbb{R})) \to [0, +\infty)$ is a quasi-metric if the following two conditions hold:

- For all $F_U, F_{U'} \in \mathcal{L}_\infty(\mathbb{R}, \mathbb{B}(\mathbb{R}))$, $d(F_U, F_{U'}) = 0$ if and only if $F_U = F_{U'}$;
- For all $F_U, F_{U'}, F_{Z''} \in \mathcal{L}_\infty(\mathbb{R}, \mathbb{B}(\mathbb{R}))$, $d(F_U, F_{Z''}) \leq d(F_U, F_{U'}) + d(F_{U'}, F_{Z''})$.

If a quasimetric is symmetic, i.e., for all $F_U, F_{U'} \in \mathcal{L}_\infty(\mathbb{R}, \mathbb{B}(\mathbb{R}))$, $d(F_U, F_{U'}) = d(F_{U'}, F_U)$, it is also a *metric*. The set of quasi-metrics contains symmetric quasi-metics, e.g., sup norms $\mathcal{L}_\infty$, Wasserstein distance, along with non-symmetric quasi-metrics, e.g., Kullback-Leibler divergence.

We will now discuss why optimized certainty equivalent (OCE) risks (e.g., mean-variance, entropic risk, CVaR) and spectral risks with bounded spectrum (e.g., CVaR, certain CPT-inspired Risks) are Lipschitz on bounded random variables. OCE risks, first introduced by Ben-Tal and Teboulle (1986), are defined as

$$\rho_{\text{oce}}(F_U) := \inf_{\lambda \in \mathbb{R}} \{\lambda + \mathbb{E}[\phi(U - \lambda)]\},$$

where $\phi : \mathbb{R} \to \mathbb{R} \cup \{+\infty\}$ is a nondecreasing, closed and convex function with $\phi(0) = 0$ and $1 \in \partial\phi(0)$. To complement the risk-averse OCEs, a risk-seeking version (inverted OCE) is proposed:

$$\rho_{\overline{\text{oce}}}(F_U) := \sup_{\lambda \in \mathbb{R}} \{\lambda - \mathbb{E}[\phi(\lambda - U)]\}.$$

Proposition E.1

If ϕ is continuously differentiable, then the OCE risks ρ_{oce} and inverted OCE risks $\rho_{\overline{\text{oce}}}$ are Lipschitz on the space of bounded random variables with support $[0, D]$:

$$|\rho_{\text{oce}}(F_U) - \rho_{\text{oce}}(F_{U'})| \leq \max_{x\in[0,D]} (\phi(D-x) - \phi(-x)) \|F_U - F_{U'}\|_\infty,$$
$$|\rho_{\overline{\text{oce}}}(F_U) - \rho_{\overline{\text{oce}}}(F_{U'})| \leq \max_{x\in[0,D]} (\phi(x-D) - \phi(x)) \|F_U - F_{U'}\|_\infty.$$

Remark E.1. *Similar to Huang et al. (2021, Lemma 4.1), Proposition E.1 shows that the expected value and CVaR$_\alpha$ are D- and $\frac{D}{\alpha}$-Lipschtiz on random variables with support $[0, D]$, respectively. In addition, this result also provides Lipschitzness of the entropic risks (since the corresponding ϕ is continuously differentiable (Lee et al., 2020, Table 1)) and other OCE and inverted OCE risks that do not belong to distortion risk functionals.*

Proof. When U has support $[0, D]$, as shown in Lee et al. (2020, Lemma 9), we can re-write the OCE and inverted OCE risks as follows:

$$\rho_{\text{oce}}(F_U) = \min_{\lambda\in[0,D]} \{\lambda + \mathbb{E}[\phi(U-\lambda)]\}$$
$$\rho_{\overline{\text{oce}}}(F_U) = \max_{\lambda\in[0,D]} \{\lambda - \mathbb{E}[\phi(\lambda-U)]\}.$$

For any $U, U' \in [0, D]$, denote $\lambda_U \in \arg\min_{\lambda\in[0,D]} \lambda + \mathbb{E}_U[\phi(U-\lambda)]$ and $\lambda_{U'} \in \arg\min_{\lambda\in[0,D]} \lambda + \mathbb{E}_{U'}[\phi(U'-\lambda)]$. Consider the case where $\rho_{\text{oce}}(F_U) < \rho_{\text{oce}}(F_{U'})$.

$$\begin{aligned}
&|\rho_{\text{oce}}(F_U) - \rho_{\text{oce}}(F_{U'})| \\
&= \rho_{\text{oce}}(F_{U'}) - \rho_{\text{oce}}(F_U) \\
&= \lambda_{U'} + \mathbb{E}_{U'}[\phi(U'-\lambda_{U'})] - \lambda_U - \mathbb{E}_U[\phi(U-\lambda_U)] \\
&\overset{(i)}{\leq} \lambda_U + \mathbb{E}_{U'}[\phi(U'-\lambda_U)] - \lambda_U - \mathbb{E}_U[\phi(U-\lambda_U)] \\
&= \int_0^D \phi(u-\lambda_U) d\left(F_{U'}(u) - F_U(u)\right) \\
&\overset{(ii)}{=} \phi(u-\lambda_U)\left(F_{U'}(u) - F_U(u)\right)\Big|_{u=0}^{u=D} - \int_0^D \phi'(u-\lambda_U)\left(F_{U'}(u) - F_U(u)\right) du \\
&\overset{(iii)}{\leq} \|F_U - F_{U'}\|_\infty \int_0^D \phi'(u-\lambda_U) du \\
&= \left(\phi(D-\lambda_U) - \phi(-\lambda_U)\right) \|F_U - F_{U'}\|_\infty,
\end{aligned}$$

where (i) comes from the definition of $\lambda_{U'}$, (ii) uses integration by parts, and (iii) uses the fact that ϕ is non-decreasing, i.e., ϕ' is non-negative, and $F_U(0) = F_{U'}(0) = 0, F_U(D) = F_{U'}(D) = 1$. The case when $\rho_{\text{oce}}(F_{U'}) < \rho_{\text{oce}}(F_U)$ proceeds similarly:

$$|\rho_{\text{oce}}(F_U) - \rho_{\text{oce}}(F_{U'})|$$

$$
\begin{aligned}
&= \rho_{\text{oce}}(F_U) - \rho_{\text{oce}}(F_{U'}) \\
&\leq \int_0^D \phi(u - \lambda_{U'}) d\left(F_U(u) - dF_{U'}(u)\right) \\
&= \phi(u - \lambda_{U'})\left(F_U(u) - F_{U'}(u)\right)\Big|_{u=0}^{u=D} - \int_0^D \phi'(u - \lambda_{U'})\left(F_U(u) - F_{U'}(u)\right) du \\
&\leq \left(\phi(D - \lambda_{U'}) - \phi(-\lambda_{U'})\right) \|F_U - F_{U'}\|_\infty.
\end{aligned}
$$

Putting it together, we have that

$$
|\rho_{\text{oce}}(F_U) - \rho_{\text{oce}}(F_{U'})| \leq \max_{\lambda \in [0,D]} (\phi(D - x) - \phi(-x)) \|F_U - F_{U'}\|_\infty.
$$

For inverted OCE risks, denote $\lambda_U \in \arg\max_{\lambda \in [0,D]} \lambda + \mathbb{E}_U[\phi(\lambda - U)]$ and $\lambda_{U'} \in \arg\max_{\lambda \in [0,D]} \lambda + \mathbb{E}_{U'}[\phi(\lambda - U)]$. The proof proceeds similarly by using the fact that $\rho_{\overline{\text{oce}}}(F_{U'}) - \rho_{\overline{\text{oce}}}(F_U) \leq \mathbb{E}_U[\phi(\lambda_{U'} - U)] - \mathbb{E}_{U'}[\phi(\lambda_{U'} - U')]$ and $\rho_{\overline{\text{oce}}}(F_U) - \rho_{\overline{\text{oce}}}(F'_U) \leq \mathbb{E}_{U'}[\phi(\lambda_U - U')] - \mathbb{E}_U[\phi(\lambda_U - U)]$. Following similar steps, we obtain that

$$
|\rho_{\overline{\text{oce}}}(F_U) - \rho_{\overline{\text{oce}}}(F_{U'})| \leq \max_{\lambda \in [0,D]} (\phi(x - D) - \phi(x)) \|F_U - F_{U'}\|_\infty.
$$

□

Spectral risks (also known as L-risks or rank-weighted risks) are a subset of distortion risk functionals. As noted in Bäuerle and Glauner (2021), a spectral risk can be written as a distortion risk (Equation (6.4)) with the following distortion function: for $t \in [0, 1]$,

$$
g(t) = \int_0^t h(s) ds,
$$

where $h : [0, 1] \to \mathbb{R}_+$ is the non-decreasing spectrum function that integrates to 1. Since g is Lipschitz when h is bounded (i.e., $g'(t) = h(t)$ for $t \in (0, 1)$), spectral risks are Lipschitz on the space of bounded random variables when their spectrum is bounded.

Finally, examples of risk functionals that are Hölder but are not Lipschitz include distortion risks whose distortion functions are Hölder but not Lipschitz.

E.2 Proof of Results in Section 6.4

We note that in the following proofs, Z is used to denote a generic random variable and the loss function is denoted by $\ell : \mathcal{X} \times \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$. (The loss function presented in the main text is a special case of this.) For a given $(X, Y, f(X))$, the loss is denoted by $\ell(X, Y, f(X))$.

E.2.1 Auxillary Lemmas

The below two auxillary lemmas are mainly adaptations to the class note from Professor Roberto Imbuzeiro Oliveira (Oliveira and Yang, n.d.).

Lemma E.1

Let $\mathbb{G}(1) := \left\{ g(\cdot\,; r) := \mathbb{1}_{\{\cdot \le r\}} : \forall r \in \mathbb{R} \right\}$. For a fixed $f \in \mathbb{F}$, iid sample $\{X_i, Y_i\}_{i=1}^n$ with joint probability measure $\mathbb{P}$, and a loss function $\ell : \mathcal{X} \times \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$, we have that

$$\sup_{g \in \mathbb{G}(1)} \left| \sum_{i=1}^n \xi_i g(\ell(X_i, Y_i, f(X_i))) \right| = \max_{j \in [n]} \left| \sum_{i=1}^j \xi_i \right|, \tag{E.1}$$

and further

$$\begin{aligned} &\mathbb{E}_{\mathbb{P},\mathbb{R}} \left[\exp\left(\frac{\lambda}{n} \sup_{g \in \mathbb{G}(1)} \left| \sum_{i=1}^n \xi_i g(\ell(X_i, Y_i, f(X_i))) \right| \right) \right] \\ &\le 2\mathbb{E}_{\mathbb{R}} \left[\max_{j \in [n]} \left(\exp\left(\frac{\lambda}{n} \sum_{i=1}^j \xi_i \right) \mathbb{1}_{\{\sum_{i=1}^j \xi_i \ge 0\}} \right) \right]. \end{aligned} \tag{E.2}$$

Remark E.2. *We note that an important property of Lemma E.1 is that the bound is independent of the samples* $\{X_i, Y_i\}_{i=1}^n$.

Proof. Let $\{R_i\}_{i=1}^n$ denote the sorted sequence of $\{\ell(X_i, Y_i, f(X_i))\}_{i=1}^n$, where $R_1 \le R_2 \ldots \le R_n$. Using $\{R_i\}_{i=1}^n$, we have

$$\mathbb{E}_{\mathbb{P},\mathbb{R}} \left[\exp\left(\frac{\lambda}{n} \sup_{g \in \mathbb{G}(1)} \left| \sum_{i=1}^n \xi_i g(\ell(X_i, Y_i, f(X_i))) \right| \right) \right] = \mathbb{E}_{\mathbb{P},\mathbb{R}} \left[\exp\left(\frac{\lambda}{n} \sup_{g \in \mathbb{G}(1)} \left| \sum_{i=1}^n \xi_i g(R_i) \right| \right) \right]$$

Consider a function $g(t; r) = \mathbb{1}_{\{t \le r\}}$. For such a function, $\sum_{i=1}^n \xi_i g(R_i; r)$ is equal to

- 0 if $r < \min_{i \in [n]} R_i$,
- $\sum_{i=1}^j \xi_i$ when $R_j \le r < R_{j+1}$ for some $j \in \{1, \ldots, n-1\}$,
- $\sum_{i=1}^n \xi_i$ otherwise.

Therefore, we have that

$$\sup_{g\in\mathbb{G}(1)}\left|\sum_{i=1}^{n}\xi_i g(\ell(X_i,Y_i,f(X_i)))\right| = \max_{j\in[n]}\left|\sum_{i=1}^{j}\xi_i\right|.$$

Finally, we notice that

$$\begin{aligned}
&\mathbb{E}_{\mathbb{P},\mathbb{R}}\left[\sup_{g\in\mathbb{G}(1)}\exp\left(\frac{\lambda}{n}\left|\sum_{i=1}^{n}\xi_i g(R_i)\right|\right)\right] = \mathbb{E}_{\mathbb{P},\mathbb{R}}\left[\max_{j\in[n]}\exp\left(\frac{\lambda}{n}\left|\sum_{i=1}^{j}\xi_i\right|\right)\right]\\
=&\mathbb{E}_{\mathbb{P},\mathbb{R}}\left[\max_{j\in[n]}\left(\exp\left(\frac{\lambda}{n}\sum_{i=1}^{j}\xi_i\right)\mathbb{1}_{\{\sum_{i=1}^{j}\xi_i\geq 0\}} + \exp\left(-\frac{\lambda}{n}\sum i=1^{j}\xi_i\right)\mathbb{1}_{\{\sum_{i=1}^{j}\xi_i<0\}}\right)\right]\\
=&\mathbb{E}_{\mathbb{P},\mathbb{R}}\left[\max_{j\in[n]}\left(\exp\left(\frac{\lambda}{n}\sum_{i=1}^{j}\xi_i\right)\mathbb{1}_{\{\sum_{i=1}^{j}\xi_i\geq 0\}}\right)\right] + \mathbb{E}_{\mathbb{P},\mathbb{R}}\left[\max_{j\in[n]}\left(\exp\left(-\frac{\lambda}{n}\sum_{i=1}^{j}\xi_i\right)\mathbb{1}_{\{\sum_{i=1}^{j}\xi_i<0\}}\right)\right]\\
\leq&2\mathbb{E}_{\mathbb{P},\mathbb{R}}\left[\max_{j\in[n]}\left(\exp\left(\frac{\lambda}{n}\sum_{i=1}^{j}\xi_i\right)\mathbb{1}_{\{\sum_{i=1}^{j}\xi_i\geq 0\}}\right)\right]. \qquad \text{(E.3)}
\end{aligned}$$

□

Lemma E.2

Let $\{Z_i\}_{i=1}^n$ taking values in $\mathcal{Z}$ denote independent samples drawn from $\mathbb{P}$ and $w:\mathcal{Z}\to\mathbb{R}_+$. For n independent Rademacher random variables $\{\xi_i\}_{i=1}^n$, we have that for all $\lambda\geq 0$,

$$\mathbb{E}_{\mathbb{P},\mathbb{R}}\left[\exp\left(\max_{j\in[n]}\frac{\lambda}{n}\sum_{i=1}^{j}w(Z_i)\xi_i\right)\right] \leq 2\mathbb{E}_{\mathbb{P},\mathbb{R}}\left[\exp\left(\frac{\lambda}{n}\sum_{i=1}^{n}w(Z_i)\xi_i\right)\right].$$

Further, if $\frac{1}{n}\sum_{i=1}^{n}w(Z_i)\xi_i$ is mean zero $\frac{\gamma^2}{n}$-subGaussian, then

$$\mathbb{E}_{\mathbb{R}}\left[\exp\left(\max_{j\in[n]}\frac{\lambda}{n}\sum_{i}^{j}\xi_i\right)\right] \leq 2\exp\left(\frac{\lambda^2\gamma^2}{2n}\right).$$

Remark E.3. *We note that an immediate consequence of Lemma E.2 is*

$$\begin{aligned}
&\mathbb{E}_{\mathbb{P},\mathbb{R}}\left[\max_{j\in[n]}\left(\exp\left(\frac{\lambda}{n}\sum_{i=1}^{j}w(Z_i)\xi_i\right)\mathbb{1}_{\{\sum_{i=1}^{j}w(Z_i)\xi_i\geq 0\}}\right)\right]\\
\leq&\mathbb{E}_{\mathbb{P},\mathbb{R}}\left[\exp\left(\max_{j\in[n]}\frac{\lambda}{n}\sum_{i=1}^{j}w(Z_i)\xi_i\right)\right] \leq 2\mathbb{E}_{\mathbb{P},\mathbb{R}}\left[\exp\left(\frac{\lambda}{n}\sum_{i=1}^{n}w(Z_i)\xi_i\right)\right].
\end{aligned}$$

Proof. The proof contains two main steps:
Step 1: We will show that for all $t > 0$,

$$\mathbb{P}\left(\max_{j\in[n]}\sum_{i=1}^{j} w(Z_i)\xi_i \geq t\right) \leq 2\mathbb{P}\left(\sum_{i=1}^{n} w(Z_i)\xi_i \geq t\right). \tag{E.4}$$

To show (E.4), for $t > 0$, consider events $E_0 := \emptyset$ and $E_j := \{\sum_{i=1}^{j} w(Z_i)\xi_i \geq t, \sum_{i=1}^{l} w(Z_i)\xi_i < t, \forall l < j\}$ for $j \in [n]$, which states that j is the first index such that the partial sum $\sum_{i=1}^{j} w(Z_i)\xi_i$ is at least t. We first notice that

$$\left\{\max_{j}\sum_{i=1}^{j} w(Z_i)\xi_i \geq t\right\} \subset \bigcup_{j=0}^{n} E_j.$$

Since E_j and $\sum_{i=j+1}^{n} w(Z_i)\xi_i \geq 0$ implies that $\sum_{i=1}^{n} w(Z_i)\xi_i \geq t$, we obtain

$$\bigcup_{j=0}^{n}\left(E_j \bigcap \left\{\sum_{i=j+1}^{n} w(Z_i)\xi_i \geq 0\right\}\right) \subset \left\{\sum_{i=1}^{n} w(Z_i)\xi_i \geq t\right\}. \tag{E.5}$$

Moreover, for any $j \in [n]$, we have

$$\mathbb{P}\left(\sum_{i=j+1}^{n} w(Z_i)\xi_i \geq 0\right) \geq \frac{1}{2}, \tag{E.6}$$

since $\sum_{i=j+1}^{n} w(Z_i)\xi_i$ is symmetic around 0 (i.e., $\sum_{i=j+1}^{n} w(Z_i)\xi_i \overset{d}{=} -\sum_{i=j+1}^{n} w(Z_i)\xi_i$) for all $i \in \{0, \ldots, n\}$. For any $j \in [n]$, since the event E_j (dependeing on $\{w(Z_i), \xi_i\}_{i\leq j}$) is independent of $\{\sum_{i=j+1}^{n} w(Z_i)\xi_i \geq 0\}$, we have

$$\mathbb{P}\left(E_j \bigcap \left\{\sum_{i=j+1}^{n} w(Z_i)\xi_i \geq 0\right\}\right) = \mathbb{P}\left(E_j\right)\mathbb{P}\left(\sum_{i=j+1}^{n} w(Z_i)\xi_i \geq 0\right) \geq \frac{\mathbb{P}\left(E_j\right)}{2}.$$

As a result, we have

$$\begin{aligned}
\mathbb{P}\left(\sum_{i=1}^{n} w(Z_i)\xi_i \geq t\right) &\geq \mathbb{P}\left(\bigcup_{j=0}^{n}\left(E_j \bigcap \left\{\sum_{i=j+1}^{n} w(Z_i)\xi_i \geq 0\right\}\right)\right) \\
&= \sum_{j=0}^{n}\mathbb{P}\left(E_j \bigcap \left\{\sum_{i=j+1}^{n} w(Z_i)\xi_i \geq 0\right\}\right) \\
&\geq \sum_{j=1}^{n}\frac{\mathbb{P}\left(E_j\right)}{2} \\
&\geq \frac{1}{2}\mathbb{P}\left(\max_{j\in[n]}\sum_{i=1}^{j} w(Z_i)\xi_i \geq t\right),
\end{aligned}$$

where the first equality holds because for $i \neq j$, $E_i \cap E_j = \emptyset$, the first inequality follows from (E.5) and the last inequality comes from the union bound.

Step 2: For any differentiable f and random variable X,

$$\begin{aligned}\mathbb{E}[f(X)\mathbb{1}_{\{X\geq 0\}}] &= \mathbb{E}\left[\left(f(0) + \int_0^X f'(t)dt\right)\right] \\ &= f(0)\mathbb{E}[\mathbb{1}_{\{X\geq 0\}}] + \mathbb{E}\left[\int_0^\infty f'(t)\mathbb{1}_{\{X\geq t\}}dt\right] \\ &= f(0)\mathbb{P}(X \geq 0) + \int_0^\infty f'(t)\mathbb{P}(X \geq t)dt, \end{aligned} \tag{E.7}$$

where the last equality follows from Fubini's theorem. Putting it altogether, we have

$$\begin{aligned}&\mathbb{E}_{\mathbb{P},\mathbb{R}}\left[\exp\left(\max_{j\in[n]}\frac{\lambda}{n}\sum_{i=1}^{j} w(Z_i)\xi_i\right)\right] \\ =&1 + \lambda\int_0^\infty \exp(\lambda t)\,\mathbb{P}\left(\max_{j\in[n]}\frac{1}{n}\sum_{i=1}^{j} w(Z_i)\xi_i \geq t\right)dt \\ \leq&1 + 2\lambda\int_0^\infty \exp(\lambda t)\,\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n} w(Z_i)\xi_i \geq t\right)dt \\ =&1 + 2\mathbb{E}_{\mathbb{P},\mathbb{R}}\left[\left(\exp\left(\frac{\lambda}{n}\sum_{i=1}^{n} w(Z_i)\xi_i\right) - 1\right)\mathbb{1}_{\{\sum_{i=1}^n w(Z_i)\xi_i\geq 0\}}\right] \\ =&1 + 2\mathbb{E}_{\mathbb{P},\mathbb{R}}\left[\exp\left(\frac{\lambda}{n}\sum_{i=1}^{n} w(Z_i)\xi_i\right)\right] - 2\mathbb{P}\left(\sum_{i=1}^{n} w(Z_i)\xi_i \geq 0\right) \\ \leq&2\mathbb{E}_{\mathbb{P},\mathbb{R}}\left[\exp\left(\frac{\lambda}{n}\sum_{i=1}^{n} w(Z_i)\xi_i\right)\right],\end{aligned}$$

where the first inequality follows from (E.4), the second equality uses (E.7) with $f(t) = e^{\lambda t}$ and $X = \max_{j\in[n]}\frac{1}{n}\sum_{i=1}^{j} w(Z_i)\xi_i$, and the last inequality follows from (E.6). Finally, if $\frac{1}{n}\sum_{i=1}^{n} w(Z_i)\xi_i$ is mean zero $\frac{\gamma^2}{n}$-subGaussian, then using the definition of a subGaussian random variable, we obtain

$$2\mathbb{E}_{\mathbb{P},\mathbb{R}}\left[\exp\left(\frac{\lambda}{n}\sum_{i=1}^{n} w(Z_i)\xi_i\right)\right] \leq 2\exp\left(\frac{\lambda^2\gamma^2}{2n}\right).$$

□

E.2.2 Proof of Theorem 6.1

Theorem 6.1

Given a hypothesis class $\mathbb{F}$, any loss function $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$, and n samples $\{Z_i\}_{i=1}^n$, we have that with probability at least $1-\delta$,

$$e_n(\mathbb{F}, \ell) \leq 2\mathcal{R}(n, \mathbb{F}) + \sqrt{\frac{\log(\frac{1}{\delta})}{2n}}.$$

Proof. We first analyze the sensitivity of the sup-norm of the CDF estimator over $\mathbb{F}$ and $\mathbb{G}(1)$. For a given two sets $\{x_i, y_i\}_{i=1}^n$ and $\{x_i', y_i'\}_{i=1}^n$, which just differ in j'th entry, let

$$\widehat{F}_1(r; f) := \frac{1}{n}\sum_{i=1}^n \mathbb{1}_{\{\ell(y_i, f(x_i)) \leq r\}} \text{ and, } \widehat{F}_2(r; f) := \frac{1}{n}\sum_{i=1}^n \mathbb{1}_{\{\ell(y_i', f(x_i')) \leq r\}}$$

Then, if $\sup_{f\in\mathbb{F}} \sup_{r\in\mathbb{R}} \left|\widehat{F}_1(r; f) - F(r; f)\right| \geq \sup_{f\in\mathbb{F}} \sup_{r\in\mathbb{R}} \left|\widehat{F}_2(r; f) - F(r; f)\right|$, we have

$$\begin{aligned}
&\sup_{f\in\mathbb{F}} \sup_{r\in\mathbb{R}} \left|\widehat{F}_1(r; f) - F(r; f)\right| - \sup_{f\in\mathbb{F}} \sup_{r\in\mathbb{R}} \left|\widehat{F}_2(r; f) - F(r; f)\right| \\
&= \sup_{f\in\mathbb{F}} \sup_{r\in\mathbb{R}} \left|\frac{1}{n}\sum_{i=1}^n \mathbb{1}_{\{\ell(X_i, y_i, f(x_i)) \leq r\}} - F(r; f)\right| - \sup_{f\in\mathbb{F}} \sup_{r\in\mathbb{R}} \left|\frac{1}{n}\sum_{i=1}^n \mathbb{1}_{\{\ell(x_i, y_i', f(x_i')) \leq r\}} - F(r; f)\right| \\
&= \sup_{f\in\mathbb{F}} \sup_{r\in\mathbb{R}} \left|\frac{1}{n}\sum_{i=1}^n \mathbb{1}_{\{\ell(x_i', y_i', f(x_i')) \leq r\}} + \frac{1}{n}\left(\mathbb{1}_{\{\ell(x_j, y_j, f(x_j)) \leq r\}} - \mathbb{1}_{\{\ell(x_j', y_j', f(x_j')) \leq r\}}\right) - F(r; f)\right| \\
&\qquad\qquad - \sup_{f\in\mathbb{F}} \sup_{r\in\mathbb{R}} \left|\frac{1}{n}\sum_{i=1}^n \mathbb{1}_{\{\ell(x_i', y_i', f(x_i')) \leq r\}} - F(r; f)\right| \\
&= \sup_{f\in\mathbb{F}} \sup_{r\in\mathbb{R}} \frac{1}{n}\left|\mathbb{1}_{\{\ell(x_j, y_j, f(x_j)) \leq r\}} - \mathbb{1}_{\{\ell(x_j', y_j', f(x_j')) \leq r\}}\right| \leq \frac{1}{n}.
\end{aligned}$$

$$\begin{aligned}
&\left|\sup_{f\in\mathbb{F}} \sup_{r\in\mathbb{R}} \left|\widehat{F}_1(r; f) - F(r; f)\right| - \sup_{f\in\mathbb{F}} \sup_{r\in\mathbb{R}} \left|\widehat{F}_2(r; f) - F(r; f)\right|\right| \\
&\qquad \leq \sup_{f\in\mathbb{F}} \sup_{r\in\mathbb{R}} \frac{1}{n}\left|\mathbb{1}_{\{\ell(x_j, y_j, f(x_j)) \leq r\}} - \mathbb{1}_{\{\ell(x_j', y_j', f(x_j')) \leq r\}}\right| = \frac{1}{n}
\end{aligned}$$

This bound holds no matter what j and what data set we choose. Using bounded difference inequality, i.e., McDiarmid's inequality (Boucheron et al., 2013), we have,

$$\mathbb{P}\left(\sup_{f\in\mathbb{F}} \sup_{r\in\mathbb{R}} \left|\widehat{F}(r; f) - F(r; f)\right| - \mathbb{E}_{\mathbb{P}}\left[\sup_{f\in\mathbb{F}} \sup_{r\in\mathbb{R}} \left|\widehat{F}(r; f) - F(r; f)\right|\right] \leq \sqrt{\frac{\log(\frac{1}{\delta})}{2n}}\right) \tag{E.8}$$

with probability at least $1-\delta$. Using a ghost sample set $\{X_i', Y_i'\}_{i=1}^n$ we have

$$
\begin{aligned}
&\mathbb{E}_{\mathbb{P}}\left[\sup_{f\in\mathbb{F}}\sup_{r\in\mathbb{R}}\left|\widehat{F}(r;f)-F(r;f)\right|\right] \\
=&\mathbb{E}_{\mathbb{P}}\left[\sup_{f\in\mathbb{F}}\sup_{r\in\mathbb{R}}\left|\frac{1}{n}\sum_{i=1}^n \mathbb{1}_{\{\ell(X_i,Y_i,f(X_i))\le r\}}-\mathbb{E}_{\mathbb{P}}\left[\frac{1}{n}\sum_{i=1}^n \mathbb{1}_{\{\ell(X_i',Y_i',f(X_i'))\le r\}}\right]\right|\right] \\
=&\mathbb{E}_{\mathbb{P}}\left[\sup_{f\in\mathbb{F}}\sup_{g\in\mathbb{G}(1)}\left|\frac{1}{n}\sum_{i=1}^n g(\ell(X_i,Y_i,f(X_i)))-\mathbb{E}_{\mathbb{P}}\left[\frac{1}{n}\sum_{i=1}^n g(\ell(X_i',Y_i',f(X_i')))\right]\right|\right] \\
=&\mathbb{E}_{\mathbb{P}}\left[\sup_{f\in\mathbb{F}}\sup_{g\in\mathbb{G}(1)}\left|\frac{1}{n}\mathbb{E}_{\mathbb{P}}\left[\sum_{i=1}^n g(\ell(X_i,Y_i,f(X_i)))-\sum_{i=1}^n g(\ell(X_i',Y_i',f(X_i')))\middle|\sigma\left(\{X_i,Y_i\}_{i=1}^n\right)\right]\right|\right] \\
\le&\mathbb{E}_{\mathbb{P}}\left[\sup_{f\in\mathbb{F}}\sup_{g\in\mathbb{G}(1)}\left|\frac{1}{n}\sum_{i=1}^n\left(g(\ell(X_i,Y_i,f(X_i)))-g(\ell(X_i',Y_i',f(X_i')))\right)\right|\right] \\
\le&\mathbb{E}_{\mathbb{P},\mathbb{R}}\left[\sup_{f\in\mathbb{F}}\sup_{g\in\mathbb{G}(1)}\left|\frac{1}{n}\sum_{i=1}^n\xi_i\left(g(\ell(X_i,Y_i,f(X_i)))-g(\ell(X_i',Y_i',f(X_i')))\right)\right|\right] \\
\le&2\mathbb{E}_{\mathbb{P},\mathbb{R}}\left[\sup_{f\in\mathbb{F}}\sup_{g\in\mathbb{G}(1)}\left|\frac{1}{n}\sum_{i=1}^n\xi_i g(\ell(X_i,Y_i,f(X_i)))\right|\right]=2\mathcal{R}(n,\mathbb{F}).
\end{aligned}
\tag{E.9}
$$

Putting (E.8) and (E.9) together concludes the proof. □

E.2.3 Permuation Complexity

We note that this proof, along with many other proofs for Section 6.4, is based on the machinery and techniques in Massart's finite class Lemma (Massart, 2000) and DKW inequality in (Devroye et al., 2013).

Theorem 6.2

For any hypothesis class $\mathbb{F}$ and loss function $\ell : \mathcal{Y}\times\mathcal{Y}\to\mathbb{R}$, we have that

$$\mathcal{R}(n,\mathbb{F})\le\sqrt{\frac{\log(4\mathcal{N}_\Pi(n,\mathbb{F}_\ell,\mathbb{P}))}{2n}}.$$

Proof. For a positive λ, we have,

$$
\begin{aligned}
\exp\left(\lambda\mathcal{R}(n,\mathbb{F})\right)&=\exp\left(\frac{\lambda}{n}\mathbb{E}_{\mathbb{P},\mathbb{R}}\left[\sup_{f\in\mathbb{F}}\sup_{g\in\mathbb{G}(1)}\left|\sum_{i=1}^n\xi_i g(\ell(X_i,Y_i,f(X_i)))\right|\right]\right) \\
&\le\mathbb{E}_{\mathbb{P},\mathbb{R}}\left[\exp\left(\frac{\lambda}{n}\sup_{f\in\mathbb{F}}\sup_{g\in\mathbb{G}(1)}\left|\sum_{i=1}^n\xi_i g(\ell(X_i,Y_i,f(X_i)))\right|\right)\right]
\end{aligned}
$$

$$\leq \mathbb{E}_{\mathbb{P},\mathbb{R}}\left[\mathbb{E}_{\mathbb{P},\mathbb{R}}\left[\exp\left(\frac{\lambda}{n}\sup_{f\in\mathbb{F}}\sup_{g\in\mathbb{G}(1)}\left|\sum_{i=1}^{n}\xi_i g(\ell(X_i,Y_i,f(X_i)))\right|\right)\Big|\sigma\left(\{X_i,Y_i\}_{i=1}^n\right)\right]\right] \tag{E.10}$$

For any $f \in \mathbb{F}$, let π_f denote a permutation such that $\ell(X_{\pi_f(i)}, Y_{\pi_f(i)}, f(X_{\pi_f(i)})) \leq \ell(X_{\pi_f(j)}, Y_{\pi_f(j)}, f(X_{\pi_f(j)}))$ for any $i, j \in [n]$ and $i \leq j$. Therefore we have,

$$\sup_{g\in\mathbb{G}(1)}\left|\sum_{i=1}^{n}\xi_i g(\ell(X_i,Y_i,f(X_i)))\right| = \sup_{g\in\mathbb{G}(1)}\left|\sum_{i=1}^{n}\xi_{\pi_f(i)} g(\ell(X_{\pi_f(i)},Y_{\pi_f(i)},f(X_{\pi_f(i)})))\right|$$

Consider a function $g(r') = \mathbb{1}_{\{r'\leq r\}}$. For such a function, $\sum_{i=1}^n \xi_{\pi_f(i)} g(\ell(X_{\pi_f(i)}, Y_{\pi_f(i)}, f(X_{\pi_f(i)})))$ is equal to,

- 0 if $r < \min_i\{\ell(X_{\pi_f(i)}, Y_{\pi_f(i)}, f(X_{\pi_f(i)}))\}_i^n$,
- $\sum_i^j \xi_{\pi_f(i)}$ when $\ell(X_{\pi_f(j)}, Y_{\pi_f(j)}, f(X_{\pi_f(j)})) \leq r < \ell(X_{\pi_f(j+1)}, Y_{\pi_f(j+1)}, f(X_{\pi_f(j+1)}))$ for a $j \in \{1,\ldots,n-1\}$,
- $\sum_i^n \xi_{\pi_f(i)}$ otherwise.

Using this property, we have,

$$\sup_{g\in\mathbb{G}(1)}\left|\sum_{i=1}^{n}\xi_{\pi_f(i)} g(\ell(X_{\pi_f(i)},Y_{\pi_f(i)},f(X_{\pi_f(i)})))\right| = \max_j\left|\sum_{i}^{j}\xi_{\pi_f(i)}\right| \tag{E.11}$$

Using this equality, we can further extend the Eq. E.10,

$$\begin{aligned}
\exp\left(\lambda\mathcal{R}(n,\mathbb{F})\right) &\leq \mathbb{E}_{\mathbb{P},\mathbb{R}}\left[\mathbb{E}_{\mathbb{P},\mathbb{R}}\left[\exp\left(\frac{\lambda}{n}\sup_{f\in\mathbb{F}}\max_j\left|\sum_{i}^{j}\xi_{\pi_f(i)}\right|\right)\Big|\sigma\left(\{X_i,Y_i\}_{i=1}^n\right)\right]\right] \\
&\leq \mathbb{E}_{\mathbb{P},\mathbb{R}}\left[\mathbb{E}_{\mathbb{P},\mathbb{R}}\left[\mathcal{N}_\Pi(\mathbb{F}_\ell,\{X_i,Y_i\}_{i=1}^n)\exp\left(\frac{\lambda}{n}\max_j\left|\sum_{i}^{j}\xi_i\right|\right)\Big|\sigma\left(\{X_i,Y_i\}_{i=1}^n\right)\right]\right] \\
&\leq \mathcal{N}_\Pi(n,\mathbb{F}_\ell,\mathbb{P})\mathbb{E}_{\mathbb{P},\mathbb{R}}\left[\exp\left(\frac{\lambda}{n}\max_j\left|\sum_{i}^{j}\xi_i\right|\right)\right],
\end{aligned} \tag{E.12}$$

where the second inequality follows from the fact that effectively, there are at most $\mathcal{N}_\Pi(\mathbb{F}_\ell, \{X_i, Y_i\}_{i=1}^n)$ number of π_f's. Using the same derivation for (E.3) (Lemma E.1), we have

$$\exp\left(\lambda\mathcal{R}(n,\mathbb{F})\right) \leq 2\mathcal{N}_\Pi(n,\mathbb{F}_\ell,\mathbb{P})\mathbb{E}_{\mathbb{P},\mathbb{R}}\left[\max_j\left(\exp\left(\frac{\lambda}{n}\sum_{i}^{j}\xi_i\right)\mathbb{1}_{\{\sum_i^j\xi_i\geq 0\}}\right)\right].$$

By Lemma E.2, we have

$$\exp\left(\lambda \mathcal{R}(n, \mathbb{F})\right) \leq 4\mathcal{N}_{\Pi}(n, \mathbb{F}_{\ell}, \mathbb{P}) \exp\left(\frac{\lambda^2}{2n}\right)$$

Now, taking the log from both sides, and dividing by λ, we have $\mathcal{R}(n, \mathbb{F}) \leq \frac{\log(4\mathcal{N}_{\Pi}(n, \mathbb{F}_{\ell}, \mathbb{P}))}{\lambda} + \frac{\lambda}{2n}$. Choosing $\lambda = \sqrt{2n \log(2\mathcal{N}_{\Pi}(n, \mathbb{F}_{\ell}, \mathbb{P}))}$, we obtain the final result

$$\mathcal{R}(n, \mathbb{F}) \leq \sqrt{\frac{\log(4\mathcal{N}_{\Pi}(n, \mathbb{F}_{\ell}, \mathbb{P}))}{2n}}.$$

□

Corollary 6.1. *For a finite hypothesis class $\mathbb{F}$,*

$$\mathcal{R}(n, \mathbb{F}) \leq \sqrt{\frac{\log(4|\mathbb{F}|)}{2n}}.$$

Proof. The result follows since when $|\mathbb{F}| < \infty$, $\mathcal{N}_{\Pi}(n, \mathbb{F}_{\ell}, \mathbb{P}) \leq |\mathbb{F}|$, i.e., we need at most one permutation function to sort the losses for each $f \in \mathbb{F}$.

□

E.3 Proof of Results in Section 6.5

Theorem 6.3

For a hypothesis class $\mathbb{F}$, a bounded loss function ℓ, and $\delta \in (0, 1]$, if $\mathbb{P}(e_n(\mathbb{F}, \ell) \leq \epsilon) \geq 1-\delta$, then with probability $1 - \delta$, for all $\rho \in \mathbb{T}$, we have

$$\sup_{f \in \mathbb{F}} |\rho(F(\cdot; f)) - \rho(\widehat{F}(\cdot; f))| \leq L(\rho, 1, \mathcal{L}_\infty)\epsilon,$$

where $\mathbb{T}$ is the set of $L(\cdot, 1, \mathcal{L}_\infty)$ Hölder risk functionals on the space of bounded random variables.

Proof. For all $\rho \in \mathbb{T}$, since ρ is $L(\rho, 1, \mathcal{L}_\infty)$ Hölder, we have that $|\rho(F(\cdot; f)) - \rho(\widehat{F}(\cdot; f))| \leq L(\rho, 1, \mathcal{L}_\infty)\|F - \widehat{F}\|_\infty$. The desired result then follows. □

The error of risk assessment can be bounded using distances other than the sup-norm, such as the Wasserstein distance. For two random variables U and U' with bounded support $[0, D]$, the dual form of the Wasserstein distance $W_1(F_U, F_{U'})$ is given by (Vallender, 1974),

$$W_1(F_U, F_{U'}) = \int_0^D |F_U(t) - F_{U'}(t)|dt \leq D\|F_U - F_U\|_\infty.$$

This inequality suggests the following corollary.

Corollary E.1. *Under the setting of Theorem 6.3 where the loss has support $[0, D]$, with probability $1 - \delta$, for all $\rho \in \mathbb{T}$, we have*

$$\sup_{f \in \mathbb{F}} |\rho(F(\cdot; f)) - \rho(\widehat{F}(\cdot; f))| \leq L(\rho, p, W_1)D^p\epsilon^p,$$

where $\mathbb{T}$ is the set of $L(\cdot, p, W_1)$ Hölder risk functionals on the space of bounded random variables.

Proof. For all $\rho \in \mathbb{T}$, since ρ is $L(\rho, p, W_1)$ Hölder, we have that $|\rho(F(\cdot; f)) - \rho(\widehat{F}(\cdot; f))| \leq L(\rho, p, W_1)W_1(F, \widehat{F})^p \leq L(\rho, p, W_1)D^p\|F - \widehat{F}\|_\infty^p$. The desired result then follows. □

Corollary 6.2. *For a hypothesis class $\mathbb{F}$, a bounded loss function $\ell : \mathcal{Y} \times \mathcal{Y} \to [0, D]$, and $\delta \in (0, 1]$, if $\mathbb{P}(e_n(\mathbb{F}, \ell) \leq \epsilon) \geq 1 - \delta$, then with probability $1 - \delta$, we have*

$$\sup_{f \in \mathbb{F}} |\rho(F(\cdot; f)) - \rho(\widehat{F}(\cdot; f))| \leq L\epsilon,$$

where ρ is a distortion risk with $\frac{L}{D}$-Lipschitz distortion function.

Proof. As is shown in Huang et al. (2021, Lemma 4.1), ρ is L-Lipschitz. Thus, directly applying Theorem 6.3 concludes the proof. □

E.4 Proofs for results in Section 6.6

E.4.1 Proof of Lemma 6.3

Before proving Lemma 6.3, we first present two auxiliary lemmas.

Lemma E.3

For any continuous function $g, h : \mathbb{R}^d \to \mathbb{R}$, $\min(g, h)$ and $\max(g, h)$ are continuous. Similarly, for any Lipschitz continuous function $g, h : \mathbb{R}^d \to \mathbb{R}$, $\min(g, h)$ and $\max(g, h)$ are Lipschitz continuous.

Proof. Denote $f_{\min} = \min(g, h)$ and $f_{\max} = \max(g, h)$.
Continuity: Continuity: We first consider the case when both g and h are continuous. Define $\overline{\Theta} = \{\theta \in \mathbb{R}^d : g(\theta) = h(\theta)\}$. For $\theta \notin \overline{\Theta}$, $f_{\min}$ and $f_{\max}$ are continuous since g, h are continuous. Consider $\theta \in \overline{\Theta}$. For every $\epsilon > 0$, there exists $\delta_g, \delta_h > 0$ such that for all $\theta' \in \mathbb{R}^d$, $\|\theta - \theta'\| \leq \epsilon \implies |g(\theta) - g(\theta')| \leq \delta_g, |h(\theta) - h(\theta')| \leq \delta_h$. In addition, $f_{\min}(\theta) = g(\theta) = h(\theta)$, $f_{\max}(\theta) = g(\theta) = h(\theta)$ and $f_{\min}(\theta')$, $f_{\max}(\theta')$ can be either $g(\theta')$ or $h(\theta')$. Combining both facts gives us that $|f_{\min}(\theta) - f_{\min}(\theta')| \leq \max(\delta_g, \delta_h)$ and $|f_{\max}(\theta) - f_{\max}(\theta')| \leq \max(\delta_g, \delta_h)$.

Lipschitz Continuity: We next work with the case where both g and h are Lipschitz continuous. When $|f_{\max}(\theta) - f_{\max}(\theta')| = |g(\theta) - h(\theta')|$, we have the following two cases:

1. $g(\theta) > h(\theta')$: $|g(\theta) - h(\theta')| = g(\theta) - h(\theta') \leq g(\theta) - g(\theta')$, since $f_{\max}(\theta') = h(\theta')$.
2. $g(\theta) \leq h(\theta')$: $|g(\theta) - h(\theta')| = h(\theta') - g(\theta) \leq h(\theta') - h(\theta)$, since $f_{\max}(\theta) = g(\theta)$.

Since both h and g are Lipschitz continuous, we obtain that

$$|f_{\max}(\theta) - f_{\max}(\theta')| \leq \max_{f \in \{g,h\}} |f(\theta) - f(\theta')| \lesssim \|\theta - \theta'\|,$$

showing that $f_{\max}$ is Lipschitz continuous. The proof completes with the fact that $\min(f, g) = -\max(-f, -g)$. □

Lemma E.4

If $\{\ell_\theta(x_j, y_j)\}_{j=1}^n$ are Lipschitz continuous in θ, then for all $i \in [n]$, $\ell_\theta(\pi_\theta(i))$, i.e., the i-th smallest loss evaluated using data points $\{x_j, y_j\}_{j=1}^n$, is Lipschitz continuous in θ.

Proof. The key observation is that the i-th smallest loss can be defined as $\ell_\theta(\pi_\theta(i)) = \min\{\max\{\ell_\theta(j) : j \in J\} : J \subseteq [n], |J| = i\}$. Since each $\ell_\theta(j)$ is Lipschitz continuous in θ and that for $j' \in J$, $\max\{\ell_\theta(j) : j \in J\} = \max\{\ell_\theta(j'), \max\{\ell_\theta(j) : j \in J \setminus \{j'\}\}\}$, by Lemma E.3, we have $\max\{\ell_\theta(j) : j \in J\}$ to be Lipschitz continuous in θ. Similarly, since $\max\{\ell_\theta(j) : j \in J\}$ is Lipschitz continuous in θ, we have $\min\{\max\{\ell_\theta(j) : j \in J\} : J \subseteq [n], |J| = i\}$ to be Lipschitz continuous. □

Lemma 6.3

If $\{\ell_\theta(z_i)\}_{i=1}^n$ are Lipschitz continuous in $\theta \in \Theta$, then for all $i \in [n]$, $\ell_\theta(\pi_\theta(i))$, i.e., the i-th smallest loss, is Lipschitz continuous in θ and $\rho(\widehat{F}_\theta)$ is differentiable in θ almost everywhere.

Proof. Using Lemma E.4, we obtain that $\ell_\theta(\pi_\theta(i))$ is Lipschitz in $\theta \in \Theta$. Following from a classical result of Rademacher (Rockafellar and Wets, 2009, Theorem 9.60), i.e., a locally Lipschitz function is differentiable almost everywhere, we have that $\rho(\widehat{F}_\theta)$ is differentiable almost everywhere. □

E.4.2 Local Convergence

The proofs for Corollary 6.3 are standard (Bottou et al., 2018), which we provide for completeness.

Corollary 6.3. *If $\{\ell_\theta(z_i)\}_{i=1}^n$ are Lipschitz continuous and $\rho(\widehat{F}_\theta)$ is β-smooth in θ, then the following holds almost surely when the learning rate in (6.7) is $\eta = \frac{1}{\beta\sqrt{T}}$:*

$$\frac{1}{T}\sum_{t=1}^{T} \mathbb{E}\left[\|\nabla_\theta \rho(\widehat{F}_{\theta_t})\|^2\right] \leq \frac{2\beta}{\sqrt{T}}\left(\rho(\widehat{F}_{\theta_1}) - \rho(\widehat{F}_{\theta_\star}) + \frac{1}{2\beta}\right).$$

Proof. For notation simplicity, we use $h(\theta)$ to denote $\rho(\widehat{F}_\theta)$ and g_t to denote $\nabla_\theta \rho(\widehat{F}_\theta)$ when $\theta = \theta_t$. Since $h(\theta)$ is differentiable almost everywhere, following (6.7), the sequence $\{h(\theta_t)\}_{t=1}^T$ will be differentiable almost surely. Since h is β-smooth, we have that

$$h(\theta_{t+1}) - h(\theta_t) \leq \nabla_\theta h(\theta_t)^\top(\theta_{t+1} - \theta_t) + \frac{\beta}{2}\|\theta_{t+1} - \theta_t\|^2.$$

We denote the filtration for the stochastic process $\{\theta_t\}_{t=1}^T$ to be $\{\mathcal{F}_t\}_{t=1}^T$. Extending the above inequality, we have

$$h(\theta_{t+1}) - h(\theta_t) \leq \nabla_\theta h(\theta_t)^\top(-\eta(g_t + w_t)) + \frac{\beta}{2}\| - \eta(g_t + w_t)\|^2,$$

which suggests that

$$\mathbb{E}[h(\theta_{t+1}) - h(\theta_t)] \leq -\eta \nabla_\theta h(\theta_t)^\top g_t + \eta^2 \frac{\beta}{2}(\|g_t\|^2 + d\sigma_w^2),$$

where $\sigma_w^2 = 1/d$ is the variance of w_t. Therefore, for the conditional expectation, we have

$$\begin{aligned}\mathbb{E}\left[h(\theta_{t+1}) - h(\theta_t)\middle|\mathcal{F}_t\right] &\leq \mathbb{E}\left[-\eta \nabla_\theta h(\theta_t)^\top g_t + \eta^2 \frac{\beta}{2}(\|g_t\|^2 + 1)\middle|\mathcal{F}_k\right] \\ &= -(\eta - \eta^2\frac{\beta}{2})\mathbb{E}\left[\|\nabla_\theta h(\theta_t)\|^2\middle|\mathcal{F}_t\right] + \eta^2\frac{\beta}{2}.\end{aligned}$$

Using the telescoping sum and law of total expectation, we obtain

$$\mathbb{E}\left[h(\theta_T) - h(\theta_1)\middle|\mathcal{F}_1\right] = \mathbb{E}\left[\sum_{t=1}^{T} h(\theta_1) - h(\theta_{t-1})\middle|\mathcal{F}_1\right]$$

$$\leq (\eta - \eta^2 \frac{\beta}{2}) \mathbb{E}\left[\sum_{t=1}^{T} -\|\nabla_\theta h(\theta_t)\|^2 \Big| \mathcal{F}_1\right] + T\eta^2 \frac{\beta}{2}. \tag{E.13}$$

Use the fact that $h(\theta_\star) \leq h(\theta_T)$, we have $\mathbb{E}\left[h(\theta_1) - h(\theta_T)\right] \leq h(\theta_1) - h(\theta_\star)$, which implies

$$(\eta - \eta^2 \frac{\beta}{2}) \mathbb{E}\left[\sum_{t=1}^{T} \|\nabla_\theta h(\theta_t)\|^2\right] \leq h(\theta_1) - h(\theta_\star) + T\eta^2 \frac{\beta}{2}.$$

Plugging the learning rate $\eta = \frac{1}{\beta\sqrt{T}}$, we have $\eta - \eta^2 \frac{\beta}{2} = \frac{1}{\beta\sqrt{T}} - \frac{1}{2\beta T} \geq \frac{1}{\beta\sqrt{T}} - \frac{1}{2\beta\sqrt{T}} \geq \frac{1}{2\beta\sqrt{T}} > 0$. Therefore we have

$$\frac{1}{2\beta\sqrt{T}} \mathbb{E}\left[\sum_{t=1}^{T} \|\nabla_\theta h(\theta_t)\|^2\right] \leq h(\theta_1) - h(\theta_\star) + \frac{1}{2\beta}.$$

Rearranging the above inequality gives the result. □

E.5 Additional Experimental Details

E.5.1 Risk Assessment on ImageNet Models

Figure E.1 shows the CVaR of models presented in Table 6.1 under different α's. We note that CVaR_α is the expected value above the top 100α percent losses.

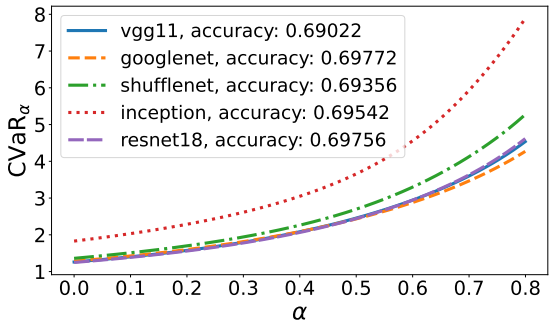


Figure E.1: $\text{CVaR}_\alpha(\ell_f(Z))$ across different α's where ℓ_f is the cross-entropy loss evaluated on the ImageNet validation dataset.

E.5.2 Empirical Distortion Risk Minimization

Toy Example The data used in this experiment is generated using the make_blobs function from sklearn.datasets with the following parameters: n_samples = [1000, 50], centers = [[0.0, 0.0], [1.0, 1.0]], cluster_std = [1.5, 0.5], random_state = 0, shuffle = False.

CIFAR-10 For completeness, we report the average test accuracy for the VGG-16 models obtained through minimizing the empirical risks for expected loss, $\text{CVaR}_{.05}$, $\text{CVaR}_{.7}$ and $\text{HRM}_{.3,.4}$. The models are trained over 150 epochs and the learning rate is chosen to be 0.005. The accuracy is 55.2%, 13.2%, 51.3%, and 53.9% respectively. We note that the goal of this experiment is to illustrate the efficacy of our proposed optimization procedure for minimizing distortion risks instead of arguing for the usage of a particular risk functional for finding a model with the highest test accuracy.

Appendix F

Median Optimal Treatment Regimes: Additional Details

F.1 Proofs

F.1.1 Proofs in Section 7.4

Proof of Proposition 7.1. Pick $x \in \mathcal{X}$. For all $a \in \{0,1\}$, we have $|\mu_a(x) - m_a(x)| \leq \sigma_a(x)$. If $\mu_1(x) > \mu_0(x)$, then $m_1(x) \geq \mu_1(x) - \sigma_1(x) > \mu_0(x) + \sigma_0(x) \geq m_0(x)$, where the second inequality holds because of the assumption that $|\mu_1(x) - \mu_0(x)| > \sigma_1(x) + \sigma_0(x)$. Similarly, $m_1(x) > m_0(x)$ implies that $\mu_1(x) \geq m_1(x) - \sigma_1(x) > m_0(x) + \sigma_0(x) \geq \mu_0(x)$. Thus, we have that $\mu_1(x) > \mu_0(x) \Leftrightarrow m_1(x) > m_0(x)$, which completes the proof. □

Proof of Proposition 7.2. By definition, the marginal median optimal policy d^*_{MME} satisfies that for all $d \in \mathcal{D}$, $m(Y^{d^*_{\text{MME}}}) \geq m(Y^d)$. We use a tuple (a_1, a_0) to represent a policy d where $a_i = d(i)$, which gives us the following table:

(a_0, a_1)	$m(Y^d)$
$(0,1)$	$\frac{\sigma_1(1)\mu_0(0)+\sigma_0(0)\mu_1(1)}{\sigma_0(0)+\sigma_1(1)}$
$(1,1)$	$\frac{\sigma_1(1)\mu_1(0)+\sigma_1(0)\mu_1(1)}{\sigma_1(0)+\sigma_1(1)}$
$(0,0)$	$\frac{\sigma_0(1)\mu_0(0)+\sigma_0(0)\mu_0(1)}{\sigma_0(0)+\sigma_0(1)}$
$(1,0)$	$\frac{\sigma_0(1)\mu_1(0)+\sigma_1(0)\mu_0(1)}{\sigma_1(0)+\sigma_0(1)}$

Table F.1: Marginal medians of all possible policies in the setting described in Section 7.4.2.

Given $\sigma_1(1) > \sigma_0(1)$ and $\mu_1(1) > \mu_0(1)$ and $\frac{\mu_1(1)}{\mu_0(1)} < \frac{\sigma_1(1)}{\sigma_0(1)}$, when we pick $\{\mu_a(0), \sigma_a(0)\}_{a=0}^1$

such that for $a = 0$ and $a = 1$,

$$(\sigma_1(1) - \sigma_0(1))\,\mu_a(0) + (\mu_1(1) - \mu_0(1))\,\sigma_a(0) > \sigma_1(1)\mu_0(1) - \sigma_0(1)\mu_1(1), \tag{F.1}$$

the optimal policy d^*_{MME} will assign the decision for $x = 1$ according to $d^*_{\text{MME}}(1) = \mathbb{1}\{\mu_1(1) > \mu_0(1)\}$ since the conditions have ensured that $m(Y^{(1,1)}) > m(Y^{(1,0)})$ and $m(Y^{(0,1)}) > m(Y^{(0,0)})$. On the other hand, if (F.1) does not hold for $a = 0$ and $a = 1$, then the marginal median optimal policy $d^*_{\text{MME}}(1) = \mathbb{1}\{\mu_1(1) \leq \mu_0(1)\}$. □

F.1.2 Proofs in Section 7.5

Derivation of the influence function under discrete covariates and continuous outcome. Let $\text{IF}(\cdot)$ denote the operator that returns the influence function of an input functional. In this analysis, we assume X to be discrete and Y to be continuous. Let $p(x)$ denote the probability mass function of X. We start with the chain rule of $\text{IF}(\cdot)$.

$$\begin{aligned}\phi_d(Z) &= \text{IF}(\psi_d(\mathbb{P})) = \text{IF}\left(\sum_{x\in\mathcal{X}} p(x)m_d(x)\right)\\ &= \sum_{x\in\mathcal{X}} p(x)\Big\{\text{IF}(m_1(x))d(x) + \text{IF}(m_0(x))(1-d(x))\Big\} + \text{IF}(p(x))m_d(x),\end{aligned}$$

where $\text{IF}(p(x)) = \mathbb{1}\{X = x\} - p(x)$. Thus, it suffices to find $\text{IF}(m_a(x))$ where $a \in \{0, 1\}$.

Let δ_z denote the Dirac measure at z. We use the submodel $\mathbb{P}_\epsilon(z) = (1-\epsilon)\mathbb{P}(z) + \delta_{z'}$ for some $z' = (x', a', y')$ to find the influence function, which gives that

$$f_\epsilon(y|a,x) = \frac{(1-\epsilon)f(y|x,a)\mathbb{P}(X = x, A = a) + \epsilon\delta_{z'}}{(1-\epsilon)\mathbb{P}(X = x, A = a) + \epsilon\mathbb{1}\{x = x', a = a'\}},$$

where f_ϵ and f are densities with respect to $\mathbb{P}_\epsilon$ and $\mathbb{P}$ respectively. Let $m_{a,\epsilon}(x)$ denote the median of the distribution $\mathbb{P}_\epsilon(Y|X = x, A = a)$. By the definition of the median, we have that

$$\begin{aligned}F_\epsilon(m_{a,\epsilon}(x)|x,a) &= \int_{y\leq m_{a,\epsilon}(x)} f_\epsilon(y|x,a)dy\\ &= \int_{y\leq m_{a,\epsilon}(x)} \frac{(1-\epsilon)f(y|x,a)\mathbb{P}(X = x, A = a) + \epsilon\delta_{z'}}{(1-\epsilon)\mathbb{P}(X = x, A = a) + \epsilon\mathbb{1}\{x = x', a = a'\}}dy = \frac{1}{2}\end{aligned}$$

When $x \neq x'$ or $a \neq a'$, we have that

$$\mathbb{P}(X = x, A = a)\int_{y\leq m_{a,\epsilon}(x)} f(y|x,a)dy = \frac{\mathbb{P}(X = x, A = a)}{2},$$

which suggests that $m_a(x) = m_{a,\epsilon}(x)$. When $x = x'$ and $a = a'$, we have that

$$\int_{y\leq m_{a,\epsilon}(x)} \frac{(1-\epsilon)f(y|x,a)\mathbb{P}(X = x, A = a) + \epsilon\delta_{y'}}{(1-\epsilon)\mathbb{P}(X = x, A = a) + \epsilon}dy = \frac{1}{2}. \tag{F.2}$$

It has two cases

- When $y' > m_{a,\epsilon}(x)$, (F.2) can be simplified into

$$F(m_{a,\epsilon}(x)|x,a) = \int_{y \le m_{a,\epsilon}(x)} f(y|x,a)dy = \frac{(1-\epsilon)\mathbb{P}(X=x,A=a)+\epsilon}{2(1-\epsilon)\mathbb{P}(X=x,A=a)},$$

and $m_{a,\epsilon}(x) = F^{-1}\left(\frac{(1-\epsilon)\mathbb{P}(X=x,A=a)+\epsilon}{2(1-\epsilon)\mathbb{P}(X=x,A=a)}\middle| x,a\right)$.

- When $y' \le m_{a,\epsilon}(x)$, (F.2) can be simplified into

$$F(m_{a,\epsilon}(x)|x,a) = \int_{y \le m_{a,\epsilon}(x)} f(y|x,a)dy = \frac{(1-\epsilon)\mathbb{P}(X=x,A=a)-\epsilon}{2(1-\epsilon)\mathbb{P}(X=x,A=a)},$$

and $m_{a,\epsilon}(x) = F^{-1}\left(\frac{(1-\epsilon)\mathbb{P}(X=x,A=a)-\epsilon}{2(1-\epsilon)\mathbb{P}(X=x,A=a)}\middle| x,a\right)$.

Putting it altogether, we have that when $x \neq x'$ or $a \neq a'$:

$$\text{IF}(m_a(x)) = \frac{d}{d\epsilon} m_{a,\epsilon}(x)\Big|_{\epsilon=0} = 0 = \mathbb{1}\{x=x'\}\mathbb{1}\{a=a'\}.$$

When $x = x'$ and $a = a'$, we have that

$$\begin{aligned}\text{IF}(m_a(x)) =& \mathbb{1}\{x=x'\}\mathbb{1}\{a=a'\}\frac{d}{d\epsilon} m_{a,\epsilon}(x)\Big|_{\epsilon=0} \\ =& -\frac{\mathbb{1}\{x=x'\}\mathbb{1}\{a=a'\}}{\mathbb{P}(X=x)\mathbb{P}(A=a|X=x)}\frac{\mathbb{1}\{y' \le m_a(x)\}-1/2}{f(m_a(x)|x,a)},\end{aligned}$$

where the last equality holds since for all $m_{a,\epsilon}(x) = F^{-1}(g(\epsilon)|x,a)$,

$$\frac{d}{d\epsilon} m_{a,\epsilon}(x)\Big|_{\epsilon=0} = \left(\frac{1}{f(m_{a,\epsilon}(x)|x,a)}\frac{d}{d\epsilon} g(\epsilon)\right)\Big|_{\epsilon=0},$$

and for example when $g(\epsilon) = \frac{(1-\epsilon)\mathbb{P}(X=x,A=a)+\epsilon}{2(1-\epsilon)\mathbb{P}(X=x,A=a)}$, i.e., when $y' > F^{-1}\left(\frac{(1-\epsilon)\mathbb{P}(X=x,A=a)+\epsilon}{2(1-\epsilon)\mathbb{P}(X=x,A=a)}\middle| x,a\right)$, gives that $\frac{d}{d\epsilon} g(\epsilon) = \frac{1}{2\mathbb{P}(X=x,A=a)} \cdot \frac{d}{d\epsilon}\frac{\epsilon}{1-\epsilon} = \frac{1}{2(1-\epsilon)^2\mathbb{P}(X=x,A=a)}$.

Therefore, we have obtained that

$$\begin{aligned}&\phi_d(Z) = \sum_{x\in\mathcal{X}} p(x)\Big\{\text{IF}\left(m_1(x)\right)d(x) + \text{IF}\left(m_0(x)\right)\left(1-d(x)\right)\Big\} + \text{IF}(p(x))m_d(X) \\ &= \sum_{x\in\mathcal{X}} p(x)\Big\{\frac{\mathbb{1}\{X=x\}\mathbb{1}\{A=1\}}{\mathbb{P}(X=x)\mathbb{P}(A=1|X=x)}\frac{1/2-\mathbb{1}\{Y\le m_1(x)\}}{f(m_1(x)|x,1)}d(x) \\ &+\frac{\mathbb{1}\{X=x\}\mathbb{1}\{A=0\}}{\mathbb{P}(X=x)\mathbb{P}(A=0|X=x)}\frac{1/2-\mathbb{1}\{Y\le m_0(x)\}}{f(m_0(x)|x,0)}(1-d(x))\Big\} + (\mathbb{1}\{X=x\}-p(x))m_d(x) \\ &= \frac{Ad(X)}{\pi_1(X)}\frac{1/2-\mathbb{1}\{Y\le m_1(X)\}}{f_{1,m}(X)} + \frac{(1-A)(1-d(X))}{\pi_0(X)}\frac{1/2-\mathbb{1}\{Y\le m_0(X)\}}{f_{0,m}(X)} + m_d(X) - \psi_d.\end{aligned}$$

where the last equality follows since the indicator function $\mathbb{1}\{X=x\}$ picks out the values $x \in \mathcal{X}$ such that it equals X and $p(x) = \mathbb{P}(X=x)$.

Proof of Lemma 7.1. By definition, since $\int \phi_d(Z;\overline{\mathbb{P}})d\overline{\mathbb{P}} = 0$, we have that

$$
\begin{aligned}
&R_d(\overline{\mathbb{P}},\mathbb{P}) = \psi_d(\overline{\mathbb{P}}) - \psi_d(\mathbb{P}) + \int \phi_d(Z;\overline{\mathbb{P}})d\mathbb{P} \\
&= -\psi_d(\mathbb{P}) + \int_{\mathcal{X}} \frac{\pi_1(x)d(x)}{\overline{\pi}_1(x)} \frac{1/2 - F_{1,\overline{m}_1}(x)}{\overline{f}_{1,\overline{m}}(x)} + \frac{\pi_0(x)(1-d(x))}{\overline{\pi}_0(x)} \frac{1/2 - F_{0,\overline{m}}(x)}{\overline{f}_{0,\overline{m}}(x)} + \overline{m}_d(x)d\mathbb{P}(x) \\
&= \int_{\mathcal{X}} d(x)\Big(\overline{m}_1(x) - m_1(x) - \frac{\pi_1(x)}{\overline{\pi}_1(x)} \frac{F_{1,\overline{m}}(x) - F_{1,m}(x)}{\overline{f}_{1,\overline{m}}(x)}\Big) \\
&\qquad + (1-d(x))\Big(\overline{m}_0(x) - m_0(x) - \frac{\pi_0(x)}{\overline{\pi}_0(x)} \frac{F_{0,\overline{m}}(x) - F_{0,m}(x)}{\overline{f}_{0,\overline{m}}(x)}\Big)d\mathbb{P}(x),
\end{aligned}
$$

where the second equality follows from iterated expectation, i.e., $\mathbb{E}_{\mathbb{P}}[\mathbb{1}\{Y \le \overline{m}_a(X)\}|X = x, A = a] = F_{a,\overline{m}}(x)$, and the third equality follows from rearranging terms and the fact that $F_{a,m}(x) = 1/2$. □

Proof of Corollary 7.1. Recall that for (7.4), we have

$$
\begin{aligned}
R_d(\overline{\mathbb{P}},\mathbb{P}) = \int_{\mathcal{X}} d(x)\Big(\underbrace{\overline{m}_1(x) - m_1(x) - \frac{\pi_1(x)}{\overline{\pi}_1(x)} \overbrace{\frac{F_{1,\overline{m}}(x) - F_{1,m}(x)}{\overline{f}_{1,\overline{m}}(x)}}^{A_2}}_{A_1}\Big) \\
+ (1-d(x))\Big(\underbrace{\overline{m}_0(x) - m_0(x) - \frac{\pi_0(x)}{\overline{\pi}_0(x)} \frac{F_{0,\overline{m}}(x) - F_{0,m}(x)}{\overline{f}_{0,\overline{m}}(x)}}_{A_0}\Big)d\mathbb{P}(x),
\end{aligned}
$$

For $x \in \mathcal{X}$ such that $f_a(y|x)$ is differentiable and L-Lipschitz continuous in y, we can simplify A_2 as:

$$
\begin{aligned}
&A_2 = \frac{F_{1,\overline{m}}(x) - F_{1,m}(x)}{\overline{f}_{1,\overline{m}}(x)} \\
&= \frac{1}{\overline{f}_{1,\overline{m}}(x)}\left(F_{1,m}(x) + f_{1,m}(x)(\overline{m}_1(x) - m_1(x)) + \frac{f_1'(c_1(x)|x)}{2}(\overline{m}_1(x) - m_1(x))^2 - F_{1,m}(x)\right) \\
&= \frac{1}{\overline{f}_{1,\overline{m}}(x)}\left(f_{1,m}(x)(\overline{m}_1(x) - m_1(x)) + \frac{f_{1,c}'(x)}{2}(\overline{m}_1(x) - m_1(x))^2\right),
\end{aligned}
$$

where the second equality follows from Taylor's theorem and $c_1(x)$ is a value in between $\overline{m}_1(x)$ and $m_1(x)$. To simplify the notations, in the third equality (and the following usage of Taylor's theorem), we denote the derivative of the conditional probability density function $f_a(\cdot|X = x)$ at $c_a(x)$ to be $f_{a,c}'(x)$ where $c_a(x)$ is a value between $\overline{m}_a(x)$ and $m_a(x)$ and satisfies that $F_{a,\overline{m}}(x) = F_{a,m}(x) + f_{a,m}(x)(\overline{m}_a(x) - m_a(x)) + f_{a,c}'(x)(\overline{m}_a(x) - m_a(x))^2/2$. This gives that

$$
A_1 = \overline{m}_1(x) - m_1(x) - \frac{\pi_1(x)}{\overline{\pi}_1(x)} \frac{1}{\overline{f}_{1,\overline{m}}(x)}\left(f_{1,m}(x)(\overline{m}_1(x) - m_1(x)) + \frac{f_{1,c}'(x)}{2}(\overline{m}_1(x) - m_1(x))^2\right)
$$

$$= (\overline{m}_1(x) - m_1(x)) \left(1 - \frac{\pi_1(x)}{\overline{\pi}_1(x)} \frac{f_{1,m}(x)}{\overline{f}_{1,\overline{m}}(x)}\right) - \frac{f'_{1,c}(x)\pi_1(x)(\overline{m}_1(x) - m_1(x))^2}{2\overline{\pi}_1(x)\overline{f}_{1,\overline{m}}(x)}.$$

Applying the same logic to A_0, we obtain that

$$\begin{aligned}
&R_d(\overline{\mathbb{P}}, \mathbb{P}) \\
&= \int d(x) \left((\overline{m}_1(x) - m_1(x)) \left(1 - \frac{\pi_1(x)}{\overline{\pi}_1(x)} \frac{f_{1,m}(x)}{\overline{f}_{1,\overline{m}}(x)}\right) - \frac{f'_{1,c}(x)\pi_1(x)(\overline{m}_1(x) - m_1(x))^2}{2\overline{\pi}_1(x)\overline{f}_{1,\overline{m}}(x)} \right) \\
&+(1 - d(x)) \left((\overline{m}_0(x) - m_0(x)) \left(1 - \frac{\pi_0(x)}{\overline{\pi}_0(x)} \frac{f_{0,m}(x)}{\overline{f}_{0,\overline{m}}(x)}\right) - \frac{f'_{0,c}(x)\pi_0(x)(\overline{m}_0(x) - m_0(x))^2}{2\overline{\pi}_0(x)\overline{f}_{0,\overline{m}}(x)} \right) d\mathbb{P}(x) \\
&= \mathbb{P}\Bigg\{ \frac{d}{\overline{\pi}_1 \overline{f}_{1,\overline{m}}} \left((\overline{m}_1 - m_1) \left(\overline{\pi}_1 \overline{f}_{1,\overline{m}} - \pi_1 f_{1,m}\right) - \frac{f'_{1,c}\pi_1(\overline{m}_1 - m_1)^2}{2} \right) \\
&\qquad + \frac{(1-d)}{\overline{\pi}_0 \overline{f}_{0,\overline{m}}} \left((\overline{m}_0 - m_0) \left(\overline{\pi}_0 \overline{f}_{0,\overline{m}} - \pi_0 f_{0,m}\right) - \frac{f'_{0,c}\pi_0(\overline{m}_0 - m_0)^2}{2} \right) \Bigg\} \\
&= \mathbb{P}\Bigg\{ \frac{d\,(\overline{m}_1 - m_1)}{\overline{\pi}_1 \overline{f}_{1,\overline{m}}} \left((\overline{\pi}_1 - \pi_1)\,\overline{f}_{1,\overline{m}} + \left(\overline{f}_{1,\overline{m}} - f_{1,m}\right)\pi_1 - (\overline{m}_1 - m_1)\frac{f'_{1,c}\pi_1}{2} \right) \\
&\qquad + \frac{(1-d)\,(\overline{m}_0 - m_0)}{\overline{\pi}_0 \overline{f}_{0,\overline{m}}} \left((\overline{\pi}_0 - \pi_0)\overline{f}_{0,\overline{m}} + (\overline{f}_{0,\overline{m}} - f_{0,m})\pi_0 - (\overline{m}_0 - m_0)\frac{f'_{0,c}\pi_0}{2} \right) \Bigg\}.
\end{aligned}$$

□

Proof of Corollary 7.2. Consider a parametric submodel $\mathbb{P}_\epsilon$, e.g., the probability density functions p_ϵ and p of $\mathbb{P}_\epsilon$ and $\mathbb{P}$ satisfy that $p_\epsilon(z) = p(z)(1 + \epsilon h(z))$ where $\mathbb{E}[h] = 0$, $\|h\|_\infty < M$ and $\epsilon > 1/M$. To show that ϕ_d is an influence function, it suffices to show that the mean-zero ϕ_d satisfies path-wise differentiability, i.e.,

$$\left.\frac{d\psi_d(\mathbb{P}_\epsilon)}{d\epsilon}\right|_{\epsilon=0} = \int_Z \phi_d(Z;\mathbb{P}) \left(\frac{d}{d\epsilon} \log d\mathbb{P}_\epsilon\right) \Bigg|_{\epsilon=0} d\mathbb{P}.$$

As shown in Lemma 7.1, we have that $\psi_d(\mathbb{P}_\epsilon) = \psi_d(\mathbb{P}) + \int_Z \phi_d(Z;\mathbb{P}) d(\mathbb{P}_\epsilon - \mathbb{P}) - R_d(\mathbb{P}, \mathbb{P}_\epsilon)$, which gives that

$$\begin{aligned}
\left.\frac{d\psi_d(\mathbb{P}_\epsilon)}{d\epsilon}\right|_{\epsilon=0} &= \left.\frac{d}{d\epsilon} \int \phi_d(Z;\mathbb{P}) d\mathbb{P}_\epsilon\right|_{\epsilon=0} - \left.\frac{d}{d\epsilon} R_d(\mathbb{P}, \mathbb{P}_\epsilon)\right|_{\epsilon=0} \\
&= \int_Z \phi_d(Z;\mathbb{P}) \left(\frac{d}{d\epsilon} \log d\mathbb{P}_\epsilon\right) \Bigg|_{\epsilon=0} d\mathbb{P}_\epsilon - \left.\frac{d}{d\epsilon} R_d(\mathbb{P}, \mathbb{P}_\epsilon)\right|_{\epsilon=0}.
\end{aligned}$$

Finally, as shown in Corollary 7.1, $R_d(\mathbb{P}, \mathbb{P}_\epsilon)$ only contains second-order products of errors between $\mathbb{P}$ and $\mathbb{P}_\epsilon$ and thus using the product rule on $\frac{d}{d\epsilon} R_d(\mathbb{P}, \mathbb{P}_\epsilon)$, one can get that

$$\left.\frac{d}{d\epsilon} R_d(\mathbb{P}, \mathbb{P}_\epsilon)\right|_{\epsilon=0} = 0.$$

Since our model is nonparametric, there is only one influence function which is efficient (Bickel et al., 1993; Tsiatis, 2007). Thus, ϕ_d is the efficient influence function of ψ_d. □

Proof of Theorem 7.1. For a given policy $d \in \mathcal{D}$, we have that

$$
\begin{aligned}
&\operatorname{Var}\{\phi_d(Z)\} = \mathbb{E}[\phi_d(Z)^2] = \mathbb{E}[(\xi_d(Z) - \psi_d)^2] \\
=&\mathbb{E}\left\{\left(\frac{Ad(X)}{\pi_1(X)}\frac{1/2 - \mathbb{1}\{Y \le m_1(X)\}}{f_{1,m}(X)}\right)^2 + \left(\frac{(1-A)(1-d(X))}{\pi_0(X)}\frac{1/2 - \mathbb{1}\{Y \le m_0(X)\}}{f_{0,m}(X)}\right)^2 + (m_d(X) - \psi_d)^2\right\} \\
=&\mathbb{E}\left\{\left(\frac{Ad(X)}{\pi_1(X)}\frac{1/2 - \mathbb{1}\{Y \le m_1(X)\}}{f_{1,m}(X)}\right)^2\right\} \\
&\quad + \mathbb{E}\left\{\left(\frac{(1-A)(1-d(X))}{\pi_0(X)}\frac{1/2 - \mathbb{1}\{Y \le m_0(X)\}}{f_{0,m}(X \mid X, 0)}\right)^2\right\} + \operatorname{Var}\{m_d(X)\} \\
=&\mathbb{E}_{X,A}\left\{\frac{Ad(X)}{\pi_1^2(X)}\frac{\mathbb{E}_{Y|X,A}\left[(1/2 - \mathbb{1}\{Y \le m_1(X)\})^2\right]}{f_{1,m}^2(X)}\right\} \\
&\quad + \mathbb{E}_{X,A}\left\{\frac{(1-A)(1-d(X))}{\pi_0(X)^2}\frac{\mathbb{E}_{Y|X,A}\left[(1/2 - \mathbb{1}\{Y \le m_0(X)\})^2\right]}{f_{0,m}(X)}\right\} + \operatorname{Var}\{m_d(X)\} \\
=&\mathbb{E}_X\left\{\frac{d(X)\operatorname{Var}\{\mathbb{1}\{Y \le m_1(X)|X, A=1\}\}}{\pi_1(X)f_{1,m}^2(X)} + \frac{(1-d(X))\operatorname{Var}\{\mathbb{1}\{Y \le m_0(X)\}|X, A=0\}}{\pi_0(X)f_{0,m}^2(X)}\right\} + \operatorname{Var}\{m_d(X)\} \\
=&\mathbb{E}\left\{\frac{d(X)}{4\pi_1(X)f_{1,m}^2(X)} + \frac{(1-d(X))}{4\pi_0(X)f_{0,m}^2(X)}\right\} + \operatorname{Var}\{m_d(X)\},
\end{aligned}
$$

where the first equality is true since $\mathbb{E}[\phi_d(Z)] = 0$, the forth equality uses the fact $A^2 = A$, the fifth equality holds as $\mathbb{E}_{Y|X,A=a}[\mathbb{1}\{Y \le m_a(X)\}] = 1/2$, and the last equality is true since $\operatorname{Var}\{\mathbb{1}\{Y \le m_1(X)|X, A=1\} = \mathbb{P}(Y \le m_1(X)|X, A=1)\,(1 - \mathbb{P}(Y \le m_1(X)|X, A=1)) = 1/4$. □

F.1.3 Proofs in Section 7.6

Proof of Theorem 7.2. We start with the following decomposition

$$
\begin{aligned}
\widehat{\psi}_{d,\mathrm{dr}} - \psi_d &= \mathbb{P}_n\xi_d(\widehat{\mathbb{P}}) - \mathbb{P}\xi_d(\widehat{\mathbb{P}}) \\
&= \underbrace{\left(\mathbb{P}_n - \mathbb{P}\right)\left(\xi_d(\widehat{\mathbb{P}}) - \xi_d(\mathbb{P})\right)}_{T_1} + \underbrace{\left(\mathbb{P}_n - \mathbb{P}\right)\xi_d(\mathbb{P})}_{T_2} + \underbrace{\mathbb{P}\left(\xi_d(\widehat{\mathbb{P}}) - \xi_d(\mathbb{P})\right)}_{T_3}.
\end{aligned}
$$

When $\widehat{\xi}_d$ is learned from a separate sample from the empirical measure $\mathbb{P}_n$, by Lemma F.2, we obtain that

$$
T_1 = O_{\mathbb{P}}\left(\frac{\|\widehat{\xi}_d - \xi_d\|}{\sqrt{n}}\right) = o_{\mathbb{P}}(1/\sqrt{n}),
$$

where the last equality follows from our assumption that $\|\widehat{\xi}_d - \xi_d\| = o_{\mathbb{P}}(1)$. On the other hand, when ξ_d and $\widehat{\xi}_d$ are contained in a Donsker class and $\|\widehat{\xi}_d - \xi_d\| = o_{\mathbb{P}}(1)$, by Vaart (2000, Lemma 19.24), we have that $T_1 = o_{\mathbb{P}}(1/\sqrt{n})$. Since $\mathbb{P}\xi_d - \psi_d = 0$, we have that

$$\begin{aligned} T_3 &= \mathbb{P}\left(\widehat{\xi}_d - \xi_d\right) = \mathbb{P}\left\{ \frac{\pi_1 d}{\widehat{\pi}_1} \frac{F_{1,m} - F_{1,\widehat{m}}}{\widehat{f}_{1,\widehat{m}}} + \frac{\pi_0(1-d)}{\widehat{\pi}_0} \frac{F_{0,m} - F_{0,\widehat{m}}}{\widehat{f}_{0,\widehat{m}}} + \widehat{m}_d - m_d \right\} \\ &= \mathbb{P}\left\{ d\left(\widehat{m}_1 - m_1 - \frac{\pi_1}{\widehat{\pi}_1} \frac{F_{1,\widehat{m}} - F_{1,m}}{\widehat{f}_{1,\widehat{m}}} \right) + (1-d)\left(\widehat{m}_0 - m_0 - \frac{\pi_0}{\widehat{\pi}_0} \frac{F_{0,\widehat{m}} - F_{0,m}}{\widehat{f}_{0,\widehat{m}}} \right) \right\} \\ &= R_d\left(\widehat{\mathbb{P}}, \mathbb{P}\right), \end{aligned}$$

where the second equality holds since $\mathbb{P}\xi_d = \psi_d = \mathbb{P}\{m_d\}$ and $\mathbb{E}_{Y|X,A=a}[\mathbb{1}\{Y \leq \widehat{m}_a(X)\}|X, A=a] = F_{a,\widehat{m}}(X)$. Finally, using Lemma 7.1, we obtain that

$$\begin{aligned} R_d(\widehat{\mathbb{P}}, \mathbb{P}) = \mathbb{P}\Bigg\{ & \frac{d\,(\widehat{m}_1 - m_1)}{\widehat{\pi}_1 \widehat{f}_{1,\widehat{m}}} \left((\widehat{\pi}_1 - \pi_1)\,\widehat{f}_{1,\widehat{m}} + \left(\widehat{f}_{1,\widehat{m}} - f_{1,m}\right)\pi_1 - (\widehat{m}_1 - m_1)\frac{f'_{1,c}\pi_1}{2} \right) \\ & + \frac{(1-d)\,(\widehat{m}_0 - m_0)}{\widehat{\pi}_0 \widehat{f}_{0,\widehat{m}}} \left((\widehat{\pi}_0 - \pi_0)\widehat{f}_{0,\widehat{m}} + (\widehat{f}_{0,\widehat{m}} - f_{0,m})\pi_0 - (\widehat{m}_0 - m_0)\frac{f'_{0,c}\pi_0}{2} \right) \Bigg\} \\ = O_{\mathbb{P}} & \left(\sum_{a=0}^{1} \|\widehat{m}_a - m_a\| \ \left(\left\| \widehat{\pi}_a \widehat{f}_{a,\widehat{m}} - \pi_a f_{a,m} \right\| + \|\widehat{m}_a - m_a\| \right) \right). \end{aligned}$$

□

Proof of Corollary 7.4. Following from the proof of Theorem 7.2, we have that $\widehat{\psi}_{d,\mathrm{dr}} - \psi_d = T_1 + T_2 + T_3$ where $T_1 = o_{\mathbb{P}}(1/\sqrt{n})$. From Theorem 7.1, we have that $\mathrm{Var}\{\xi_d\} = \mathrm{Var}\{\phi_d\} = \sigma_d^2 < +\infty$ since $\mathbb{P} \in \mathcal{P}$. By Central Limit Theorem, we have that $T_2 = o_{\mathbb{P}}(1/\sqrt{n})$ and

$$\sqrt{n}T_2 = \sqrt{n}(\mathbb{P}_n - \mathbb{P})\xi_d(\mathbb{P}) = \sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n} \xi_d(Z_i) - \mathbb{E}[\xi_d(Z)] \right) \rightsquigarrow \mathcal{N}(0, \mathrm{Var}(\xi_d)).$$

Using Slutsky's theorem and the assumptions that ensure $T_3 = o_{\mathbb{P}}(1/\sqrt{n})$, we have

$$\sqrt{n}(\widehat{\psi}_{d,\mathrm{dr}} - \psi_d) \rightsquigarrow \mathcal{N}(0, \sigma_d^2).$$

□

Proof of Theorem 7.3. For any $d \in \mathcal{D}$, by the fact that $\mathbb{P}\xi_d = \psi_d$, we have that

$$\begin{aligned} \widehat{\psi}_{\widehat{d^*},\mathrm{dr}} - \psi_{d^*} &= \mathbb{P}_n \widehat{\xi}_{\widehat{d^*}} - \mathbb{P}\xi_{\widehat{d^*}} + \mathbb{P}\xi_{\widehat{d^*}} - \mathbb{P}\xi_{d^*} \\ &= \mathbb{P}_n \widehat{\xi}_{\widehat{d^*}} - \mathbb{P}\xi_{\widehat{d^*}} + \mathbb{P}\gamma(\widehat{d^*} - d^*), \end{aligned}$$

where the second equality uses that

$$\begin{aligned}\mathbb{P}\xi_{\widehat{d}^*} - \mathbb{P}\xi_{d^*} &= \mathbb{P}\Bigg\{ \frac{A\left(\widehat{d}^*(X) - d^*(X)\right)}{\pi_1(X)} \frac{1/2 - \mathbb{1}\{Y \le m_1(X)\}}{f_{1,m}(X)} \\ &\qquad + \frac{(1-A)(d^*(X) - \widehat{d}^*(X))}{\pi_0(X)} \frac{1/2 - \mathbb{1}\{Y \le m_0(X)\}}{f_{0,m}(X)} + m_{\widehat{d}^*}(X) - m_{d^*}(X) \Bigg\} \\ &= \mathbb{P}\left\{ m_1(\widehat{d}^* - d^*) + m_0(d^* - \widehat{d}^*) \right\} = \mathbb{P}\gamma(\widehat{d}^* - d^*).\end{aligned}$$

The second equality above comes from the fact that $\mathbb{E}_{Y|X,A=a}[\mathbb{1}\{Y \le m_a(X)\}] = 1/2$. Further, by Lemma F.1 and the fact that $d^* = \mathbb{1}\{\gamma > 0\}$, we obtain that

$$\begin{aligned}\mathbb{P}\gamma(\widehat{d}^* - d^*) &\le \mathbb{P}\gamma\mathbb{1}\{|\gamma| \le |\widehat{\gamma} - \gamma|\} \le \mathbb{P}|\widehat{\gamma} - \gamma|\mathbb{1}\{|\gamma| \le |\widehat{\gamma} - \gamma|\} \\ &\le \|\widehat{\gamma} - \gamma\|_\infty \mathbb{P}\left(|\gamma| \le \|\widehat{\gamma} - \gamma\|_\infty\right) \lesssim \|\widehat{\gamma} - \gamma\|_\infty^{1+\alpha},\end{aligned}$$

where the last inequality follows from the margin condition. On the other hand, similar to the proof of Theorem 7.2, we use the decomposition

$$\mathbb{P}_n\widehat{\xi}_{\widehat{d}^*} - \mathbb{P}\xi_{\widehat{d}^*} = (\mathbb{P}_n - \mathbb{P})\left(\widehat{\xi}_{\widehat{d}^*} - \xi_{d^*}\right) + (\mathbb{P}_n - \mathbb{P})\,\xi_{d^*} + \mathbb{P}\left(\widehat{\xi}_{\widehat{d}^*} - \xi_{\widehat{d}^*}\right).$$

By the fact that $\mathbb{P}\phi_{\widehat{d}^*} = \psi_{\widehat{d}^*} + \mathbb{P}\xi_{\widehat{d}^*} = 0$, we have that

$$\begin{aligned}&\mathbb{P}\left(\widehat{\xi}_{\widehat{d}^*} - \xi_{\widehat{d}^*}\right) \\ &= \mathbb{P}\left\{ \frac{\pi_1\widehat{d}^*}{\widehat{\pi}_1} \frac{F_{1,m} - F_{1,\widehat{m}}}{\widehat{f}_{1,\widehat{m}}} + \frac{\pi_0(1-\widehat{d}^*)}{\widehat{\pi}_0} \frac{F_{0,m} - F_{0,\widehat{m}}}{\widehat{f}_{0,\widehat{m}}} + \widehat{m}_{\widehat{d}^*} - m_{\widehat{d}^*} \right\} \\ &= \mathbb{P}\left\{ \widehat{d}^* \left(\widehat{m}_1 - m_1 - \frac{\pi_1}{\widehat{\pi}_1} \frac{F_{1,\widehat{m}} - F_{1,m}}{\widehat{f}_{1,\widehat{m}}} \right) + (1-\widehat{d}^*) \left(\widehat{m}_0 - m_0 - \frac{\pi_0}{\widehat{\pi}_0} \frac{F_{0,\widehat{m}} - F_{0,m}}{\widehat{f}_{0,\widehat{m}}} \right) \right\} \\ &= R_{\widehat{d}^*}\left(\widehat{\mathbb{P}}, \mathbb{P}\right) = O_{\mathbb{P}}\left(\sum_{a=0}^{1} \|\widehat{m}_a - m_a\| \left(\left\| \widehat{\pi}_a \widehat{f}_{a,\widehat{m}} - \pi_a f_{a,m} \right\| + \|\widehat{m}_a - m_a\| \right) \right).\end{aligned}$$

Finally, for the term $(\mathbb{P}_n - \mathbb{P})\left(\widehat{\xi}_{\widehat{d}^*} - \xi_{d^*}\right)$: when $\widehat{\xi}_{\widehat{d}^*}$ is estimated from a separate sample $Z^{n,0}$ from the empirical measure over Z^n, by Lemma F.2 we have that

$$(\mathbb{P}_n - \mathbb{P})\left(\widehat{\xi}_{\widehat{d}^*} - \xi_{d^*}\right) = O_{\mathbb{P}}\left(\frac{\|\widehat{\xi}_{\widehat{d}^*} - \xi_{d^*}\|}{\sqrt{n}} \right) = o_{\mathbb{P}}(1/\sqrt{n}),$$

where the last equality follows from our assumption that $\|\widehat{\xi}_{\widehat{d}^*} - \xi_{d^*}\| = o_{\mathbb{P}}(1)$. When ξ_{d^*} and $\widehat{\xi}_{\widehat{d}^*}$ are contained in a Donsker class and $\|\widehat{\xi}_{\widehat{d}^*} - \xi_{d^*}\| = o_{\mathbb{P}}(1)$, by Vaart (2000, Lemma 19.24), we have that $(\mathbb{P}_n - \mathbb{P})\left(\widehat{\xi}_{\widehat{d}^*} - \xi_{d^*}\right) = o_{\mathbb{P}}(1/\sqrt{n})$. □

Proof of Corollary 7.5. From Theorem 7.3, we have that

$$\widehat{\psi}_{\widehat{d}^*,\mathrm{dr}} - \psi_{d^*} = (\mathbb{P}_n - \mathbb{P})\,\xi_{d^*} + O_{\mathbb{P}}\Big(o_{\mathbb{P}}(1/\sqrt{n}) + \underbrace{\|\widehat{\gamma} - \gamma\|_\infty^{1+\alpha} + \sum_{a=0}^{1} \|\widehat{m}_a - m_a\| \left(\|\widehat{\pi}_a \widehat{f}_{a,\widehat{m}} - \pi_a f_{a,m}\| + \|\widehat{m}_a - m_a\|\right)}_{R}\Big).$$

The conditions in Corollary 7.5 ensure that $R = o_{\mathbb{P}}(1/\sqrt{n})$, which gives that

$$\sqrt{n}(\widehat{\psi}_{\widehat{d}^*,\mathrm{dr}} - \psi_{d^*}) \rightsquigarrow \mathcal{N}\left(0, \sigma_{d^*}^2\right),$$

since $\mathrm{Var}\{\xi_{d^*}\} = \mathrm{Var}\{\phi_{d^*}\} = \sigma_{d^*}^2$. □

F.1.4 Proofs in Section 7.7

Proof of Theorem 7.4. We begin with a decomposition:

$$\begin{aligned}
&\widehat{\gamma}_{\mathrm{dr}}(x) - \gamma(x) = \sum_{i=1}^{n} w_i(x; X^n)\widehat{g}(Z_i) - \gamma(x) \\
&= \left(\sum_{i=1}^{n} w_i(x; X^n) g(Z_i) - \gamma(x)\right) + \sum_{i=1}^{n} w_i(x; X^n) h(X_i) + \sum_{i=1}^{n} w_i(x; X^n)\left(\widehat{g}(Z_i) - g(Z_i) - h(X_i)\right) \\
&= (\widetilde{\gamma}(x) - \gamma(x)) + \sum_{i=1}^{n} w_i(x; X^n) h(X_i) + \sum_{i=1}^{n} w_i(x; X^n)\left(\widehat{g}(Z_i) - g(Z_i) - h(X_i)\right),
\end{aligned}$$

where $h(x) = \mathbb{E}[\widehat{g}(Z) - g(Z) | D^n, X = x]$ and $D^n = (D_1, D_2)$ is the training data for obtaining $\widehat{g}$.

To bound $h(x)$, we obtain that

$$\begin{aligned}
&h(x) = \mathbb{E}[\widehat{g}(Z) - g(Z) | D^n, X = x] = \mathbb{E}[\widehat{g}(Z) | D^n, X = x] - \gamma(x) \\
&= \left(\widehat{m}_1(x) - m_1(x) - \frac{\pi_1(x)}{\widehat{\pi}_1(x)} \frac{F_{1,\widehat{m}}(x) - F_{1,m}(x)}{\widehat{f}_{1,\widehat{m}}(x)}\right) - \left(\widehat{m}_0(x) - m_0(x) - \frac{\pi_0(x)}{\widehat{\pi}_0(x)} \frac{F_{0,\widehat{m}}(x) - F_{0,m}(x)}{\widehat{f}_{0,\widehat{m}}(x)}\right) \\
&= \left((\widehat{m}_1(x) - m_1(x))\left(1 - \frac{\pi_1(x)}{\widehat{\pi}_1(x)} \frac{f_{1,m}(x)}{\widehat{f}_{1,\widehat{m}}(x)}\right) - \frac{f'_{1,c}(x)\pi_1(x)(\widehat{m}_1(x) - m_1(x))^2}{2\widehat{\pi}_1(x)\widehat{f}_{1,\widehat{m}}(x)}\right) \\
&\quad - \left((\widehat{m}_0(x) - m_0(x))\left(1 - \frac{\pi_0(x)}{\widehat{\pi}_0(x)} \frac{f_{0,m}(x)}{\widehat{f}_{0,\widehat{m}}(x)}\right) - \frac{f'_{0,c}(x)\pi_0(x)(\widehat{m}_0(x) - m_0(x))^2}{2\widehat{\pi}_0(x)\widehat{f}_{0,\widehat{m}}(x)}\right) \\
&\lesssim \sum_{a=0}^{1} |\widehat{m}_a(x) - m_a(x)| \left(|\widehat{\pi}_a(x)\widehat{f}_{a,\widehat{m}}(x) - \pi_a(x) f_{a,m}(x)| + |\widehat{m}_a(x) - m_a(x)|\right),
\end{aligned}$$

where $f'_{a,c}(x) \le L$ is the derivative of $f_a(y \mid x)$ at $y = c_a(x)$ for some value $c_a(x)$ between $m_a(x)$ and $\widehat{m}_a(x)$ and the third equality follows similarly to the proof of Corollary 7.1. This gives that

$$\sum_{i=1}^{n} w_i(x; X^n) h(X_i)$$

$$
\begin{aligned}
&\lesssim \sum_{i=1}^{n} |w_i(x;X^n)| \left(\sum_{a=0}^{1} |\widehat{m}_a(X_i) - m_a(X_i)| \left(|\widehat{\pi}_a(X_i)\widehat{f}_{a,\widehat{m}}(X_i) - \pi_a(X_i) f_{a,m}(X_i)| + |\widehat{m}_a(X_i) - m_a(X_i)| \right) \right) \\
&= \sum_{a=0}^{1} \sum_{i=1}^{n} |w_i(x;X^n)| |\widehat{m}_a(X_i) - m_a(X_i)| \left(\left| \pi_a(X_i) f_{a,m}(X_i) - \widehat{\pi}_a(X_i)\widehat{f}_{a,\widehat{m}}(X_i) \right| + |\widehat{m}_a(X_i) - m_a(X_i)| \right) \\
&= \sum_{a=0}^{1} \sum_{i=1}^{n} |w_i(x;X^n)| |\widehat{m}_a(X_i) - m_a(X_i)| \cdot \left| \pi_a(X_i) f_{a,m}(X_i) - \widehat{\pi}_a(X_i)\widehat{f}_{a,\widehat{m}}(X_i) \right| \\
&\quad + \sum_{a=0}^{1} \sum_{i=1}^{n} |w_i(x;X^n)| \cdot |\widehat{m}_a(X_i) - m_a(X_i)| \cdot |\widehat{m}_a(X_i) - m_a(X_i)| \\
&\leq \sum_{a=0}^{1} \sqrt{\sum_{i=1}^{n} |w_i(x;X^n)| |\widehat{m}_a(X_i) - m_a(X_i)|^2} \cdot \sqrt{\sum_{i=1}^{n} |w_i(x;X^n)| \left| \pi_a(X_i) f_{a,m}(X_i) - \widehat{\pi}_a(X_i)\widehat{f}_{a,\widehat{m}}(X_i) \right|^2} \\
&\quad + \sum_{a=0}^{1} \sqrt{\sum_{i=1}^{n} |w_i(x;X^n)| |\widehat{m}_a(X_i) - m_a(X_i)|^2} \cdot \sqrt{\sum_{i=1}^{n} |w_i(x;X^n)| \, |\widehat{m}_a(X_i) - m_a(X_i)|^2} \\
&= \left(\sum_{i=1}^{n} |w_i(x;X^n)| \right) \cdot \left(\sum_{a=0}^{1} \|\widehat{m}_a - m_a\|_w \left(\|\widehat{\pi}_a \widehat{f}_{a,\widehat{m}} - \pi_a f_{a,m}\|_w + \|\widehat{m}_a - m_a\|_w \right) \right).
\end{aligned}
$$

To bound $G(x) := \sum_{i=1}^{n} w_i(x;X^n) \left(\widehat{g}(Z_i) - g(Z_i) - h(X_i) \right)$, we observe that by definition,

$$
\mathbb{E}[\widehat{g}(Z_i) - g(Z_i) - h(X_i) | D^n, X^n] = 0,
$$

where X^n are the covariates in D_3. We also have

$$
\begin{aligned}
\text{Var}\{G(x)|D^n, X^n\} &= \text{Var}\left\{ \sum_{i=1}^{n} w_i(x;X^n) \left(\widehat{g}(Z_i) - g(Z_i) - h(X_i) \right) | D^n, X^n \right\} \\
&= \sum_{i=1}^{n} w_i(x;X^n)^2 \text{Var}\left\{ \widehat{g}(Z_i) - g(Z_i) - h(X_i) | D^n, X^n \right\} \\
&= \sum_{i=1}^{n} w_i(x;X^n)^2 \text{Var}\left\{ \widehat{g}(Z_i) - g(Z_i) | D^n, X^n \right\} \\
&\leq \|\widehat{g} - g\|_{w^2}^2 \sum_{i=1}^{n} w_i(x;X^n)^2,
\end{aligned}
$$

where the third equality holds since $\text{Var}\left\{ \widehat{g}(Z_i) - g(Z_i) | D^n, X^n \right\} = \mathbb{E}[(\widehat{g}(Z_i) - g(Z_i))^2 | D^n, X^n] - \mathbb{E}[\widehat{g}(Z_i) - g(Z_i) | D^n, X^n]^2 \leq \mathbb{E}[(\widehat{g}(Z_i) - g(Z_i))^2 | D^n, X^n]$. We note that

$$
\begin{aligned}
&\mathbb{E}[(\widetilde{\gamma}(x) - \gamma(x))^2] \\
&= \mathbb{E}\left[\left(\sum_{i=1}^{n} w_i(x;X^n) \left(g(Z_i) - \gamma(X_i) \right) + \sum_{i=1}^{n} w_i(x;X^n) \gamma(X_i) - \gamma(x) \right)^2 \right] \\
&= \mathbb{E}\left[\left(\sum_{i=1}^{n} w_i(x;X^n) \left(g(Z_i) - \gamma(X_i) \right) \right)^2 \right] + \mathbb{E}\left[\left(\sum_{i=1}^{n} w_i(x;X^n) \gamma(X_i) - \gamma(x) \right)^2 \right]
\end{aligned}
$$

$$= \mathbb{E}\left[\sum_{i=1}^{n} w_i(x;X^n)^2 \mathrm{Var}\{g(Z_i)|X_i\}\right] + \mathbb{E}\left[\left(\sum_{i=1}^{n} w_i(x;X^n)\gamma(X_i) - \gamma(x)\right)^2\right]$$
$$\geq \sigma^2_{\min} \sum_{i=1}^{n} \mathbb{E}\left[w_i(x;X^n)^2\right],$$

where the second and third equality hold since $\mathbb{E}[g(Z_i)|X^n] = \gamma(X_i)$ and all samples are independent. Putting it altogether, using Markov's inequality, we obtain

$$\mathbb{P}\left(\frac{|G(x)|}{\|\widehat{g}-g\|_{w^2}\sqrt{\mathbb{E}[(\widetilde{\gamma}(x)-\gamma(x))^2]}} \geq t\right)$$
$$= \mathbb{E}\left[\mathbb{P}\left(\frac{|G(x)|}{\|\widehat{g}-g\|_{w^2}\sqrt{\mathbb{E}[(\widetilde{\gamma}(x)-\gamma(x))^2]}} \geq t \Big| D^n, X^n\right)\right]$$
$$\leq \frac{1}{t^2}\cdot\mathbb{E}\left[\frac{\mathrm{Var}\{G(x)|D^n,X^n\}}{\|\widehat{g}-g\|^2_{w^2}\mathbb{E}[(\widetilde{\gamma}(x)-\gamma(x))^2]}\right] \leq \frac{1}{\sigma^2_{\min}t^2}.$$

Thus, we have that $G(x) = O_{\mathbb{P}}(\|\widehat{g}-g\|_{w^2}\sqrt{\mathbb{E}[(\widetilde{\gamma}(x)-\gamma(x))^2]})$ and

$$G(x)^2 = O_{\mathbb{P}}\left(\|\widehat{g}-g\|^2_{w^2}\mathbb{E}[(\widetilde{\gamma}(x)-\gamma(x))^2]\right).$$

The proof completes by realizing that $(a+b+c)^2 \leq 3a^2+3b^2+3c^2$. □

F.2 Auxiliary Lemmas

Below are the auxiliary lemmas we have used to prove the results in this paper. For completeness, we include their proofs here as well.

Lemma F.1

(Kennedy et al., 2020, Lemma 1) Let $\widehat{f}, f$ be functions taking any real values. Then

$$|\mathbb{1}\{\widehat{f}>0\} - \mathbb{1}\{f>0\}| \leq \mathbb{1}\{|f| \leq |\widehat{f}-f|\}.$$

Proof. It follows that

$$|\mathbb{1}\{\widehat{f}>0\} - \mathbb{1}\{f>0\}| = \mathbb{1}\{f, \widehat{f} \text{ have opposite signs}\} \leq \mathbb{1}\{|f| \leq |\widehat{f}-f|\},$$

since $|\widehat{f}-f| = |\widehat{f}| + |f|$ when $f, \widehat{f}$ have opposite signs. □

Lemma F.2

(Kennedy et al., 2020, Lemma 2) Let $\mathbb{P}_n$ denote the empirical measure over $Z^n = (Z_1,\ldots,Z_n)$ and $\widehat{\phi}$ be a function estimated from a sample $Z^{n,0} = (Z^0_1,\ldots,Z^0_n)$, which

is independent of Z^n, then for any function ϕ,

$$(\mathbb{P}_n - \mathbb{P})(\widehat{\phi} - \phi) = O_{\mathbb{P}}\left(\frac{\|\widehat{\phi} - \phi\|}{\sqrt{n}}\right).$$

Proof. First, we notice that

$$\text{Var}\left\{(\mathbb{P}_n - \mathbb{P})(\widehat{\phi} - \phi)\middle| Z^{n,0}\right\} = \text{Var}\left\{\mathbb{P}_n(\widehat{\phi} - \phi)\middle| Z^{n,0}\right\} = \frac{1}{n}\text{Var}\left\{\widehat{\phi} - \phi | Z^{n,0}\right\} \leq \frac{\|\widehat{\phi} - \phi\|_2^2}{n},$$

where the first equality is true since conditioned on $Z^{n,0}$, $\mathbb{P}(\widehat{\phi} - \phi)$ is a constant. Then by applying the law of total expectation and Chebyshev's inequality, we obtain that

$$\begin{aligned}\mathbb{P}\left(\frac{|(\mathbb{P}_n - \mathbb{P})(\widehat{\phi} - \phi)|}{\|\widehat{\phi} - \phi\|_2/\sqrt{n}} \geq t\right) &= \mathbb{E}\left\{\mathbb{P}\left(\frac{|(\mathbb{P}_n - \mathbb{P})(\widehat{\phi} - \phi)|}{\|\widehat{\phi} - \phi\|_2/\sqrt{n}} \geq t\middle| Z^{n,0}\right)\right\} \\ &\leq \mathbb{E}\left\{\frac{\text{Var}\left\{(\mathbb{P}_n - \mathbb{P})(\widehat{\phi} - \phi)\middle| Z^{n,0}\right\}}{\|\widehat{\phi} - \phi\|_2^2 t^2/n}\right\} \leq \frac{1}{t^2},\end{aligned}$$

where we have utilized the fact that $\mathbb{E}\left\{(\mathbb{P}_n - \mathbb{P})(\widehat{\phi} - \phi)\middle| Z^{n,0}\right\} = \mathbb{E}\{\widehat{\phi} - \phi | Z^{n,0}\} - \mathbb{P}(\widehat{\phi} - \phi) = 0$. □

Bibliography

[1] Ronald Aylmer Fisher. “Design of experiments”. In: *Br Med J* 1.3923 (1936), pp. 554–554 (cit. on p. 18).

[2] Herbert Robbins. “Some aspects of the sequential design of experiments”. In: *Bulletin of the American Mathematical Society* 58.5 (1952), pp. 527–535 (cit. on pp. 9, 17, 40).

[3] William T Tucker. “The development of brand loyalty”. In: *Journal of Marketing research* 1.3 (1964), pp. 32–35 (cit. on pp. 4, 11, 39, 44).

[4] William F Sharpe. “Mutual fund performance”. In: *The Journal of business* 39.1 (1966), pp. 119–138 (cit. on p. 71).

[5] Peter J Huber et al. “The behavior of maximum likelihood estimates under nonstandard conditions”. In: *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*. Vol. 1. 1. University of California Press. 1967, pp. 221–233 (cit. on pp. 83, 89).

[6] J Douglas McConnell. “The development of brand loyalty: an experimental study”. In: *Journal of Marketing Research* 5.1 (1968), pp. 13–19 (cit. on pp. 39, 44).

[7] Philip M Groves and Richard F Thompson. “Habituation: a dual-process theory.” In: *Psychological review* 77.5 (1970), p. 419 (cit. on p. 9).

[8] Ronald A Howard and James E Matheson. “Risk-sensitive Markov decision processes”. In: *Management science* 18.7 (1972), pp. 356–369 (cit. on p. 26).

[9] Donald B Rubin. “Estimating causal effects of treatments in randomized and nonrandomized studies.” In: *Journal of Educational Psychology* 66.5 (1974), p. 688 (cit. on p. 85).

[10] SS Vallender. “Calculation of the Wasserstein distance between probability distributions on the line”. In: *Theory of Probability & Its Applications* 18.4 (1974), pp. 784–786 (cit. on p. 183).

[11] Charles J Stone. “Consistent nonparametric regression”. In: *The annals of statistics* (1977), pp. 595–620 (cit. on p. 102).

[12] GE Box. “All models are wrong, but some are useful”. In: *Robustness in Statistics* 202.1979 (1979), p. 549 (cit. on p. 24).

[13] Peter Gänssler and Winfried Stute. “Empirical processes: a survey of results for independent and identically distributed random variables”. In: *The Annals of Probability* (1979), pp. 193–243 (cit. on p. 71).

[14] John C Gittins. "Bandit processes and dynamic allocation indices". In: *Journal of the Royal Statistical Society: Series B (Methodological)* 41.2 (1979), pp. 148–164 (cit. on p. 41).

[15] Eli P Cox III. "The optimal number of response alternatives for a scale: A review". In: *Journal of marketing research* 17.4 (1980), pp. 407–422 (cit. on pp. 9, 16).

[16] Leigh McAlister. "A dynamic attribute satiation model of variety-seeking behavior". In: *Journal of Consumer Research* 9.2 (1982), pp. 141–150 (cit. on p. 40).

[17] Leigh McAlister and Edgar Pessemier. "Variety seeking behavior: An interdisciplinary review". In: *Journal of Consumer research* 9.3 (1982), pp. 311–322 (cit. on p. 40).

[18] Kenneth S Alexander. "Probability inequalities for empirical processes and a law of the iterated logarithm". In: *The Annals of Probability* (1984), pp. 1041–1067 (cit. on p. 71).

[19] Aharon Ben-Tal and Marc Teboulle. "Expected utility, penalty functions, and duality in stochastic nonlinear programming". In: *Management Science* 32.11 (1986), pp. 1445–1466 (cit. on pp. 71, 172).

[20] Bernard W Silverman. *Density estimation for statistics and data analysis*. Vol. 26. CRC press, 1986 (cit. on p. 105).

[21] Peter J Bickel and Yaacov Ritov. "Estimating integrated squared density derivatives: sharp best order of convergence estimates". In: *Sankhyā: The Indian Journal of Statistics, Series A* (1988), pp. 381–393 (cit. on p. 96).

[22] Peter Whittle. "Restless bandits: Activity allocation in a changing world". In: *Journal of applied probability* 25.A (1988), pp. 287–298 (cit. on p. 41).

[23] Dieter Denneberg. "Distorted probabilities and insurance premiums". In: *Methods of Operations Research* 63.3 (1990), pp. 3–5 (cit. on pp. 70, 76).

[24] Pascal Massart. "The tight constant in the Dvoretzky-Kiefer-Wolfowitz inequality". In: *The annals of Probability* (1990), pp. 1269–1283 (cit. on p. 71).

[25] Whitney K Newey. "Semiparametric efficiency bounds". In: *Journal of Applied Econometrics* 5.2 (1990), pp. 99–135 (cit. on pp. 83, 92).

[26] Colin Camerer and Martin Weber. "Recent developments in modeling preferences: Uncertainty and ambiguity". In: *Journal of risk and uncertainty* 5.4 (1992), pp. 325–370 (cit. on p. 33).

[27] Amos Tversky and Daniel Kahneman. "Advances in prospect theory: Cumulative representation of uncertainty". In: *Journal of Risk and uncertainty* 5.4 (1992), pp. 297–323 (cit. on pp. 4, 27, 28, 31, 35).

[28] Peter J Bickel, Chris AJ Klaassen, Peter J Bickel, Ya'acov Ritov, J Klaassen, Jon A Wellner, and YA'Acov Ritov. *Efficient and adaptive estimation for semiparametric models*. Vol. 4. Johns Hopkins University Press Baltimore, 1993 (cit. on pp. 83, 92, 96, 193).

[29] James Hamilton. *Time series analysis*. Princeton, N.J: Princeton University Press, 1994. ISBN: 978-0691042893 (cit. on pp. 50, 133).

[30] Jerzy A Filar, Dmitry Krass, and Keith W Ross. "Percentile performance criteria for limiting average Markov decision processes". In: *IEEE Transactions on Automatic Control* 40.1 (1995), pp. 2–10 (cit. on p. 26).

[31] Barbara E Kahn. "Consumer variety-seeking among goods and services: An integrative review". In: *Journal of retailing and consumer services* 2.3 (1995), pp. 139–148 (cit. on p. 40).
[32] Gary S Becker. *Accounting for tastes*. Harvard University Press, 1996 (cit. on p. 19).
[33] Shaun Wang. "Premium calculation by transforming the layer premium density". In: *ASTIN Bulletin: The Journal of the IAA* 26.1 (1996), pp. 71–92 (cit. on pp. 70, 76).
[34] George Wu and Richard Gonzalez. "Curvature of the probability weighting function". In: *Management science* 42.12 (1996), pp. 1676–1690 (cit. on p. 27).
[35] Alan Frieze and Mark Jerrum. "Improved approximation algorithms for max k-cut and max bisection". In: *Algorithmica* 18.1 (1997), pp. 67–81 (cit. on p. 46).
[36] Shaun S Wang, Virginia R Young, and Harry H Panjer. "Axiomatic characterization of insurance prices". In: *Insurance: Mathematics and economics* 21.2 (1997), pp. 173–183 (cit. on p. 76).
[37] Jinyong Hahn. "On the role of the propensity score in efficient semiparametric estimation of average treatment effects". In: *Econometrica* (1998), pp. 315–331 (cit. on p. 94).
[38] Leslie Pack Kaelbling, Michael L Littman, and Anthony R Cassandra. "Planning and acting in partially observable stochastic domains". In: *Artificial intelligence* 101.1-2 (1998), pp. 99–134 (cit. on p. 63).
[39] Drazen Prelec et al. "The probability weighting function". In: *ECONOMETRICA-EVANSTON ILL-* 66 (1998), pp. 497–528 (cit. on p. 27).
[40] Philippe Artzner, Freddy Delbaen, Jean-Marc Eber, and David Heath. "Coherent measures of risk". In: *Mathematical finance* 9.3 (1999), pp. 203–228 (cit. on pp. 69, 71).
[41] Ronald M Epstein. "Mindful practice". In: *Jama* 282.9 (1999), pp. 833–839 (cit. on p. 23).
[42] Richard Gonzalez and George Wu. "On the shape of the probability weighting function". In: *Cognitive psychology* 38.1 (1999), pp. 129–166 (cit. on p. 28).
[43] Lennart Ljung. "System identification". In: *Wiley encyclopedia of electrical and electronics engineering* (1999), pp. 1–19 (cit. on pp. 50, 133).
[44] John W Payne, James R Bettman, David A Schkade, Norbert Schwarz, and Robin Gregory. "Measuring constructed preferences: Towards a building code". In: *Elicitation of preferences*. Springer, 1999, pp. 243–275 (cit. on p. 22).
[45] Rebecca K Ratner, Barbara E Kahn, and Daniel Kahneman. "Choosing less-preferred experiences for the sake of variety". In: *Journal of consumer research* 26.1 (1999), pp. 1–15 (cit. on p. 39).
[46] Vladimir Naumovich Vapnik. "An overview of statistical learning theory". In: *IEEE transactions on neural networks* 10.5 (1999), pp. 988–999 (cit. on p. 70).
[47] Julia Lynn Wirch and Mary R Hardy. "A synthesis of risk measures for capital adequacy". In: *Insurance: mathematics and Economics* 25.3 (1999), pp. 337–347 (cit. on p. 76).
[48] Pascal Massart. "Some applications of concentration inequalities to statistics". In: *Annales de la Faculté des sciences de Toulouse: Mathématiques*. Vol. 9. 2000, pp. 245–303 (cit. on p. 180).
[49] R Tyrrell Rockafellar, Stanislav Uryasev, et al. "Optimization of conditional value-at-risk". In: *Journal of risk* 2 (2000), pp. 21–42 (cit. on pp. 29, 37, 69, 71).

[50] Aad W van der Vaart. *Asymptotic statistics*. Vol. 3. Cambridge university press, 2000 (cit. on pp. 194, 195).

[51] Enrico Diecidue and Peter P Wakker. "On the intuition of rank-dependent utility". In: *Journal of Risk and Uncertainty* 23.3 (2001), pp. 281–298 (cit. on pp. 28, 118).

[52] Shigeo Kusuoka. "On law invariant coherent risk measures". In: *Advances in mathematical economics*. Springer, 2001, pp. 83–95 (cit. on p. 72).

[53] Julia L Wirch and Mary R Hardy. "Distortion risk measures: Coherence and stochastic dominance". In: *International congress on insurance: Mathematics and economics*. 2001, pp. 15–17 (cit. on pp. 69, 76).

[54] Carlo Acerbi. "Spectral measures of risk: A coherent representation of subjective risk aversion". In: *Journal of Banking & Finance* 26.7 (2002), pp. 1505–1518 (cit. on pp. 69, 71).

[55] Peter Auer, Nicolo Cesa-Bianchi, and Paul Fischer. "Finite-time analysis of the multiarmed bandit problem". In: *Machine learning* 47.2 (2002), pp. 235–256 (cit. on p. 14).

[56] Peter L Bartlett and Shahar Mendelson. "Rademacher and Gaussian complexities: Risk bounds and structural results". In: *Journal of Machine Learning Research* 3.Nov (2002), pp. 463–482 (cit. on p. 70).

[57] George Casella and Roger L Berger. *Statistical inference*. Vol. 2. Duxbury Pacific Grove, CA, 2002 (cit. on pp. 83, 89).

[58] László Györfi, Michael Kohler, Adam Krzyzak, Harro Walk, et al. *A distribution-free theory of nonparametric regression*. Vol. 1. Springer, 2002 (cit. on p. 102).

[59] R Tyrrell Rockafellar and Stanislav Uryasev. "Conditional value-at-risk for general loss distributions". In: *Journal of Banking & Finance* 26.7 (2002), pp. 1443–1471 (cit. on p. 84).

[60] Aad W van der Vaart. "Semiparametric statistics". In: *Lecture Notes in Math.* 1781 (2002) (cit. on pp. 83, 92, 94–96).

[61] Susan A Murphy. "Optimal dynamic treatment regimes". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 65.2 (2003), pp. 331–355 (cit. on pp. 83, 86, 103).

[62] Mark J van der Laan and James M Robins. *Unified methods for censored longitudinal data and causality*. Springer Science & Business Media, 2003 (cit. on pp. 83, 92).

[63] Colin F Camerer, George Loewenstein, and Matthew Rabin. *Advances in behavioral economics*. Princeton university press, 2004 (cit. on p. 33).

[64] Peter J Huber. *Robust statistics*. Vol. 523. John Wiley & Sons, 2004 (cit. on pp. 83, 89).

[65] Drazen Prelec. "Decreasing impatience: a criterion for Non-stationary time preference and "hyperbolic" discounting". In: *Scandinavian Journal of Economics* 106.3 (2004), pp. 511–532 (cit. on p. 11).

[66] James M Robins. "Optimal structural nested models for optimal sequential decisions". In: *Proceedings of the second seattle Symposium in Biostatistics*. Springer. 2004, pp. 189–326 (cit. on pp. 83, 103).

[67] Alexander B Tsybakov et al. "Optimal aggregation of classifiers in statistical learning". In: *The Annals of Statistics* 32.1 (2004), pp. 135–166 (cit. on p. 98).

[68] Victor Chernozhukov and Christian Hansen. "An IV model of quantile treatment effects". In: *Econometrica* 73.1 (2005), pp. 245–261 (cit. on p. 100).

[69] Damien Ernst, Pierre Geurts, and Louis Wehenkel. "Tree-based batch mode reinforcement learning". In: *Journal of Machine Learning Research* 6.Apr (2005), pp. 503–556 (cit. on p. 48).

[70] Faruk Gul and Wolfgang Pesendorfer. "The revealed preference theory of changing tastes". In: *The Review of Economic Studies* 72.2 (2005), pp. 429–448 (cit. on p. 11).

[71] José AF Machado and José Mata. "Counterfactual decomposition of changes in wage distributions using quantile regression". In: *Journal of applied Econometrics* 20.4 (2005), pp. 445–465 (cit. on p. 100).

[72] Blaise Melly. "Decomposition of differences in distribution using quantile regression". In: *Labour economics* 12.4 (2005), pp. 577–590 (cit. on p. 100).

[73] Martin Riedmiller. "Neural fitted Q iteration–first experiences with a data efficient neural reinforcement learning method". In: *European Conference on Machine Learning*. Springer. 2005, pp. 317–328 (cit. on p. 48).

[74] Guy Shani, David Heckerman, and Ronen I Brafman. "An MDP-based recommender system". In: *Journal of Machine Learning Research* 6.Sep (2005), pp. 1265–1295 (cit. on pp. 9–11, 25).

[75] Guy Shani, David Heckerman, and Ronen I. Brafman. "An MDP-Based Recommender System". In: *Journal of Machine Learning Research* 6.43 (2005), pp. 1265–1295. URL: http://jmlr.org/papers/v6/shani05a.html (cit. on p. 56).

[76] Nicol'o Cesa-Bianchi and G´abor Lugosi. *Prediction, learning, and games*. Cambridge Univ Press, 2006 (cit. on p. 55).

[77] Larry Christensen and Alisa Brooks. "Changing food preference as a function of mood". In: *The journal of psychology* 140.4 (2006), pp. 293–306 (cit. on p. 11).

[78] A Hitchhiker's Guide. *Infinite dimensional analysis*. Springer, 2006 (cit. on p. 73).

[79] Nicolai Meinshausen. "Quantile regression forests". In: *Journal of Machine Learning Research* 7.Jun (2006), pp. 983–999 (cit. on p. 105).

[80] Marc Oliver Rieger and Mei Wang. "Cumulative prospect theory and the St. Petersburg paradox". In: *Economic Theory* 28.3 (2006), pp. 665–679 (cit. on pp. 27, 28).

[81] Vladimir Vapnik. *Estimation of dependences based on empirical data*. Springer Science & Business Media, 2006 (cit. on p. 71).

[82] Larry Wasserman. *All of nonparametric statistics*. Springer Science & Business Media, 2006 (cit. on p. 101).

[83] Jean-Yves Audibert and Alexandre B Tsybakov. "Fast learning rates for plug-in classifiers". In: *The Annals of statistics* 35.2 (2007), pp. 608–633 (cit. on p. 103).

[84] Harriet E Baber. "Adaptive preference". In: *Social Theory and Practice* 33.1 (2007), pp. 105–126 (cit. on p. 11).

[85] Manel Baucells and Rakesh K Sarin. "Satiation in discounted utility". In: *Operations research* 55.1 (2007), pp. 170–181 (cit. on pp. 11, 19, 39, 40).

[86] James Bennett, Stan Lanning, et al. "The netflix prize". In: *Proceedings of KDD cup and workshop*. Vol. 2007. New York. 2007, p. 35 (cit. on pp. 8, 39).

[87] Sergio Firpo. "Efficient semiparametric estimation of quantile treatment effects". In: *Econometrica* 75.1 (2007), pp. 259–276 (cit. on p. 100).

[88] Pavlo A Krokhmal. "Higher moment coherent risk measures". In: (2007) (cit. on p. 71).
[89] Fenrong Liu et al. *Changing for the better: Preference dynamics and agent diversity*. Institute for Logic, Language and Computation, 2007 (cit. on p. 11).
[90] Rémi Munos. "Performance bounds in ℓ_p-norm for approximate value iteration". In: *SIAM journal on control and optimization* 46.2 (2007), pp. 541–561 (cit. on p. 48).
[91] Anastasios Tsiatis. *Semiparametric theory and missing data*. Springer Science & Business Media, 2007 (cit. on pp. 83, 92, 193).
[92] Daniel Acuna and Paul Schrater. "Bayesian modeling of human sequential decision-making on the multi-armed bandit problem". In: *Proceedings of the 30th annual conference of the cognitive science society*. Vol. 100. Washington, DC: Cognitive Science Society. 2008, pp. 200–300 (cit. on p. 12).
[93] Paul Grossman. "On measuring mindfulness in psychosomatic and psychological research." In: (2008) (cit. on p. 23).
[94] Dino Pedreshi, Salvatore Ruggieri, and Franco Turini. "Discrimination-aware data mining". In: *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM. 2008, pp. 560–568 (cit. on p. 34).
[95] James Robins, Lingling Li, Eric Tchetgen, Aad van der Vaart, et al. "Higher order influence functions and minimax estimation of nonlinear functionals". In: *Probability and statistics: essays in honor of David A. Freedman*. Institute of Mathematical Statistics, 2008, pp. 335–421 (cit. on p. 96).
[96] Alejandro Balbás, José Garrido, and Silvia Mayoral. "Properties of distortion risk measures". In: *Methodology and Computing in Applied Probability* 11.3 (2009), pp. 385–399 (cit. on p. 76).
[97] Yehuda Koren, Robert Bell, and Chris Volinsky. "Matrix factorization techniques for recommender systems". In: *Computer* 42.8 (2009), pp. 30–37 (cit. on p. 8).
[98] Victor H Peña, Tze Leung Lai, and Qi-Man Shao. *Self-normalized processes: Limit theory and Statistical Applications*. Springer, 2009 (cit. on p. 41).
[99] Joaquin Quionero-Candela, Masashi Sugiyama, Anton Schwaighofer, and Neil D Lawrence. *Dataset shift in machine learning*. The MIT Press, 2009 (cit. on p. 90).
[100] R Tyrrell Rockafellar and Roger J-B Wets. *Variational analysis*. Vol. 317. Springer Science & Business Media, 2009 (cit. on p. 185).
[101] Hervé Abdi. "Holm's sequential Bonferroni procedure". In: *Encyclopedia of research design* 1.8 (2010), pp. 1–8 (cit. on p. 19).
[102] Bibhas Chakraborty, Susan Murphy, and Victor Strecher. "Inference for non-regular parameters in optimal dynamic treatment regimes". In: *Statistical methods in medical research* 19.3 (2010), pp. 317–343 (cit. on p. 98).
[103] Thomas Jaksch, Ronald Ortner, and Peter Auer. "Near-optimal Regret Bounds for Reinforcement Learning." In: *Journal of Machine Learning Research* 11.4 (2010) (cit. on p. 48).
[104] Lihong Li, Wei Chu, John Langford, and Robert E Schapire. "A contextual-bandit approach to personalized news article recommendation". In: *Proceedings of the 19th international conference on World wide web*. 2010, pp. 661–670 (cit. on p. 9).

[105] Christoph Rothe. "Nonparametric estimation of distributional policy effects". In: *Journal of Econometrics* 155.1 (2010), pp. 56–70 (cit. on p. 100).

[106] Thomas Schäfer and Peter Sedlmeier. "What makes us like music? Determinants of music preference." In: *Psychology of Aesthetics, Creativity, and the Arts* 4.4 (2010), p. 223 (cit. on pp. 14, 19).

[107] Wenjing Zheng and Mark J van der Laan. "Asymptotic theory for cross-validated targeted maximum likelihood estimation". In: (2010) (cit. on p. 96).

[108] Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. "Improved algorithms for linear stochastic bandits". In: *Advances in Neural Information Processing Systems*. 2011, pp. 2312–2320 (cit. on pp. 41, 55).

[109] Nicole Fortin, Thomas Lemieux, and Sergio Firpo. "Decomposition methods in economics". In: *Handbook of labor economics*. Vol. 4. Elsevier, 2011, pp. 1–102 (cit. on p. 100).

[110] Laurie A Greco, Ruth A Baer, and Gregory T Smith. "Assessing mindfulness in children and adolescents: development and validation of the Child and Adolescent Mindfulness Measure (CAMM)." In: *Psychological assessment* 23.3 (2011), p. 606 (cit. on p. 23).

[111] Toshihiro Kamishima, Shotaro Akaho, and Jun Sakuma. "Fairness-aware learning through regularization approach". In: *2011 IEEE 11th International Conference on Data Mining Workshops*. IEEE. 2011, pp. 643–650 (cit. on pp. 26, 34).

[112] Eric B Laber and Susan A Murphy. "Adaptive confidence intervals for the test error in classification". In: *Journal of the American Statistical Association* 106.495 (2011), pp. 904–913 (cit. on p. 98).

[113] Michael D Lee, Shunan Zhang, Miles Munro, and Mark Steyvers. "Psychological models of human and optimal performance in bandit problems". In: *Cognitive Systems Research* 12.2 (2011), pp. 164–174 (cit. on p. 12).

[114] Sébastien Bubeck, Nicolo Cesa-Bianchi, et al. "Regret analysis of stochastic and nonstochastic multi-armed bandit problems". In: *Foundations and Trends® in Machine Learning* 5.1 (2012), pp. 1–122 (cit. on p. 55).

[115] Felipe Caro and Victor Martıénez-de-Albéniz. "Product and price competition with satiation effects". In: *Management Science* 58.7 (2012), pp. 1357–1373 (cit. on p. 39).

[116] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. "Fairness through awareness". In: *Proceedings of the 3rd innovations in theoretical computer science conference*. ACM. 2012, pp. 214–226 (cit. on pp. 34, 84).

[117] Yonina Eldar and Gitta Kutyniok. *Compressed Sensing: Theory and Applications*. Jan. 2012. ISBN: 978-1107005587. DOI: 10.1017/CBO9780511794308 (cit. on p. 60).

[118] Keisuke Hirano and Jack R Porter. "Impossibility results for nondifferentiable functionals". In: *Econometrica* 80.4 (2012), pp. 1769–1790 (cit. on p. 98).

[119] M Juraska, PB Gilbert, X Lu, M Zhang, M Davidian, and AA Tsiatis. "speff2trial: Semiparametric efficient estimation for a two-sample treatment effect". In: *R package version* 1.4 (2012) (cit. on p. 104).

[120] Faisal Kamiran and Toon Calders. "Data preprocessing techniques for classification without discrimination". In: *Knowledge and Information Systems* 33.1 (2012), pp. 1–33 (cit. on pp. 36, 37).

[121] Cheekiat Low, Dessislava Pachamanova, and Melvyn Sim. “Skewness-Aware Asset Allocation: A New Theoretical Framework and Empirical Evidence”. In: *Mathematical Finance: An International Journal of Mathematics, Statistics and Financial Economics* 22.2 (2012), pp. 379–410 (cit. on p. 32).

[122] Ronald Ortner and Daniil Ryabko. “Online regret bounds for undiscounted continuous reinforcement learning”. In: *Advances in Neural Information Processing Systems*. 2012, pp. 1763–1771 (cit. on p. 48).

[123] Ronald Ortner, Daniil Ryabko, Peter Auer, and Rémi Munos. “Regret bounds for restless markov bandits”. In: *International Conference on Algorithmic Learning Theory*. Springer. 2012, pp. 214–228 (cit. on pp. 47, 129).

[124] Takayuki Osogami. “Robustness and risk-sensitivity in Markov decision processes”. In: *Advances in Neural Information Processing Systems*. 2012, pp. 233–241 (cit. on p. 26).

[125] Suvrit Sra, Sebastian Nowozin, and Stephen J Wright. *Optimization for machine learning*. Mit Press, 2012 (cit. on p. 26).

[126] Baqun Zhang, Anastasios A Tsiatis, Eric B Laber, and Marie Davidian. “A robust method for estimating optimal treatment regimes”. In: *Biometrics* 68.4 (2012), pp. 1010–1018 (cit. on p. 86).

[127] Yingqi Zhao, Donglin Zeng, A John Rush, and Michael R Kosorok. “Estimating individualized treatment rules using outcome weighted learning”. In: *Journal of the American Statistical Association* 107.499 (2012), pp. 1106–1118 (cit. on p. 103).

[128] Mohammad Gheshlaghi Azar, Alessandro Lazaric, and Emma Brunskill. “Sequential transfer in multi-armed bandit with finite set of models”. In: *Proceedings of the 26th International Conference on Neural Information Processing Systems-Volume 2*. 2013, pp. 2220–2228 (cit. on pp. 55, 56).

[129] Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration inequalities: A nonasymptotic theory of independence*. Oxford university press, 2013 (cit. on p. 179).

[130] Bibhas Chakraborty. *Statistical methods for dynamic treatment regimes*. Springer, 2013 (cit. on p. 83).

[131] Luc Devroye, László Györfi, and Gábor Lugosi. *A probabilistic theory of pattern recognition*. Vol. 31. Springer Science & Business Media, 2013 (cit. on pp. 71, 180).

[132] Markus Frölich and Blaise Melly. “Unconditional quantile treatment effects under endogeneity”. In: *Journal of Business & Economic Statistics* 31.3 (2013), pp. 346–357 (cit. on p. 100).

[133] Julian John McAuley and Jure Leskovec. “From amateurs to connoisseurs: modeling the evolution of user expertise through online reviews”. In: *Proceedings of the 22nd international conference on World Wide Web*. ACM. 2013, pp. 897–908 (cit. on pp. 8, 39).

[134] Amir Sani, Alessandro Lazaric, and Rémi Munos. “Risk-aversion in multi-armed bandits”. In: *arXiv preprint arXiv:1301.1936* (2013) (cit. on p. 71).

[135] Vladimir Vapnik. *The nature of statistical learning theory*. Springer science & business media, 2013 (cit. on p. 70).

[136] Baqun Zhang, Anastasios A Tsiatis, Eric B Laber, and Marie Davidian. "Robust estimation of optimal dynamic treatment regimes for sequential treatment decisions". In: *Biometrika* 100.3 (2013), pp. 681–694 (cit. on p. 103).

[137] Tomas Björk, Agatha Murgoci, and Xun Yu Zhou. "Mean–variance portfolio optimization with state-dependent risk aversion". In: *Mathematical Finance: An International Journal of Mathematics, Statistics and Financial Economics* 24.1 (2014), pp. 1–24 (cit. on p. 69).

[138] Ilan Kremer, Yishay Mansour, and Motty Perry. "Implementing the wisdom of the crowd". In: *Journal of Political Economy* 122 (2014), pp. 988–1012 (cit. on p. 55).

[139] Mark J van der Laan and Alexander R Luedtke. "Targeted learning of an optimal dynamic treatment, and statistical inference for its mean outcome". In: (2014) (cit. on pp. 83, 98, 103).

[140] Eric B Laber, Daniel J Lizotte, Min Qian, William E Pelham, and Susan A Murphy. "Dynamic treatment regimes: Technical challenges and applications". In: *Electronic Journal of Statistics* 8.1 (2014), p. 1225 (cit. on pp. 83, 98).

[141] Pei Lee, Laks VS Lakshmanan, Mitul Tiwari, and Sam Shah. "Modeling impression discounting in large-scale recommender systems". In: *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*. 2014, pp. 1837–1846 (cit. on p. 11).

[142] Rémi Munos. "From bandits to Monte-Carlo Tree Search: The optimistic principle applied to optimization and planning". In: (2014) (cit. on p. 71).

[143] Paul B Reverdy, Vaibhav Srivastava, and Naomi Ehrich Leonard. "Modeling human decision making in generalized Gaussian multiarmed bandits". In: *Proceedings of the IEEE* 102.4 (2014), pp. 544–571 (cit. on p. 12).

[144] Phillip J Schulte, Anastasios A Tsiatis, Eric B Laber, and Marie Davidian. "Q-and A-learning methods for estimating optimal dynamic treatment regimes". In: *Statistical Science: a review Journal of the Institute of Mathematical Statistics* 29.4 (2014), p. 640 (cit. on p. 83).

[145] Alexander Shapiro, Darinka Dentcheva, and Andrzej Ruszczyński. *Lectures on stochastic programming: modeling and theory*. SIAM, 2014 (cit. on p. 71).

[146] Karen Simonyan and Andrew Zisserman. "Very deep convolutional networks for large-scale image recognition". In: *arXiv preprint arXiv:1409.1556* (2014) (cit. on p. 81).

[147] Alexander Zimin, Rasmus Ibsen-Jensen, and Krishnendu Chatterjee. "Generalized risk-aversion in stochastic multi-armed bandits". In: *arXiv preprint arXiv:1405.0833* (2014) (cit. on p. 71).

[148] Yinlam Chow, Aviv Tamar, Shie Mannor, and Marco Pavone. "Risk-sensitive and robust decision-making: a CVaR optimization approach". In: *Advances in Neural Information Processing Systems*. 2015, pp. 1522–1530 (cit. on p. 26).

[149] Javier Garcıa and Fernando Fernández. "A comprehensive survey on safe reinforcement learning". In: *Journal of Machine Learning Research* 16.1 (2015), pp. 1437–1480 (cit. on p. 26).

[150] Carlos A Gomez-Uribe and Neil Hunt. "The netflix recommender system: Algorithms, business value, and innovation". In: *ACM Transactions on Management Information Systems (TMIS)* 6.4 (2015), pp. 1–19 (cit. on p. 8).

[151] F. Maxwell Harper and Joseph A. Konstan. “The MovieLens Datasets: History and Context”. In: *ACM Trans. Interact. Intell. Syst.* (2015) (cit. on p. 156).

[152] Sarika Jain, Anjali Grover, Praveen Singh Thakur, and Sourabh Kumar Choudhary. “Trends, problems and solutions of recommender system”. In: *International conference on computing, communication & automation*. IEEE. 2015, pp. 955–958 (cit. on p. 8).

[153] Komal Kapoor, Karthik Subbian, Jaideep Srivastava, and Paul Schrater. “Just in time recommendations: Modeling the dynamics of boredom in activity streams”. In: *Proceedings of the eighth ACM international conference on web search and data mining*. 2015, pp. 233–242 (cit. on pp. 11, 40).

[154] Yishay Mansour, Aleksandrs Slivkins, and Vasilis Syrgkanis. “Bayesian Incentive-Compatible Bandit Exploration”. In: *ACM Conf. on Economics and Computation (EC)*. 2015 (cit. on p. 55).

[155] Dimitrios Rafailidis and Alexandros Nanopoulos. “Modeling users preference dynamics and side information in recommender systems”. In: *IEEE Transactions on Systems, Man, and Cybernetics: Systems* 46.6 (2015), pp. 782–792 (cit. on p. 11).

[156] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. “ImageNet Large Scale Visual Recognition Challenge”. In: *International Journal of Computer Vision (IJCV)* 115.3 (2015), pp. 211–252. DOI: 10.1007/s11263-015-0816-y (cit. on p. 81).

[157] Adith Swaminathan and Thorsten Joachims. “Counterfactual risk minimization: Learning from logged bandit feedback”. In: *International Conference on Machine Learning*. 2015, pp. 814–823 (cit. on pp. 8, 39).

[158] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. “Going deeper with convolutions”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015, pp. 1–9 (cit. on p. 81).

[159] Sattar Vakili and Qing Zhao. “Mean-variance and value at risk in multi-armed bandit problems”. In: *2015 53rd Annual Allerton Conference on Communication, Control, and Computing (Allerton)*. IEEE. 2015, pp. 1330–1335 (cit. on p. 71).

[160] Julius A Adebayo et al. “FairML: ToolBox for diagnosing bias in predictive modeling”. PhD thesis. Massachusetts Institute of Technology, 2016 (cit. on p. 35).

[161] Gal Bahar, Rann Smorodinsky, and Moshe Tennenholtz. “Economic Recommendation Systems: One Page Abstract”. In: *Proceedings of the 2016 ACM Conference on Economics and Computation*. EC ’16. Maastricht, The Netherlands: ACM, 2016, pp. 757–757. ISBN: 978-1-4503-3936-0. DOI: 10.1145/2940716.2940719. URL: http://doi.acm.org/10.1145/2940716.2940719 (cit. on p. 55).

[162] Moritz Hardt, Nimrod Megiddo, Christos Papadimitriou, and Mary Wootters. “Strategic classification”. In: *Proceedings of the 2016 ACM conference on innovations in theoretical computer science*. 2016, pp. 111–122 (cit. on p. 2).

[163] Moritz Hardt, Eric Price, and Nathan Srebro. “Equality of opportunity in supervised learning”. In: *Proceedings of the 30th International Conference on Neural Information Processing Systems*. 2016, pp. 3323–3331 (cit. on p. 84).

[164] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. “Deep residual learning for image recognition”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 770–778 (cit. on p. 81).

[165] Hoda Heidari, Michael J Kearns, and Aaron Roth. “Tight Policy Regret Bounds for Improving and Decaying Bandits.” In: *IJCAI*. 2016, pp. 1562–1570 (cit. on p. 40).

[166] Matthew Joseph, Michael Kearns, Jamie Morgenstern, and Aaron Roth. “Fairness in learning: Classic and contextual bandits”. In: *arXiv preprint arXiv:1605.07139* (2016) (cit. on p. 54).

[167] Edward H Kennedy. “Semiparametric theory and empirical processes in causal inference”. In: *Statistical causal inferences and their applications in public health research*. Springer, 2016, pp. 141–167 (cit. on pp. 83, 92).

[168] Nathan Korda, Balázs Szörényi, and Shuai Li. “Distributed Clustering of Linear Bandits in Peer to Peer Networks”. In: *Proceedings of the 33rd International Conference on International Conference on Machine Learning*. 2016 (cit. on p. 55).

[169] Damien Lefortier, Adith Swaminathan, Xiaotao Gu, Thorsten Joachims, and Maarten de Rijke. “Large-scale validation of counterfactual learning methods: A test-bed”. In: *arXiv preprint arXiv:1612.00367* (2016) (cit. on p. 12).

[170] Zhongqi Lu and Qiang Yang. “Partially Observable Markov Decision Process for Recommender Systems”. In: *CoRR* abs/1608.07793 (2016) (cit. on p. 56).

[171] Alexander R Luedtke and Mark J van der Laan. “Statistical inference for the mean outcome under a possibly non-unique optimal treatment strategy”. In: *Annals of statistics* 44.2 (2016), p. 713 (cit. on p. 98).

[172] Hélio M de Oliveira and Renato J Cintra. “A New Information Theoretical Concept: Information-Weighted Heavy-tailed Distributions”. In: *arXiv preprint arXiv:1601.06412* (2016) (cit. on pp. 33, 34).

[173] LA Prashanth, Cheng Jie, Michael Fu, Steve Marcus, and Csaba Szepesvári. “Cumulative prospect theory meets reinforcement learning: Prediction and control”. In: *International Conference on Machine Learning*. PMLR. 2016, pp. 1406–1415 (cit. on pp. 26, 70, 71, 76).

[174] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. “Rethinking the inception architecture for computer vision”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 2818–2826 (cit. on p. 81).

[175] Iván Dıéaz. “Efficient estimation of quantiles in missing data models”. In: *Journal of Statistical Planning and Inference* 190 (2017), pp. 39–51 (cit. on pp. 92, 100).

[176] Aditya Gopalan, LA Prashanth, Michael Fu, and Steve Marcus. “Weighted bandits or: How bandits learn distorted values that are not expected”. In: *Thirty-First AAAI Conference on Artificial Intelligence*. Vol. 31. 1. 2017 (cit. on pp. 26, 71).

[177] Thorsten Joachims, Adith Swaminathan, and Tobias Schnabel. “Unbiased learning-to-rank with biased feedback”. In: *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining*. 2017, pp. 781–789 (cit. on pp. 8, 39).

[178] Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. “Counterfactual fairness”. In: *Advances in Neural Information Processing Systems*. 2017, pp. 4066–4076 (cit. on p. 34).

[179] Nir Levine, Koby Crammer, and Shie Mannor. "Rotting bandits". In: *Advances in neural information processing systems*. 2017, pp. 3074–3083 (cit. on pp. 11, 40).

[180] Kristin A Linn, Eric B Laber, and Leonard A Stefanski. "Interactive Q-learning for quantiles". In: *Journal of the American Statistical Association* 112.518 (2017), pp. 638–649 (cit. on pp. 83, 86).

[181] Jian Wei, Jianhua He, Kai Chen, Yi Zhou, and Zuoyin Tang. "Collaborative filtering and deep learning based recommendation system for cold start items". In: *Expert Systems with Applications* 69 (2017), pp. 29–39 (cit. on p. 8).

[182] Zhifei Zhang, Yang Song, and Hairong Qi. "Age Progression/Regression by Conditional Adversarial Autoencoder". In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE. 2017 (cit. on p. 36).

[183] Rachel K. E. Bellamy, Kuntal Dey, Michael Hind, Samuel C. Hoffman, Stephanie Houde, Kalapriya Kannan, Pranay Lohia, Jacquelyn Martino, Sameep Mehta, Aleksandra Mojsilovic, Seema Nagar, Karthikeyan Natesan Ramamurthy, John Richards, Diptikalyan Saha, Prasanna Sattigeri, Moninder Singh, Kush R. Varshney, and Yunfeng Zhang. *AI Fairness 360: An Extensible Toolkit for Detecting, Understanding, and Mitigating Unwanted Algorithmic Bias*. Oct. 2018. URL: https://arxiv.org/abs/1810.01943 (cit. on pp. 36, 120).

[184] Léon Bottou, Frank E Curtis, and Jorge Nocedal. "Optimization methods for large-scale machine learning". In: *Siam Review* 60.2 (2018), pp. 223–311 (cit. on p. 185).

[185] Asaf Cassel, Shie Mannor, and Assaf Zeevi. "A General Framework for Bandit Problems Beyond Cumulative Objectives". In: (2018) (cit. on p. 71).

[186] Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. "Double/debiased machine learning for treatment and structural parameters". In: *The Econometrics Journal* 21.1 (2018), pp. C1–C68 (cit. on pp. 83, 92, 96).

[187] Luc Devroye, Abbas Mehrabian, and Tommy Reddad. "The total variation distance between high-dimensional Gaussians". In: *arXiv preprint arXiv:1810.08693* (2018) (cit. on p. 151).

[188] John Duchi and Hongseok Namkoong. "Learning models with uniform performance via distributionally robust optimization". In: *Annals of Statistics* (2018) (cit. on p. 69).

[189] Jeff Galak and Joseph P Redden. "The properties and antecedents of hedonic decline". In: *Annual review of psychology* 69 (2018), pp. 1–25 (cit. on pp. 9, 11, 19, 40).

[190] Xue Dong He, Roy Kouwenberg, and Xun Yu Zhou. "Inverse S-shaped probability weighting and its impact on investment". In: *Mathematical Control & Related Fields* (2018) (cit. on p. 32).

[191] Christopher Hillar and Andre Wibisono. *Maximum entropy distributions on graphs*. 2018. arXiv: 1301.3321 [math.ST] (cit. on p. 158).

[192] Cheng Jie, LA Prashanth, Michael Fu, Steve Marcus, and Csaba Szepesvári. "Stochastic optimization in a cumulative prospect theory framework". In: *IEEE Transactions on Automatic Control* 63.9 (2018), pp. 2867–2882 (cit. on p. 71).

[193] Robert Kleinberg and Nicole Immorlica. "Recharging bandits". In: *2018 IEEE 59th Annual Symposium on Foundations of Computer Science (FOCS)*. IEEE. 2018, pp. 309–319 (cit. on pp. 11, 40).

[194] Yang Liu and Chien-Ju Ho. "Incentivizing high quality user contributions: New arm generation in bandit learning". In: *Thirty-Second AAAI Conference on Artificial Intelligence*. 2018 (cit. on p. 54).

[195] Ningning Ma, Xiangyu Zhang, Hai-Tao Zheng, and Jian Sun. "Shufflenet v2: Practical guidelines for efficient cnn architecture design". In: *Proceedings of the European conference on computer vision (ECCV)*. 2018, pp. 116–131 (cit. on p. 81).

[196] James McInerney, Benjamin Lacker, Samantha Hansen, Karl Higley, Hugues Bouchard, Alois Gruson, and Rishabh Mehrotra. "Explore, exploit, and explain: personalizing explainable recommendations with bandits". In: *Proceedings of the 12th ACM conference on recommender systems*. 2018, pp. 31–39 (cit. on pp. 9, 10, 12, 25).

[197] Max Simchowitz, Horia Mania, Stephen Tu, Michael I Jordan, and Benjamin Recht. "Learning Without Mixing: Towards A Sharp Analysis of Linear System Identification". In: *Conference On Learning Theory*. 2018, pp. 439–473 (cit. on pp. 41, 50, 133, 136, 137).

[198] Lan Wang, Yu Zhou, Rui Song, and Ben Sherwood. "Quantile-optimal treatment regimes". In: *Journal of the American Statistical Association* 113.523 (2018), pp. 1243–1254 (cit. on pp. 83, 86).

[199] Romain Warlop, Alessandro Lazaric, and Jérémie Mary. "Fighting boredom in recommender systems with linear reinforcement learning". In: *Advances in Neural Information Processing Systems*. 2018, pp. 1757–1768 (cit. on p. 41).

[200] Guanjie Zheng, Fuzheng Zhang, Zihan Zheng, Yang Xiang, Nicholas Jing Yuan, Xing Xie, and Zhenhui Li. "DRN: A deep reinforcement learning framework for news recommendation". In: *Proceedings of the 2018 World Wide Web Conference*. 2018, pp. 167–176 (cit. on p. 25).

[201] Soumya Basu, Rajat Sen, Sujay Sanghavi, and Sanjay Shakkottai. "Blocking Bandits". In: *Advances in Neural Information Processing Systems*. 2019, pp. 4785–4794 (cit. on pp. 11, 40).

[202] Omar Besbes, Yonatan Gur, and Assaf Zeevi. "Optimal exploration–exploitation in a multi-armed bandit problem with non-stationary rewards". In: *Stochastic Systems* 9.4 (2019), pp. 319–337 (cit. on p. 51).

[203] Sanjay P Bhat and Prashanth LA. "Concentration of risk measures: A Wasserstein distance approach". In: *Advances in Neural Information Processing Systems* 32 (2019), pp. 11762–11771 (cit. on p. 76).

[204] Minmin Chen, Alex Beutel, Paul Covington, Sagar Jain, Francois Belletti, and Ed H Chi. "Top-k off-policy correction for a REINFORCE recommender system". In: *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*. 2019, pp. 456–464 (cit. on pp. 9, 10, 12, 25).

[205] Lee Cohen and Yishay Mansour. "Optimal Algorithm for Bayesian Incentive-Compatible". In: *ACM Conf. on Economics and Computation (EC)*. 2019 (cit. on p. 55).

[206] Rick Durrett. *Probability: theory and examples*. Vol. 49. Cambridge university press, 2019 (cit. on pp. 34, 117).

[207] Kenneth Holstein, Jennifer Wortman Vaughan, Hal Daumé III, Miro Dudik, and Hanna Wallach. "Improving fairness in machine learning systems: What do industry practi-

tioners need?" In: *Proceedings of the 2019 CHI conference on human factors in computing systems*. 2019, pp. 1–16 (cit. on p. 2).

[208] Eugene Ie, Chih-wei Hsu, Martin Mladenov, Vihan Jain, Sanmit Narvekar, Jing Wang, Rui Wu, and Craig Boutilier. "Recsim: A configurable simulation platform for recommender systems". In: *arXiv preprint arXiv:1909.04847* (2019) (cit. on p. 25).

[209] Nathan Kallus, Xiaojie Mao, and Masatoshi Uehara. "Localized Debiased Machine Learning: Efficient Estimation of Quantile Treatment Effects, Conditional Value at Risk, and Beyond". In: *arXiv preprint arXiv:1912.12945* (2019) (cit. on p. 86).

[210] Michael R Kosorok and Eric B Laber. "Precision medicine". In: *Annual review of statistics and its application* 6 (2019), pp. 263–286 (cit. on p. 83).

[211] Liu Leqi, Adarsh Prasad, and Pradeep Ravikumar. "On Human-Aligned Risk Minimization". In: *Advances in Neural Information Processing Systems (NeurIPS)* (2019) (cit. on p. 6).

[212] Liu Leqi, Adarsh Prasad, and Pradeep Ravikumar. "On Human-Aligned Risk Minimization". In: *Advances in Neural Information Processing Systems*. 2019 (cit. on pp. 71, 76, 78, 81).

[213] Nikolai Matni and Stephen Tu. "A tutorial on concentration bounds for system identification". In: *2019 IEEE 58th Conference on Decision and Control (CDC)*. IEEE. 2019, pp. 3741–3749 (cit. on p. 132).

[214] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. "A survey on bias and fairness in machine learning". In: *arXiv preprint arXiv:1908.09635* (2019) (cit. on p. 84).

[215] Ciara Pike-Burke and Steffen Grunewalder. "Recovering Bandits". In: *Advances in Neural Information Processing Systems*. 2019, pp. 14122–14131 (cit. on pp. 40, 41, 48, 49, 56).

[216] Zhengling Qi, Jong-Shi Pang, and Yufeng Liu. "Estimating Individualized Decision Rules with Tail Controls". In: *arXiv preprint arXiv:1903.04367* (2019) (cit. on p. 84).

[217] Tuhin Sarkar and Alexander Rakhlin. "Near optimal finite time identification of arbitrary linear dynamical systems". In: *International Conference on Machine Learning*. 2019, pp. 5610–5618 (cit. on pp. 41, 50, 133).

[218] Julien Seznec, Andrea Locatelli, Alexandra Carpentier, Alessandro Lazaric, and Michal Valko. "Rotting bandits are no harder than stochastic ones". In: *The 22nd International Conference on Artificial Intelligence and Statistics*. 2019, pp. 2564–2572 (cit. on pp. 11, 40).

[219] Aleksandrs Slivkins. "Introduction to multi-armed bandits". In: *arXiv preprint arXiv:1904.07272* (2019) (cit. on pp. 9, 17, 55).

[220] Léonard Torossian, Aurélien Garivier, and Victor Picheny. "$\mathcal{X}$ -Armed Bandits: Optimizing Quantiles, CVaR and Other Risks". In: *Asian Conference on Machine Learning*. PMLR. 2019, pp. 252–267 (cit. on p. 71).

[221] Martin J Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*. Vol. 48. Cambridge University Press, 2019 (cit. on pp. 73, 75).

[222] Robert C Williamson and Aditya Krishna Menon. "Fairness risk measures". In: *arXiv preprint arXiv:1901.08665* (2019) (cit. on p. 34).

[223] Wei Xiao, Hao Helen Zhang, and Wenbin Lu. "Robust regression for optimal individualized treatment rules". In: *Statistics in medicine* 38.11 (2019), pp. 2059–2073 (cit. on p. 103).

[224] Daniel M Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. "Fine-tuning language models from human preferences". In: *arXiv preprint arXiv:1909.08593* (2019) (cit. on p. 2).

[225] Gal Bahar, Omer Ben-Porat, Kevin Leyton-Brown, and Moshe Tennenholtz. "Fiduciary bandits". In: *International Conference on Machine Learning*. PMLR. 2020, pp. 518–527 (cit. on p. 55).

[226] Andrea Barraza-Urbina and Dorota Glowacka. "Introduction to bandits in recommender systems". In: *Fourteenth ACM Conference on Recommender Systems*. 2020, pp. 748–750 (cit. on p. 9).

[227] Junyu Cao, Wei Sun, Zuo-Jun Max Shen, and Markus Ettl. "Fatigue-Aware Bandits for Dependent Click Models". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 34. 04. 2020, pp. 3341–3348 (cit. on p. 56).

[228] Leonardo Cella and Nicolò Cesa-Bianchi. "Stochastic bandits with delay-dependent payoffs". In: *International Conference on Artificial Intelligence and Statistics*. 2020, pp. 1168–1177 (cit. on pp. 11, 41, 48).

[229] Alexandra Chouldechova and Aaron Roth. "A snapshot of the frontiers of fairness in machine learning". In: *Communications of the ACM* 63.5 (2020), pp. 82–89 (cit. on p. 84).

[230] Chester Holtz, Chao Tao, and Guangyu Xi. *BanditPyLib: a lightweight python library for bandit algorithms*. Online at: https://github.com/Alanthink/banditpylib. Documentation at https://alanthink.github.io/banditpylib-doc. 2020. URL: https://github.com/Alanthink/banditpylib (cit. on p. 12).

[231] Edward H Kennedy. "Optimal doubly robust estimation of heterogeneous causal effects". In: *arXiv preprint arXiv:2004.14497* (2020) (cit. on pp. 100, 101).

[232] Edward H Kennedy, Sivaraman Balakrishnan, Max G'Sell, et al. "Sharp instruments for classifying compliers and generalizing causal effects". In: *Annals of Statistics* 48.4 (2020), pp. 2008–2030 (cit. on p. 198).

[233] Justin Khim, Liu Leqi, Adarsh Prasad, and Pradeep Ravikumar. "Uniform convergence of rank-weighted learning". In: *International Conference on Machine Learning*. PMLR. 2020, pp. 5254–5263 (cit. on pp. 70, 71, 76, 77).

[234] Jon Kleinberg and Manish Raghavan. "How do classifiers induce agents to invest effort strategically?" In: *ACM Transactions on Economics and Computation (TEAC)* 8.4 (2020), pp. 1–23 (cit. on p. 2).

[235] Tor Lattimore and Csaba Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020 (cit. on pp. 9, 12, 17, 22, 48, 51, 52, 55).

[236] Jaeho Lee, Sejun Park, and Jinwoo Shin. "Learning Bounds for Risk-sensitive Learning". In: *34th Conference on Neural Information Processing Systems (NeurIPS) 2020*. Neural Information Processing Systems. 2020 (cit. on pp. 69–71, 76–78, 173).

[237] Tian Li, Ahmad Beirami, Maziar Sanjabi, and Virginia Smith. "Tilted Empirical Risk Minimization". In: *International Conference on Learning Representations*. 2020 (cit. on p. 69).

[238] Alex Luedtke, Antoine Chambaz, et al. “Performance guarantees for policy learning”. In: *Annales de l’Institut Henri Poincaré, Probabilités et Statistiques*. Vol. 56. 3. Institut Henri Poincaré. 2020, pp. 2162–2188 (cit. on p. 83).

[239] Kanak Mahadik, Qingyun Wu, Shuai Li, and Amit Sabne. “Fast Distributed Bandits for Online Recommendation Systems”. In: *Proceedings of the 34th ACM International Conference on Supercomputing*. 2020 (cit. on p. 55).

[240] Rishabh Mehrotra, Niannan Xue, and Mounia Lalmas. “Bandit based optimization of multiple objectives on a music streaming platform”. In: *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*. 2020, pp. 3224–3233 (cit. on pp. 9, 10, 12, 25).

[241] Yonatan Mintz, Anil Aswani, Philip Kaminsky, Elena Flowers, and Yoshimi Fukuoka. “Nonstationary bandits with habituation and recovery dynamics”. In: *Operations Research* 68.5 (2020), pp. 1493–1516 (cit. on pp. 11, 41, 49).

[242] Weibin Mo, Zhengling Qi, and Yufeng Liu. “Learning Optimal Distributionally Robust Individualized Treatment Rules”. In: *Journal of the American Statistical Association* (2020), pp. 1–16 (cit. on p. 90).

[243] Vishakha Patil, Ganesh Ghalme, Vineet Nair, and Y Narahari. “Achieving fairness in the stochastic multi-armed bandit problem”. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 34. 04. 2020, pp. 5379–5386 (cit. on p. 54).

[244] Yuta Saito, Aihara Shunsuke, Matsutani Megumi, and Narita Yusuke. “Open Bandit Dataset and Pipeline: Towards Realistic and Reproducible Off-Policy Evaluation”. In: *arXiv preprint arXiv:2008.07146* (2020) (cit. on p. 12).

[245] Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. “Learning to summarize with human feedback”. In: *Advances in Neural Information Processing Systems* 33 (2020), pp. 3008–3021 (cit. on pp. 2, 108).

[246] Xiangyu Zhao, Xudong Zheng, Xiwang Yang, Xiaobing Liu, and Jiliang Tang. “Jointly Learning to Recommend and Advertise”. In: *KDD ’20: The 26th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 2020 (cit. on p. 56).

[247] Yifei Zhao, Yu-Hang Zhou, Mingdong Ou, Huan Xu, and Nan Li. “Maximizing Cumulative User Engagement in Sequential Recommendation: An Online Optimization Perspective”. In: *KDD ’20: The 26th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. Ed. by Rajesh Gupta, Yan Liu, Jiliang Tang, and B. Aditya Prakash. 2020 (cit. on p. 56).

[248] Saurabh Arora and Prashant Doshi. “A survey of inverse reinforcement learning: Challenges, methods and progress”. In: *Artificial Intelligence* 297 (2021), p. 103500 (cit. on p. 2).

[249] Susan Athey and Stefan Wager. “Policy learning with observational data”. In: *Econometrica* 89.1 (2021), pp. 133–161 (cit. on p. 103).

[250] Nicole Bäuerle and Alexander Glauner. “Minimizing spectral risk measures applied to Markov decision processes”. In: *Mathematical Methods of Operations Research* (2021), pp. 1–35 (cit. on pp. 79, 174).

[251] Yash Chandak, Shiv Shankar, and Philip S. Thomas. *High-Confidence Off-Policy (or Counterfactual) Variance Estimation*. 2021. arXiv: 2101.09847 [cs.LG] (cit. on p. 71).

[252] Minmin Chen, Bo Chang, Can Xu, and Ed H Chi. "User response models to improve a reinforce recommender system". In: *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*. 2021, pp. 121–129 (cit. on pp. 9, 10, 12, 25).

[253] Minmin Chen, Yuyan Wang, Can Xu, Ya Le, Mohit Sharma, Lee Richardson, Su-Lin Wu, and Ed Chi. "Values of User Exploration in Recommender Systems". In: *Fifteenth ACM Conference on Recommender Systems*. 2021, pp. 85–95 (cit. on pp. 9, 10, 12, 25).

[254] Jessica Dai, Sina Fazelpour, and Zachary Lipton. "Fair machine learning under partial compliance". In: *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*. 2021, pp. 55–65 (cit. on p. 2).

[255] GoComics. *GoComics*. https://www.gocomics.com. 2021 (cit. on p. 15).

[256] Nika Haghtalab, Nicole Immorlica, Brendan Lucier, and Jack Z Wang. "Maximizing welfare with incentive-aware evaluation mechanisms". In: *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence*. 2021, pp. 160–166 (cit. on p. 2).

[257] Matthew J Holland and El Mehdi Haress. "Spectral risk-based learning using unbounded losses". In: *arXiv preprint arXiv:2105.04816* (2021) (cit. on pp. 71, 76).

[258] Audrey Huang, Liu Leqi, Zachary Lipton, and Kamyar Azizzadenesheli. "Off-policy risk assessment in contextual bandits". In: *Advances in Neural Information Processing Systems* 34 (2021), pp. 23714–23726 (cit. on pp. 70, 71, 76, 77, 173, 183).

[259] Huiwen Jia, Cong Shi, and Siqian Shen. "Multi-armed Bandit with Sub-exponential Rewards". In: *Operations Research Letters* (2021). ISSN: 0167-6377. DOI: https://doi.org/10.1016/j.orl.2021.08.004. URL: https://www.sciencedirect.com/science/article/pii/S0167637721001231 (cit. on pp. 59, 60, 157).

[260] Liu Leqi and Edward H Kennedy. "Median optimal treatment regimes". In: *arXiv preprint arXiv:2103.01802* (2021) (cit. on p. 6).

[261] Liu Leqi, Fatma Kilinc Karzan, Zachary Lipton, and Alan Montgomery. "Rebounding bandits for modeling satiation effects". In: *Advances in Neural Information Processing Systems (NeurIPS)* (2021) (cit. on p. 6).

[262] Liu Leqi, Fatma Kilinc-Karzan, Zachary C Lipton, and Alan L Montgomery. "Rebounding Bandits for Modeling Satiation Effects". In: (2021) (cit. on pp. 11, 56).

[263] Gurobi Optimization LLC. *Gurobi Optimizer Reference Manual*. 2021. URL: http://www.gurobi.com (cit. on p. 53).

[264] Xinkun Nie and Stefan Wager. "Quasi-oracle estimation of heterogeneous treatment effects". In: *Biometrika* 108.2 (2021), pp. 299–319 (cit. on p. 101).

[265] Tom Ron, Omer Ben-Porat, and Uri Shalit. "Corporate Social Responsibility via Multi-Armed Bandits". In: *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. 2021, pp. 26–40 (cit. on p. 54).

[266] Jinchun Zhang, Andrea B Troxel, and Eva Petkova. "Robust index of confidence weighted learning for optimal individualized treatment rule estimation". In: *Stat* 10.1 (2021), e374 (cit. on p. 103).

[267] Omer Ben-Porat*, Lee Cohen*, Liu Leqi*, Zachary C Lipton, and Yishay Mansour. "Modeling attrition in recommender systems with departing bandits". In: *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)* (2022) (cit. on p. 6).

[268] Minmin Chen, Can Xu, Vince Gatto, Devanshu Jain, Aviral Kumar, and Ed Chi. "Off-Policy Actor-critic for Recommender Systems". In: *Proceedings of the 16th ACM Conference on Recommender Systems*. 2022, pp. 338–349 (cit. on pp. 9, 10, 12, 25).

[269] Jeff Galak, Jinwoo Kim, and Joseph P Redden. "Identifying the temporal profiles of hedonic decline". In: *Organizational Behavior and Human Decision Processes* 169 (2022), p. 104128 (cit. on p. 20).

[270] Nika Haghtalab, Michael Jordan, and Eric Zhao. "On-demand sampling: Learning optimally from multiple distributions". In: *Advances in Neural Information Processing Systems* 35 (2022), pp. 406–419 (cit. on p. 2).

[271] Amy K Heger, Liz B Marquis, Mihaela Vorvoreanu, Hanna Wallach, and Jennifer Wortman Vaughan. "Understanding Machine Learning Practitioners' Data Documentation Perceptions, Needs, Challenges, and Desiderata". In: *Proceedings of the ACM on Human-Computer Interaction* 6.CSCW2 (2022), pp. 1–29 (cit. on p. 2).

[272] Raphael Koster, Jan Balaguer, Andrea Tacchetti, Ari Weinstein, Tina Zhu, Oliver Hauser, Duncan Williams, Lucy Campbell-Gillingham, Phoebe Thacker, Matthew Botvinick, et al. "Human-centred mechanism design with Democratic AI". In: *Nature Human Behaviour* 6.10 (2022), pp. 1398–1407 (cit. on p. 108).

[273] Liu Leqi, Audrey Huang, Zachary Lipton, and Kamyar Azizzadenesheli. "Supervised Learning with General Risk Functionals". In: *International Conference on Machine Learning (ICML)* (2022) (cit. on p. 6).

[274] Charikleia Podimata. "Incentive-Aware Machine Learning for Decision Making". PhD thesis. Harvard University, 2022 (cit. on p. 2).

[275] Yuyan Wang, Mohit Sharma, Can Xu, Sriraj Badam, Qian Sun, Lee Richardson, Lisa Chung, Ed H Chi, and Minmin Chen. "Surrogate for Long-Term User Experience in Recommender Systems". In: *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 2022, pp. 4100–4109 (cit. on pp. 9, 10, 12, 25).

[276] Liu Leqi*, Giulio Zhou*, Fatma Kilinc-Karzan, Zachary Lipton, and Alan Montgomery. "A Field Test of Bandit Algorithms for Recommendations: Understanding the Validity of Assumptions on Human Preferences in Multi-armed Bandits". In: *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI)* (2023) (cit. on p. 6).

[277] Ada Martin, Valerie Chen, Sérgio Jesus, and Pedro Saleiro. "A Case Study on Designing Evaluations of ML Explanations with Simulated User Studies". In: *arXiv preprint arXiv:2302.07444* (2023) (cit. on p. 2).

[278] John C Duchi, Tatsunori Hashimoto, and Hongseok Namkoong. "Distributionally Robust Losses Against Mixture Covariate Shifts". In: () (cit. on pp. 26, 34, 35).

[279] Lecturer: Prof Roberto Oliveira and Scribes: Shangjie Yang. *Topics in statistical mathematics: Lecture 2*. https://w3.impa.br/~rimfo/estatistica_a16/notas_alunos/shangjie.pdf. Accessed: 2021-9-15 (cit. on p. 175).